AUTHOR: ALWAWI, I.

TITLE: Cognitive modelling and control of human error processes in human-computer interaction with safety critical IT systems in telehealth.

YEAR: 2017

OpenAIR citation:

ALWAWI, I. 2017. Cognitive modelling and control of human error processes in human-computer interaction with safety critical IT systems in telehealth. Robert Gordon University, PhD thesis. Held on OpenAIR [online]. Available from: https://openair.rgu.ac.uk

This work was submitted to- and approved by Robert Gordon University in partial fulfilment of the following degree:
Doctor of Philosophy, School of Computing Science and Digital Media.

Cognitive Modelling and Control of Human Error Processes in Human-Computer Interaction with Safety Critical IT Systems in Telehealth

IBRAHIM ALWAWI

A report submitted as part of the requirements
for the degree of Doctor of Philosophy
at the School of Computing Science and Digital Media
Robert Gordon University, Aberdeen, Scotland

December 2017

Supervisor : Prof. Patrik O' B. Holt
Dr Eyad Elyan

Dedication

Dedicated to the soul of my mother Haleemah and the soul of my eldest sister Ahlam
Who gave their lives in Gaza while helping others during their work for the United
Nations when I was a child

Acknowledgements

I would like to thank the following people for their assistance and support in helping me to complete this study: Prof. Patrik O Brian Holt, my main academic supervisor, for his support, constructive criticism and continuous guidance throughout the work. Dr Eyad Elyan, my second academic supervisor, for guiding me in the area of image processing and colour control. Dr Iain Pirie for his invaluable assistance in advising on and assessing the statistical work in this thesis. Mr James Ferguson and Dr Fiona Mair, clinical leads in the A&E department, Aberdeen Royal Infirmary and key staff using telehealth for follow up diagnosis, for providing clinical guidance and for supervision in obtaining medical images from the infirmary for the cellulitis (skin infection) experiments of the study.

Dr KarynWebster, Telehealth Clinical Fellow and speciality doctor in emergency medicine, A&E, NHSForth Valley, for guidance and for giving me the opportunity to deal with telehealth systems in real use and to analyse real cases. Prof. Augusto Azuara-Blanco, professor of ophthalmology, Centre for Experimental Medicine (formerly Centre for Vision and Vascular Science), School of Medicine, Dentistry and Biomedical Sciences, Queen's University Belfast, for his clinical guidance and supervision in obtaining and verifying medical images for the conjunctivitis (red eye) experiments of the study. Mr Alan Clark for his help with the English language and for his proofreading of the thesis.

I would like to thank my family back at home in Gaza City, Palestine, especially my beloved father Rezik and my noble wife Amany, who came all the way from Gaza to support me, for their patience, support and prayers for my success. Finally, my friends and congregation members in Aberdeen and at home for their continuous encouragement, patience and support.

Abstract

The field of telehealth has developed rapidly in recent years. It provides medical support particularly to those who are living in remote areas and in emergency cases. Although developments in both technology and practice have been rapid, there are still many gaps in our knowledge with regard to the effective application of telehealth. This study investigated human colour perception in telehealth, specifically the colour red as one of the key symptoms when diagnosing different pathologies. The quality of medical images is safety critical when transmitting the symptoms of pathologies in telehealth, as distorted or degraded colours may result in errors.

The study focused on the use of digital images in teleconsultation, particularly on images showing cellulitis (bacterial skin infection) and conjunctivitis (red eye) as case studies, as both of these pathologies involve the colour red in their diagnosis. The study proposed and tested the use of an image quality scale, which represented the level of image resolution; a red colour scale, which represented the intensity of redness in an image; and a confidence scale, which represented the levels of confidence that telehealth users had when judging the colour red. The research involved a series of experiments using hypothetico-deductive and formal hypothesis testing with two groups of participants, medical doctors and non-medical participants. The experiments were conducted in collaboration with the local National Health Service (NHS) Accident and Emergency (A&E) department at Aberdeen Royal Infirmary (ARI). Medical experts in ophthalmology and dermatology were also involved in selecting and verifying the relevant images. The study found that doctors and non-doctors were consistent in the majority of the experiments. The accuracy of the participants was demonstrably higher when using a colour scale with pictures, more so for the non-doctor group than the doctor group. It also found that the level of accuracy for both doctors and non-doctors was higher when using red colour scale of three divisions than when using a scale of five divisions. This result was supported by previous studies, which used telehealth for diagnosing extreme cases. The study also found that when the image quality was poor the participants had higher error rates and less consistency in their answers.

The study found poor correlation between accuracy, confidence and time for both participant groups. The study found that most participants in both doctor and non-doctor groups had high confidence most of the time, whether the accuracy was high or low. It was also found that medical background or clinical experience had no effect on the accuracy level across the experiment sets. In some cases, doctors with no or little experience had higher accuracy than those with greater experience. This result may have significant implications for the feasibility of involving non-doctors in the management of telehealth systems, especially in tasks not requiring medical skills, such as colour classification. This has the potential to provide a considerable saving in resources and costs for healthcare providers.

An auto-evaluation system was introduced, and proposed for further study, in order to improve the current telehealth diagnostic protocol and to avoid or prevent errors by making red colour classification more objective and accurate.

List of Abbreviations

The following are the Acronyms that used in the study:

A&E	Accident and Emergency
ACS	Academic Confidence Scale
ARI	Aberdeen Royal Infirmary
ATA	American Telemedicine Association
BPS	British Psychological Society
CE	Cognitive Engineering
CR	Contrast Ratio
CRT	Cathode Ray Tubes
CSE	Cognitive System Engineering
DSCQS	Double Stimulus Continuous Quality Scale Method
DSIS	Double Stimulus Impairment Scale Method
FR	Full Reference
GCH	Global Colour Histogram
GP	General Practitioner
HCI	Human Computing Interaction
IOM	Institute Of Medicine
JPEG	Joint Photographic Experts Group
LCH	Local Colour Histogram
ML	Machine Learning
MOS	Mean Opinion Score
NASA	National Aeronautics and Space Administration
NHS	National Health Service
PNG	Portable Network Graphics
PSNR	Peak Signal to Noise Ratio
RGU	Robert Gordon University
ROI	Region Of Interest
RR	Reduced Reference
SAF	Store And Forward Teleconsultation
SDS	Semantic Differential Scale
SSCQ	Single Stimulus Continuous Quality
SSIS	Single Stimulus Impairment Scale
TIFF	Tagged Image File Format

Declaration

I confirm that the work contained in this PhD project report has been composed solely by myself and has not been accepted in any previous application for a degree. All sources of information have been specifically acknowledged and all verbatim extracts are distinguished by quotation marks.

Signed .... Date ....
IBRAHIM ALWAWI

Contents

Dedication **ii**

Acknowledgements **iii**

Abstract **iv**

List of Abbreviations **vi**

Declaration **vii**

1 Introduction **1**

1.1 Research Background 1

1.2 The Aims of the Study 2

1.3 Research Problems 4

1.4 The Scope of the Study 5

1.5 Research Objectives 5

1.6 Research Questions 6

1.7 Thesis Organisation 7

2 Literature Review **9**

2.1 Introduction 9

2.2 Telehealth 10

2.2.1 The concept of telehealth 10

2.2.2 The benefits of telehealth 11

2.2.3 Challenges facing telehealth 12

2.2.4 Telehealth diagnostic methods 13

2.2.5 Use of Digital Images in Diagnosis 19

2.2.6 Telehealth applications 24

2.3 Cognitive Engineering (CE) link 31

2.3.1 Cognitive systems and human errors 33

2.3.2 Human error prediction 36

2.3.3 System responses to error 38
2.3.4 Users and individual differences in system design 39
2.4 The Use of a Confidence Scale 40
2.4.1 The concept of human confidence 40
2.4.2 The concept of overconfidence and underconfidence 40
2.4.3 The relationship between confidence and accuracy: 41
2.4.4 Time relationship with confidence and accuracy 45
2.5 Understanding Colour and Digital Images 45
2.5.1 Colour transmission 48
2.5.2 Image quality and colour resolution 49
2.6 Human Colour Perception and Diagnostic Accuracy 53
2.6.1 Human colour perception in key previous studies 54
2.6.2 Sex differences with colour perception and vision 57
2.7 The colour red and image matching in healthcare and telehealth 59
2.7.1 The importance of the colour red in healthcare and telehealth . . 59
2.7.2 Standardisation of teleconsultation 60
2.7.3 Using colour scales in medicine 64
2.8 Auto-classification Techniques for Images 70
2.8.1 Machine Learning 70
2.8.2 The use of image databases 72
2.8.3 Quarantining the region of interest and extracting colour features 74
2.8.4 Image Descriptors and Colour Histograms 75
2.8.5 Colour moment 77
2.8.6 Auto-evaluation system 79
2.9 Critical Discussion and Conclusion 81

3 Research Methodology and Methods 85
3.1 Introduction 85
3.2 Data Collection 85
3.3 Experimental Design 86
3.4 Machine Learning 87
3.5 The Experiment Sets in the Study 92
3.6 Population and Sampling Method 94
3.6.1 Medical doctors 94
3.6.2 Non-medical people 94
3.6.3 Sampling method and frame 94
3.7 Experimental Report 96
3.8 Research Hypotheses 97
3.9 Overview of Ethical Issues 97

3.9.1 Ethical codes and approvals 98
3.9.2 Rights of participants and data confidentiality 99
3.9.3 Images archive security 99
3.9.4 Risk and ethical impact 100
3.9.5 Feedback of the study 100
3.10 The Data Analysis 100
3.11 Validity and Reliability of the Methodology 101

4 Pilot Experiment 102
4.1 Introduction 102
4.2 Participants 103
4.3 Design and Variables 103
4.4 Materials 104
4.5 Procedures and Tasks 106
4.6 Results and Analysis 107
4.6.1 Results and analysis of task 1.A (colour description) 107
4.6.2 Results and analysis of task 1B (intensity and saturation values) 109
4.7 Results of experimental tasks 3A and 3B 115
4.8 Critical Discussion 120
4.9 Conclusion 121

5 Skin Infection Experimental Series 122
5.1 Introduction 122
5.2 Cellulitis Experimental Objectives 122
5.3 Experimental Tasks 123
5.4 Participants 124
5.5 Design and Variables 124
5.6 Materials 125
5.6.1 Images used for the experiments 125
5.7 The Results 127
5.7.1 Results for experiments 1A and 1B (image description) 127
5.7.2 The results of experiments 2A and 2B (image clustering) 131
5.7.3 The results of experiments 3A and 3B (image ranking) 133
5.7.4 Results of experiments 4A and 4B (image quality rating) 139
5.7.5 Results of experiments 5A and 5B (image matching) 142
5.8 Critical Discussion and Conclusion 148

6 Image Matching Experiment 150
6.1 Introduction 150

6.2 Objectives 150
6.3 Participants 151
6.4 Design and Variables 151
6.5 Procedure and Tasks 153
6.6 Results 154
6.7 Confidence and Accuracy 157
6.8 Time Results and Analysis 158
6.9 Relationship among accuracy, confidence and time 159
6.10 Critical Discussion of Results 160

7 Conjunctivitis Experiment 162

7.1 Introduction 162
7.2 Objectives of Conjunctivitis Experiments 162
7.3 Participants 163
7.4 Design and Variables 163
7.5 Materials Used in Conjunctivitis Experiments 163
7.5.1 Online question and answers sheets 163
7.5.2 Stimuli (conjunctivitis pathology images) 164
7.6 Tasks of the experiments 164
7.6.1 Experiment 1A 165
7.7 Confidence 168
7.8 Results 168
7.8.1 Results of Experiment 1D 168
7.8.2 Results of Experiment 1B 170
7.8.3 Results of Experiments 1C (with colour scale) and 1E (without using colour scale) 171
7.8.4 Results of experiments 2A and 2C 173
7.8.5 Results of experiments 2B and 2D 175
7.8.6 Results of experiment 3A, 3C and 3D 176
7.8.7 The results of experiment 3B 177
7.8.8 Results of experiment 3E 178
7.8.9 The results of experiment 3F 179
7.8.10 Results of experiment 4 180
7.8.11 Results of experiment 4A/1A 181
7.8.12 Results of experiment 4B 181
7.8.13 Results of experiments 4C/1C 182
7.8.14 Results of experiments 4D/1D 183
7.8.15 Experiments 1E and 4E. Comparing scales of 5 and 3 levels . . . 183
7.8.16 Accuracy based on the degree of redness in images 186

7.9 Relationships between Accuracy, Confidence and Time 188
7.10 Machine learning experiments 189
7.10.1 Colour feature presentation 189
7.10.2 Nearest neighbour 190
7.10.3 Training and testing 190
7.11 Summary and discussion of the results 196

8 Conclusion 199
8.1 Critical Discussion of the Main Results 199
8.2 Research Originality and Contributions 205
8.2.1 Research contribution 1: Introducing cellulitis and conjunctivitis to telehealth systems 205
8.2.2 Research contribution 2: Introducing and testing the use of colour scales in telehealth 205
8.2.3 Research contribution 3: Introducing and testing the use of image quality scales in telehealth 206
8.2.4 Research contribution 4: Introducing and testing confidence scales in telehealth 206
8.2.5 Research contribution 5: Combined design for measuring colour perception 206
8.2.6 Research contribution 6: The involvement of non-healthcare professionals in telehealth 206
8.2.7 Research contribution 7: A guide for an auto-evaluation system . 207
8.3 Limitations and Reflections 207
8.4 Recommendations for further work 209
8.4.1 Proposed auto-evaluation system for using digital images in telehealth 209
8.4.2 Mobile phone cameras experiment 209
8.4.3 Image eye tracking 211
8.4.4 Testing the results of this study in real use 211
8.5 Conclusion 211

9 References 213

A 230
A.1 Appendix 1: Proposed specification for auto evaluation system using digital images in telehealth 230
A.2 Appendix 2: Results and Analysis of Task1A (Colour Description) . . . 240
A.3 Appendix 3: Image Classification. Python source code 241

A.4 Appendix 4: Image Classification Results 243
A.5 Appendix 5: RGU Ethical Approval Form 252
A.6 Appendix 6: Example instruction and answer sheets 255
A.7 Appendix 7.1: Consent Form (Content Only) 256
A.8 Appendix 7.2: NHS Consent Form (Content Only) 257
A.9 Appendix 8: Colour Blind Test 259
A.10 Appendix 9: Instruction sheets 265
A.11 Appendix 10: Sample of master data for image matching experiment-Chapter 6 267
A.12 Appendix 11: Conjunctivitis digital images 272
A.13 Appendix 12: Red eye Online Instruction sheets-part1 279
A.14 Appendix 12.1: Red eye Online Instruction sheets and Consent-part2 . . 282
A.15 Appendix 13: Examples of red eye online experiment 284
A.16 Appendix 14: Mann-Whitney test results for joint experiments 5a and 5b 289
A.17 Appendix B-Cellulitis Results 293
A.18 Appendix D-Cellulitis Classification Result 300

List of Tables

3.1 Sample frame of the study for primary data 95

4.1 Colour characteristics in tasks 2 and 3 of the pilot experiment 105
4.2 Intensity scale used by the study 110
4.3 The accuracy of participants' results and Photoshop (intensity) 110
4.4 Saturation scale used by the study 110
4.5 The accuracy of participants' results and Photoshop (intensity) 111
4.6 The intensity of the colour chips used in experiments 2A and 2B 112
4.7 The number clusters generated by individual participants 114
4.8 The ranking of participants and values in Photoshop (Task 3A) 117
4.9 The ranking of participants and values in Photoshop (Task 3B) 118

5.1 Frequency of answers for all participants in experiments 1A and 1B . . . 128
5.2 Frequency of answers for doctor participants in experiments 1A and 1B 128
5.3 Frequency of answers for non-doctor participants in experiments 1A and 1B 129
5.4 Direct comparison of doctor and non-doctor participants in experiment 1A 129
5.5 Direct comparison of doctor and non-doctor participants in experiment 1B 130
5.6 Two-cluster solution in experiment 2A (image clustering) 131
5.7 Two-cluster solution in experiment 2B (image clustering) 132
5.8 Friedman and Kendall's values in experiment 3A 133
5.9 Average on 9-point scale by participants in experiment 3A 134
5.10 Kendall's Coefficient of Concordance for doctors for experiment 3A . . . 134
5.11 Kendall's Coefficient of Concordance for non-doctors for experiment 3A 134
5.12 Friedman and Kendall's values in experiment 3B 135
5.13 Average position of ranking by doctor and non-doctor participants in 3B 136
5.14 Kendall's Coefficient of Concordance for doctors in experiment 3B . . . 137
5.15 Kendall's Coefficient of Concordance for non-doctors in experiment 3B . 137

5.16 The correct answers when images were the same 141
5.17 The correct answers when images had slight differences 142
5.18 The correct answers when images had moderate differences 142
5.19 Friedman Test (ranking) for all participants 143
5.20 Agreement level between the image categories for all participants . . . 143
5.21 Friedman Test (ranking) for non-doctors 143
5.22 Agreement level between the image categories for non-doctors 143
5.23 Friedman Test (ranking) for doctors 144
5.24 Agreement level between the image categories for doctors 144
5.25 Mean confidence for all participants in experiment 4 147
5.26 Mean accuracy for all participants for experiment 4 147
5.27 Spearman correlation between confidence and accuracy for all participants148

6.1 Image pair descriptions 152
6.2 Mann-Whitney U test summary 155
6.3 Wilcoxon signed ranks test summary for all participants 156

7.1 Mann-Whitney 170
7.2 Mann-Whitney 170
7.3 Comparison of the accuracy of doctors and non-doctors in each image . 171
7.4 Mann-Whitney test summary 172
7.5 Wilcoxon signed ranks tests summary for doctors 172
7.6 Wilcoxon signed ranks tests summary for non-doctors 173
7.7 Mann-Whitney test summary 174
7.8 Spearman correlation in accuracy, confidence and time-experiment 3 . . 189
7.9 Classification Results for RGB and HSV - Cellulitis 192
7.10 Classification Results for RGB and HSV - Conjunctivitis 193

A.1 Image quality scale measurement 232
A.2 Red eye scale measurement 232
A.3 Operator confidence scale measurement 233
A.4 Examples of image atlas 233
A.5 Operator confidence scale measurement 237
A.6 Classification System 264
A.7 Classification System 265
A.8 Participants master data-Non-Dr female, example1 268
A.9 Participants master data-Non-Dr male, example2 269
A.10 Participants master data- Dr Male, example3 270
A.11 Participants master data- Dr Male, example4 271
A.12 Descriptive Statistics for experiments 5a and 5b 289

A.13 Ranks 289
A.14 Test Statistics 289
A.15 Classification System 1 290
A.16 Classification System 2 291
A.17 Classification System 3 292
A.18 Classification System 4 293
A.19 Classification System 5 294
A.20 Classification System 6 295
A.21 Classification System 7 296
A.22 Classification System 8 297
A.23 Classification System 9 298
A.24 Classification System 10 299
A.25 Classification System 11 300
A.26 Classification Error - Conjunctivitis two classes 300
A.27 Classification Error - Conjunctivitis three classes 300
A.28 Classification Error - Conjunctivitis five classes 301
A.29 Classification Error - Cellulitis two, three and five classes 301

List of Figures

2.1 llustrates that patient and specialist can communicate securely between physically separated locations using telecommunication technology Campos et al. (2012) 15

2.2 Medical booth in ARI Scottish Centre for Telehealth 15

2.3 Interaction in teleconsultation in ARI-taken by researcher 16

2.4 X-ray transmission in ARI-taken by researcher 16

2.5 Image (A); typical severe cellulitis from clinical resource efficiency support team (CREST) 2005. Image B; typical mild cellulitis from Newson (2015 26

2.6 The external structure of the eye showing the bulbar conjunctiva white area (Source: Bergman, 2015) 27

2.7 conjunctivitis severity levels improving gradually, taken by the researcher (July 2013) 30

2.8 conjunctivitis severity levels improving gradually, taken by the researcher (July 2013) 31

2.9 Human error systemic model. Adapted from Reason (2000) 35

2.10 (a) shot with Kodak V1003 with flash and incandescent, (b) shot with Kodak V1003 without flash under incandescent lighting. The resulting skin tone has an orange cast. 45

2.11 (a) An image made with a with Fuji F40 camera using flash demonstrates excellent (natural) skin colour, (b) An image of the same subject made with a Sony DSC-W200 camera using flash demonstrates a yellow cast to the skin colour 46

2.12 (a) The picture of a dress perceived as white or blue, (b) The picture of same dress in (a) perceived as unambiguous shades of yellow 55

2.13 Red eye grading guide designed by Allergan and used by the eye unit at ARI 67

2.14 Mild diffuse pink in allergic conjunctivitis, by Abelson (2010) 69

2.15 (a) Training stage and (b) Classification stage. Tsai (2007) 72

2.16 Region of interest selection by Rodriguez et al. 2013: (A) Dry eye image. (B) Region of interest (ROI) showing pathological redness. (C) Vascular structure in the ROI showing the prominence of graded horizontal vessels 74
2.17 Red eye auto-evaluation system 80

3.1 Agile Methodology stages-graphic designed by researcher 89
3.2 Red eye image quality scale used in the study 90
3.3 Pilot experiments 92
3.4 Cellulitis experiments 92
3.5 Image matching experiment 93
3.6 Red eye experiment 93

4.1 Non-medical images used in the colour description task 1 of pilot study 104
4.2 Colour chips used in colour description tasks 2 and 3 of pilot study . . . 105
4.3 Frequency of colour descriptions-Image 1 107
4.4 Frequency of colour descriptions-Image 2 107
4.5 Frequency of colour descriptions-Image 3 108
4.6 Frequency of colour descriptions-Image 4 108
4.7 Frequency of colour descriptions-Image 5 108
4.8 Frequency of colour descriptions-Image 6 108
4.9 Frequency of colour descriptions-Image 7 108
4.10 Frequency of colour descriptions-Image 8 108
4.11 Dendrogram for clustering by participants in task 2 A 112
4.12 Dendrogram for clustering by participants in task 2 B 113
4.13 The effect on participants' performance of using a colour guide 115
4.14 The effect on participants' performance of using a colour guide 118
4.15 The effect on participants' performance of using a colour guide 119

5.1 Examples of images from 1A and 1B 125
5.2 Examples of images from 2A and 2B 125
5.3 Examples of images from 3A and 3B 126
5.4 Examples of images from 4A and 4B 126
5.5 Examples of images from 5A and 5B 126
5.6 Dendrogram of image clustering in experiment 2A for non-doctors . . . 131
5.7 Dendrogram of image clustering in experiment 2A for doctors 131
5.8 Dendrogram of image clustering in experiment 2B for non-doctors . . . 132
5.9 Dendrogram of image clustering in experiment 2B for doctors 132
5.10 Dendrogram of image ranking by non-doctor participants in task 3A . . 135
5.11 Dendrogram of image ranking by doctor participants in task 3A 136
5.12 Dendrogram of image ranking by non-doctors in experiment 3B 137

5.13 Dendrogram of image ranking by doctors in experiment 3B 138
5.14 Boxplot of correlation in experiment 4A 140
5.15 Boxplot of correlation in experiment 4B 141
5.16 Comparison of accuracy with image differences 144
5.17 Comparison of accuracy with image differences 145
5.18 Comparison of accuracy with image differences 146

6.1 Q1 in image matching experiment 153
6.2 Q2 in image matching experiment 153
6.3 Q3 in image matching experiment 153
6.4 Relative accuracy of doctors and non-doctors at different levels of image modification 154
6.5 The confidence levels of doctors and non-doctors closely mirrored each other 157
6.6 Time taken with different redness degrees 158

7.1 Conjunctivitis grading Grad designed by Allergan and used by the eye unit at ARI 165
7.2 Image quality scale designed by the researcher 166
7.3 Degrees of redness in conjunctivitis 167
7.4 The effects on accuracy of using a colour scale 169
7.5 The effects on accuracy of using a colour scale 172
7.6 Comparison of accuracy and consistency for doctors and non-doctors . . 174
7.7 Comparison of accuracy and consistency for doctors and non-doctors . . 175
7.8 Comparison of accuracy and consistency for doctors and non-doctors . . 177
7.9 Comparison of accuracy for doctors and non-doctors 178
7.10 Comparison of accuracy of doctors and non-doctors 179
7.11 Comparison of accuracy of doctors and non-doctors 180
7.12 Comparison of accuracy between experiments 1A and 4A 181
7.13 Comparison of accuracy of doctors and non-doctors 182
7.14 Comparison of doctors and non-doctors in 1C and 4C 183
7.15 Comparison of doctors and non-doctors in 1D and 4D 184
7.16 The effect on accuracy of scale granularity 185
7.17 The effect on accuracy of scale granularity 186
7.18 Scale of three levels was employed 187
7.19 The effect of degree of redness on accuracy 188
7.20 Machine Learning Classification Results 191
7.21 HSV classification error (1 level) 193
7.22 HSV classification error (2 levels) 194

7.23 HSV classification error (5 levels) 194
7.24 HSV classification error (1 level) 194
7.25 HSV classification error (1 level) 195
7.26 HSV classification error (1 level) 195
7.27 Accuracy and consistency levels in experiments 1 to 5 196
7.28 Confidence levels in experiments 1 to 5 197

A.1 colour intensity matching method 234
A.2 Red eye telediagnosis system key stages 235
A.3 Conjunctivitis auto-evaluation system (Eye redness intensity classification)239
A.4 The participant descriptions of the colour in images 1 to 4 240
A.5 The participant descriptions of the colour in images 5 to 8 241
A.6 Section A of RGU Ethical Approval Form 252
A.7 Section B of RGU Ethical Approval Form 253
A.8 Section C and D of RGU Ethical Approval Form 254
A.9 Management Permission 255
A.10 Consent Form 257
A.11 NHS Consent Form 259
A.12 Colour Blind Test Part 1 260
A.13 Colour Blind Test Part 2 261
A.14 Colour Blind Test Part 3 262
A.15 Colour Blind Test Part 4 263
A.16 Normal red (None) 272
A.17 Conjunctivitis digital images-Trace red 273
A.18 Conjunctivitis digital images-Mild red 274
A.19 Conjunctivitis digital images-Moderate red 275
A.20 Conjunctivitis digital images-Severe red 276
A.21 Conjunctivitis digital images matching in the scale 277
A.22 Conjunctivitis digital images matching in the scale 278
A.23 Examples of image quality levels in the scale 278
A.24 Example of red eye online experiment 1 284
A.25 Example of red eye online experiment 2 285
A.26 Example of red eye online experiment 3 286
A.27 Example of red eye online experiment 4 287
A.28 Example of red eye online experiment 5 288

Chapter 1

Introduction

1.1 Research Background

In remote and offshore areas, people are struggling to have quick access to healthcare services due to a number of difficulties such as cost, communication, management, time and transportation. In recent years telehealth systems have provided a great opportunity for enabling easier access to services for both emergencies and follow up. Furthermore, telehealth technology is evolving rapidly and has enhanced its effectiveness and efficiency through advances in new technologies. One of the critical challenges facing telehealth specialists is the transmission of digital images which have similar qualities to the original, especially in terms of colour intensity, permitting a degree of clarity which will allow accurate diagnosis.

This study focused on the use of digital images in "store and forward" teleconsultation (SAF) within telehealth. There are two key critical challenges in SAF teleconsultation that may affect the accuracy of diagnosis when using digital images. Firstly, the quality of the image plays an important role in the diagnosis, which quality may decrease after the transmission to the medical professionals. Secondly, the differences in colour perception, which is a key cognitive process that occurs naturally, can also affect the diagnosis using the digital images.

This research aimed to investigate how human participants perceive the colour red when transmitted electronically via email, web link, real-time video links, or mobile phone. In certain medical pathologies, the colour red plays a major role in clinical diagnosis yet the physics of the colour makes accurate electronic transmission and representation difficult. Given the increased reliance of the National Health Service (NHS) on telehealth, this research aimed to investigate variations in the intensity of the colour red when represented and transmitted electronically in relation to how human participants perceive and use the colour in a medical context. It is hoped that the results might provide a degree of guidance to the NHS about telehealth technology and the perception of the colour red when used in diagnosis. The research was experimental in nature, with participants being presented with clinical images and asked to complete various tasks including describing, grouping, classifying and ranking digital medical images in order to determine the parameters of colour perception and cognition.

The relevance of this research to telehealth is to enhance the use of digital images to diagnose pathologies more accurately. Improving the diagnostic accuracy of digital images will reduce waiting lists, thereby reducing healthcare costs. Time-management for both healthcare workers and patients will also be improved, particularly for patients who live in remote areas, who have limited access to specialist doctors or healthcare centres, and for those with special needs or travel difficulties.

1.2 The Aims of the Study

The project focused specifically on the role of colour in digital images (i.e. the effect of image quality and colour characteristics) in human colour perception when diagnosing pathologies, in this case the dermatological conditions cellulitis or conjunctivitis.

Additionally, this study aimed to investigate human errors and cognitive limitations in relation to colour perception during the diagnosis of cellulitis and conjunctivitis. Using colour digital images to diagnose these medical conditions can be considered an example of complex safety-critical human computer interaction where cognitive overload, for example, could result in judgemental errors that may be detrimental to the wellbeing of the patient. In order to mitigate these problems, image quality and redness scaling systems are needed.

The main expected outcome of the study was to offer healthcare providers guidance about colour related issues in telehealth, in order to reduce subjectivity of image assessment and colour judgment in the current system, thereby enhancing objectivity and avoiding or minimising the possible errors associated with red colour in digital images. This study leads to greater understanding of human colour perception in the medical field and may prevent or avoid errors that are caused by the misperception of the colour red.

A further intention was that the results and the analysis of the thesis, including image and colour classification, image quality and confidence scale, would contribute to the formulation of a protocol in the development of a system for the auto-evaluation of pathology, which would simplify and speed up the assessment of medical conditions, prior to the involvement of a doctor. Although, the development of such a system would be a high complex and sophisticated project, and beyond the scope of this thesis, a simple prototype was generated and used alongside the main data analysis as a basic test of the concept. The results are briefly mentioned in the discussion section of the experiment results and the proposed specifications included in Appendix 1.

This study was a collaboration between the Cognitive Engineering Research Group at the Robert Gordon University, the Accident and Emergency (A&E) Department at Aberdeen Royal Infirmary (ARI) and the Scottish Centre for Telehealth recently incorporated into the National Health Service (NHS) 24.

Research over the past 20 years has made progress in understanding and controlling human errors in human-computer interaction (Byrne 2011; Latino 2007; Reason 2006) but technology continually develops which poses new challenges for researchers. All users of any system make errors and this is an accepted fact. The rationale for this research was that errors in telehealth can be critical; affecting the health of people when there is a misdiagnosis due to misperception of colour in the image. It is anticipated that when technology advances, the healthcare service will be enhanced. However, without careful understanding of the technology, unexpected new errors may occur. For example, when diagnosing pathologies in telehealth using digital images, the quality of the images needs to be established and confirmed before accepting them as basis for diagnosis. This study proposed an image quality scale to be used as a supportive tool when assessing image quality at a very early stage and before classifying colour as part of a diagnostic protocol.

Using high quality digital images (more specifically high resolution) in diagnosis is a recent development in telehealth that has come a long way through previous evaluation and experimental studies. However, there is still a need for more research in order to increase the confidence of decision makers in the healthcare field in using digital images in the diagnosis of a widening range of medical conditions.

This study investigated the possibility of the addition of two well-understood pathologies to store and forward (SAF) diagnosis in telehealth, by looking into the rate of errors that may occur when judging colours in images showing these two pathologies. These are cellulitis, a bacterial skin infection, and conjunctivitis (or red eye).

The colour red is one of the key symptoms of both of these conditions. During treatment, doctors measure the amount of redness in an infected area and monitor the size of the area as it changes over time, which are tracked by doctors during the diagnosis by measuring the size of the spreading area over time as well as measuring the amount of redness in the infected area. However, quantifying the image and other medical information goes through several stages during the diagnosis protocol in telehealth practice. It starts with the capturing and transmission of the image by the patient or the health care professional using a device such as mobile phone or digital camera. This information is then transferred over an internet connection to a specialist diagnosis centre, to be displayed and subsequently assessed by a medical specialist.

1.3 Research Problems

In the telehealth scenario the specialist doctor and patient communicate remotely using computer technology. In addition to medical history and patient data, doctors receive digital images of the clinical presentation of the patients. Accurate colour differentiation in digital images plays a vital role in diagnosis. Doctors may perceive the colours in these images differently due to a number of reasons, such as visual problems in relation to colour perception, or interference from lighting or reflection in the display or viewing area.

This research investigated various aspects of the colour red when displayed in a medical context. It introduced the use of an image quality scale to assess received images before using them for diagnosis. The study also used standard sample images to investigate the impact of using a colour scale on the accuracy of colour judgement. The study also investigated the relationship between the accuracy, the confidence and the time spent in classifying colours.

This study did not focus on medical diagnosis per se but rather the perception of the colour red when it is described, grouped, and ranked, since this can have direct impact on the final diagnosis. The study investigated the level of agreement between doctors and telehealth system operators (non-doctors) in their accuracy level by testing and scoring their descriptions, grouping, ranking and differentiation of colours in images of patients presenting with cellulitis and conjunctivitis.

The study group involved both medical doctors and non-medically qualified persons. Non-doctor participants were involved in the study to act as a control group in order to determine whether medical qualification and medical experience had any effect on colour perception.

1.4 The Scope of the Study

There were three key elements of this research: The first part related to cognitive knowledge. The study counted and classified the human errors of participating users of SAF systems. Any technical or technological errors were outside the scope of this study and are mentioned only in passing. The second part related to human digital image processing. The technical aspects of capturing, transmitting, representing and storing colour digitally were not included. The study only focused on how the colour red is perceived and interpreted in medical images when displayed. Furthermore, the study only focused on the colour red.

The third part related to telehealth. The study focused on the use of digital images during SAF teleconsultation as one of the common diagnostic methods in telehealth. The study did not investigate any pathologies other than cellulitis and conjunctivitis. However, the results may be generalised and applied to other medical conditions where colour perception is critical.

1.5 Research Objectives

The following were the main four objectives for the study:

1. Investigation of the impact of medical background on human colour perception in the context of SAF teleconsultations involving conjunctivitis and cellulitis. To test whether human colour perception is generic or subject to medical background through comparative analysis of the responses of the doctors and non-doctors comprising the study participants. This can create a foundation for greater involvement of non-doctors in telehealth. The study investigated three key issues in order to achieve this objective. The study tested colour perception by assessing the level of accuracy achieved in the performance of colour related tasks such as describing, grouping, ranking, and rating. The study compared the overall accuracy of the two participating groups of doctors and non-doctors and also investigated the level of agreement among the participants within their groups.

2. Investigation of the impact of using different scales of the colour red on the accuracy in SAF teleconsultation. The study investigated the impact of using a colour scale that showed different degrees of redness as a guide when diagnosing pathologies. The colour scale combined a numerical classification with descriptive text and standard sample images. An investigation was also made into the contribution of the standard sample images. These were removed from some experiment

runs and the results compared with similar experiments that had included the images. The study compared the effect on the accuracy of using a red colour scale of three divisions with that when using a similar scale with five divisions. Such a comparison could contribute to the development of existing SAF systems.

3. Investigation of the relationship between accuracy, confidence and time taken. The study investigated the relationship between the accuracy and the confidence of the participants in their performance during SAF teleconsultations for cellulitis and conjunctivitis cases. The study also investigated the relationship between the accuracy and the time spent by the participants in teleconsultation. It also introduced a confidence scale as a supportive indication and measurement tool for the accuracy of performance.

4. Introduction and testing of the use of image quality scales in SAF teleconsultations for cellulitis and conjunctivitis cases. The study investigated the impact of using an image quality scale on the participants? accuracy and confidence, and the time spent in performing colour related tasks in SAF teleconsultations. The study investigated how the image quality is rated by participants and determined the minimum acceptable quality that is needed to ensure that the colour, and other information, in the image is clear enough for the recognition and diagnosis of both cellulitis and conjunctivitis.

1.6 Research Questions

The following were the research questions in the study.

1. Are the participants consistent in their accuracy when judging the colour of cellulitis and conjunctivitis during SAF teleconsultations?

2. Is there a relationship between accuracy, confidence, and time taken when judging the colour of cellulitis and conjunctivitis during SAF teleconsultation?

3. Are there any differences between doctor and non-doctor participants in their accuracy, confidence and time taken when judging the colour of cellulitis and conjunctivitis during SAF teleconsultation?

4. Are there any differences between participants in their accuracy when judging the colour of cellulitis and conjunctivitis during SAF teleconsultation due to their medical experience?

5. Are there any differences between participants in their accuracy of colour judgment when using a numeric descriptive colour scale, both with and without standard images and also when using three or five divisions in the scale?
6. Are there any changes to the accuracy, confidence, and time taken when judging the colour of cellulitis and conjunctivitis during SAF teleconsultation due to changes in image quality?
7. Are there any differences between participants in their accuracy when judging the colour of cellulitis and conjunctivitis during SAF teleconsultation due to human perception?

1.7 Thesis Organisation

The rest of the thesis is organised as follows:

- Chapter 2 presents the literature review, where the key and relevant work carried out by others is reviewed and critically discussed.
- Chapter 3 describes the methodology of the research, where various related research approaches from previous studies are briefly mentioned and considered. The chapter also focuses on the selected approach used in carrying out the project and justifies the research design. The hypotheses of the study are also included.
- Chapter 4 presents the pilot study, which describes and presents the results of a set of experiments where non-medical images were used to describe, group, rank, rate, and differentiate between images showing the colour red. The pilot formed the basis of the structure for subsequent experiments and tested the methodology.
- Chapter 5 presents the skin infection experiments, where a series of five experiments that used medical images showing cellulitis (skin infection) were designed and performed. The experiments were critically evaluated and reported in detail.
- Chapter 6 Describes a repeat of an image matching experiment from chapter 5. This was performed under the same experimental conditions but with a larger number of participants and using an expanded range of test images. The intention was to investigate putative shortcomings in the design of the original experiment which had been indicated by analysis of the results.
- Chapter 7 is a red eye experimental study, where a series of seven experiments using medical images showing conjunctivitis are detailed and their results are presented and critically discussed. This series introduces the use of a red colour

scale in the diagnosis protocol of telehealth. Another additional section on ML was added to compare the results between human and ML Performance.

- Chapter 8 concludes the thesis by critically discussing the results in context, including the hypotheses. The chapter also presents the research recommendations, limitations and reflections.
- Appendices present the key supplementary figures, tables, and data that were generated by the study.

Chapter 2

Literature Review

2.1 Introduction

The literature review is organised into the following three main areas:

The first part focuses on telehealth systems and the current challenges of medical teleconsultation using digital images.

The second part of the literature review is about the cognitive background of the project, more specifically human errors and their classification. This section provides an explanation of measuring the level of confidence for users of a system as an indication of the accuracy of their performance, in this case diagnostic accuracy. The section ends with a brief review of human colour perception and factors that may affect the perception and resulting accuracy.

The third and last part of the review explores the colour red in medicine in general and the use of a colour scale in healthcare and telehealth, especially in supporting diagnosis of cellulitis and conjunctivitis. This section investigates extracting colour palettes from images and considers schemes for the description and classification of images. It concludes by elaborating on using image matching as an objective diagnostic method.

Because of the multidisciplinary nature of the project, each section of the literature review will be followed with a critical discussion where the problem will be defined and the integration between the different fields and their relationship to the current study will be explained.

2.2 Telehealth

The following section provides more details about telehealth systems, their methods, and applications used for this study.

2.2.1 The concept of telehealth

The term telehealth (or alternatively: telemedicine) refers to the use of information and telecommunication technologies in medicine in order to provide healthcare services remotely, without the need for a conventional face-to-face doctor to patient meeting. Telehealth includes the use of digital images and multimedia to provide healthcare services. Initially it involved consultative services only but now the concept has expanded considerably (Moore 1999; Romero, Garrido and Garcia-Arpa 2008).

The core idea of telehealth is to exchange medical data and information in order to provide a more effective and satisfactory healthcare service across geographic, social, cultural and time-based differences. Paone and Shevchik (2013) mentioned that currently telehealth is used for diagnosis, treatment, administration of therapies, and prescription.

Healthcare providers and organisations have become more interested in employing advanced communications technologies to support, enhance and expand their services. The availability and low cost of high-bandwidth internet connections and high-performance computers and preipheral devices, played significant roles in developing and spreading telehealth services. Such development made telehealth increasingly available and facilitated its adoption by healthcare providers and patients who currently demand more convenient healthcare services. According to Paone and Shevchik (2013), 80% (or 113 million adults) in the United States use the internet for healthcare information. They stated that patients, especially in remote and rural areas, expect personalised connectivity and communication with healthcare providers along with easy self-service access to their personal health information. This is also supported by Moumtzoglu and Kastania (2013). Kaplan and Litewka (2008) who stated that enhanced technology motivates people to demand more control over their lives and health by preferrring to be treated at home with minimal interaction with healthcare workers. In general, patients prefer more control over their privacy, health management, schedule and activities rather than being institutionalised. Craig and Patterson (2005) stated that telehealth is not a new branch of medicine; rather it is a new methodology by which healthcare is provided. It is widely used in advanced countries such as the USA, Australia and the UK, and there is increased interest in developing countries. They elaborated that telehealth can be classified based on the level of interaction between the patient and the medical experts. This can vary from real-time, two-way interactions to pre-recorded consultations. It can also be classified into the type of medical data that is transmitted, whether as text, audio, video or still images. Moreover, from the surveys, the authors have found that telehealth was most commonly deployed in support of emergency care for remote areas or to generally improve existing service offerings. In other words, telehealth provides opportunities for individuals to improve their health by easing communication with healthcare providers, from whom they are distant and whom they may not have met personally.

2.2.2 The benefits of telehealth

Wallace et al. (2012) explored how telehealth could increase the efficiency of healthcare services and provide a better quality of clinical care, especially through developing good electronic patient records which in turn lead to better communication.

Kaplan and Litewka (2008) added that telehealth provides an opportunity for an integrated system that links patient personal data, allows ease of access to medical history and reduces the need for any further healthcare services. Such integration helps potential information overload to be more organised, managed and retrieved.

In fact telehealth is not only a technological improvement, but a re-engineering of the entire approach to healthcare provision, which requires consideration of socio-technical aspects in the design and development of the telehealth system itself. Telehealth can expand the scope of healthcare provision, by placing patients at the heart of services and providing more interaction in different areas with the healthcare professionals (Moumtzoglu and Kastania 2013; Kaplan and Litewka 2008; Struber 2004). Dheer and Chaturvedi (2005) concluded that patients in rural areas have been the main beneficiaries of telemedicine healthcare delivery.

The need for telehealth has increased rapidly. According to Stowe and Harding (2010) the growth of people aged 85 years and over will quadruple to 5% of UK population by 2033, which will have great impact and consequences on NHS. Telehealth could become a remedial option to this problem if it allows easy access to health care in an efficient and cost effective way. Hall (2011) found that using telehealth services in the UK reduces 80% of home visits by NHS doctors and creates a significant fall of 50% in hospital admission, which in turn allows specialist healthcare professionals to apportion more of their time to the most needy patients. Wallace et al. (2012) added that telehealth also saves travel time and costs for both patients and doctors.

According to Joseph et al. (2011), the Audit Commission estimated that the use of telehealth could potentially reduce the number of bed-days in hospitals in England for certain types of cases. The Commission's estimates amounted to savings of 50 million in cases of chronic obstructive pulmonary disease and 118 million in cases of congestive heart failure.

The benefits of telehealth were highlighted in many studies such as Paone and Shevchik (2013), Kreps and Neuhauser (2010) and Wootton et al. (2006). In addition Dheer and Chaturvedi (2005) emphasised the benefits of telehealth in reducing differences in healthcare accessibility, especially in access to medical specialists for diagnosis and consultation. Furthermore, they stated that telehealth provides a unique opportunity for medical training, education and promotion, allowing healthcare professionals in medical schools and hospitals to exchange knowledge and experiences (Moumtzoglu and Kastania 2013). Also, new practitioners and medical students can observe medical procedures at a distance. Dheer and Chaturvedi (2005) added that telehealth provides great opportunities for medical doctors to get advice and supervision from a medical specialist or super-specialist.

2.2.3 Challenges facing telehealth

Even though the current evidence on telehealth suggests that it can change the mind-set of people, can improve health care services and can save time and resources, it faces a number of challenges that can affect the establishment and implementation of telehealth systems. The key challenges can be attributed to patients, medical specialists, IT systems and communication links.

The main challenges attributed to patients are their resistance to accepting this method and unfamiliarity with the concept of telehealth and the technology it utilises (Pappas and Seale 2009). Increased awareness and training could reassure patients and better motivate them to use a telehealth system.

Challenges attributed to medical specialists include resistance by some healthcare professionals to accept the concept, (Pappas and Seale 2009); the expense and management of training (Joseph et al. 2011; Pappas and Seale 2009); ethical and legal issues such as reliability and confidentiality of data, especially when the telehealth link is established between countries (Saliba et al. 2012; Kluge 2011; Kaplan and Litewka 2008; Stowe and Harding 2010); and lack of standard reference methodology (Ritcher et al. 2009).

Challenges attributed to system and communication links include the initial cost for building the system infrastructure (Saliba et al. 2012; Joseph et al. 2011; Lee et al. 2000; Brear 2006; Miller 2007) and the compatibility and appropriateness of technologies for particular healthcare requirements (Joseph et al. 2011; Lee et al. 2000; Moumtzoglou and Kastania 2013; Obstfelder and Engeseth and Wynn 2007); poor image quality (Ritcher et al. 2009; Moumtzoglou 2013; Sikka et al. 2012); unsustainable technology (Moumtzoglou and Kastania 2013; Saliba et al. 2012); and a perceived lack of long term strategy and leadership (Miller 2007; Currell et al. 2000).

2.2.4 Telehealth diagnostic methods

The following section gives a brief review of some key studies in teleconsultation, teledermatology and red eye diagnosis, and provides a critical review of these studies in relation to the use of digital images in tele-diagnosis, which is the focus of this research.

There are three types of consultations in telehealth. The first method is a real-time teleconsultation, where a patient supported by a nurse or GP and the medical consultant, or specialist doctors in the hospital, interact together at the same time in an interview, via a videoconference. This gives the patient the opportunity to ask questions and allows input from the nurse and GP. This might involve such activities as passing case-related information to the specialist or recording their clinical observations in real time. The second method is currently the most commonly encountered strategy for providing telehealth and is termed store and forward teleconsultation (SAF). During an SAF consultation both parties interact and exchange medical information and images over a telecommunications connection but not while the parties are in direct, real-time communication. The data is stored online for later retrieval by a medical specialist who will provide medical advice and diagnostic reports as feedback to the sender. The third teleconsultation method is a hybrid of the real-time and store and forward consultation methods. Medical data and images are first sent to the specialists by email or posted to a website, as in an SAF consultation. The specialists can then retrieve this stored information in preparation for a subsequent real-time consultation.

Real-time teleconsultation

Real-time teleconsultation (or alternatively: synchronous telehealth) is an interactive telehealth process that is carried out in real time. This can be used in emergencies and when GPs need specialist advice from an expert who is situated in a remote location, usually in a main hospital. In the case of Aberdeen Royal Infirmary (ARI), there are fourteen community hospitals outside the Aberdeen city which are linked to a teleconsultation room at the centrally located infirmary. Teleconsultations are provided both routinely and in response to emergencies when any of the fourteen satellite hospitals need rapid access to expert medical advice. An integrated booking system assists with the scheduling of patient appointments.

Many studies such as Whitten (2003), Campos et al. (2012), and Lopez et al. (2011) explored and underlined the benefits of real-time teleconsultation. Whitten (2003) evaluated a real-time tele-dermatology application that used low cost microwave links and an interactive TV system, to provide diagnosis and improve healthcare services for patients in geographically remote areas.

Lopez et al. (2011) found, after 281 cases of real time teleconsultations, that this method of providing healthcare and diagnosis for people in rural communities is efficient, cost-effective and satisfactory for users of the system. They found that 80% of the participants were satisfied or very satisfied with the teleconsultation services that they received. 63% of them said that they would use the service again, while 65% said that it improved their healthcare, saved waiting and transportation times and avoided the expense and inconvenience associated with commuting. Campos et al. (2012) found a similar response with a slightly higher rate of patient satisfaction. They conducted an experiment of 259 cases of teleconsultations over a period of twelve months, between a primary care hospital and a secondary care hospital. According to them, 99% of their participants were positive about the telehealth system used and mentioned financial savings, reduced waiting time, elimination of the need to travel, and quieter access to specialist advice.

The high satisfaction rate shown by Campos et al. (2012) has demonstrated the acceptance of their newly proposed telemedicine solution for remote video-electroencephalography consultation (tele-EEG), this study was specific to the newly proposed solution (shown in figure 2.1). This is in contrast to a telephone survey carried out by Lopez et al. (2011) that evaluated the general acceptance and satisfaction of patients using an existing teleconsultation service that they were already familiar with. All the subjects surveyed had previously used a real-time telemedicine consultation system that used a video camera. It was also noted that both sample sizes of the two studies were within the same range.

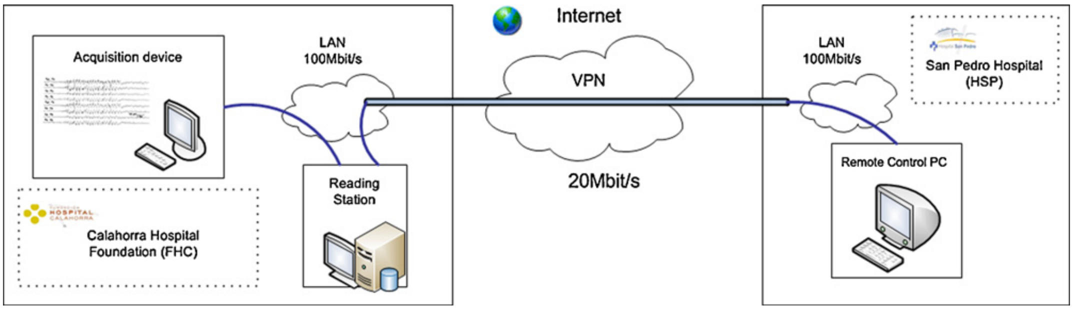


Figure 2.1: llustrates that patient and specialist can communicate securely between physically separated locations using telecommunication technology Campos et al. (2012)

Figure 2.2 shows part of a teleconsultation room in ARI where specialists can communicate with patients and GPs to provide medical advice. The equipment shown here includes a suite of audio-visual recording and display devices and is fully internet connected in order to send and receive medical data and images. The remotely situated medical booth of a patient, GP or nurse has similar teleconsultation equipment. The current system is standardised at both ends and it allows direct communication and reporting of medical data between the GP or nurse and the specialist.

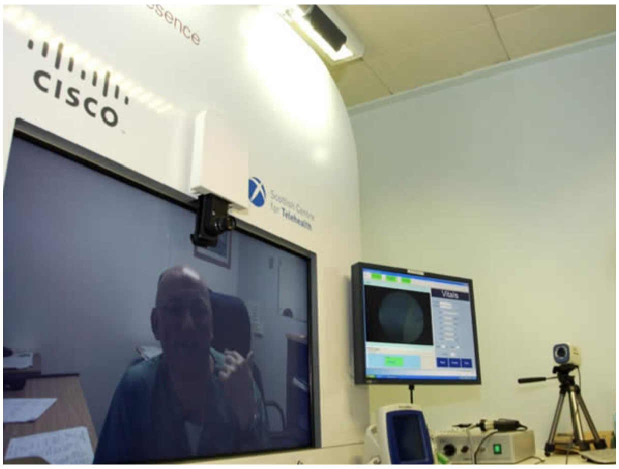


Figure 2.2: Medical booth in ARI Scottish Centre for Telehealth

In the medical booth, a high definition screen shows a live picture of the specialist to the patient and nurse and all parties are in constant two-way voice communication. A digital camera can record still or video images and transfer them directly to the specialist's computer or to their display screen. A range of medical diagnostic devices, such as blood test machines or heart-rate monitors, are also network linked and can similarly transfer data directly. Figure 2.3 Shows a live teleconsultation, observed by the researcher, between a specialist doctor in a teleconsultation room in the A&E department at ARI in Aberdeen, and a patient and nurse in Banff community hospital, which is 60 miles away from the ARI.

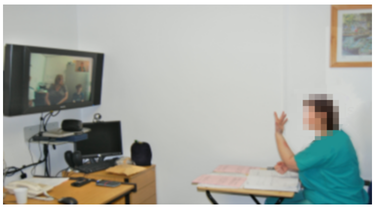


Figure 2.3: Interaction in teleconsultation in ARI-taken by researcher

Using the facilities provided by the system, the specialist was able to interview the patient directly, review their recorded medical history, hear an examination report read out by the supporting nurse and inspect medical data that was gathered from the patient at the remote location and was transmitted during the consultation over the teleconsultation systemś network link. This allowed the specialist to quickly diagnose buckle fracture in the left distal radius. Figure 2.4 further illustrates the specialist doctor in the ARI teleconsultation room reading and providing interpretation of an X-ray related to the case.

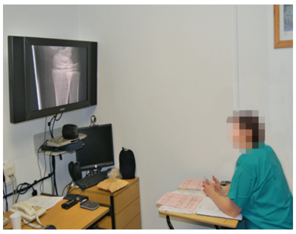


Figure 2.4: X-ray transmission in ARI-taken by researcher

At the end of this consultation, the specialist doctor gave medical advice to the patient and instruction to the nurse, and booked the next appointment for a follow up teleconsultation.

Store and forward teleconsultation (SAF)

Store and forward teleconsultation (or alternatively: asynchronous telehealth) is a very frequently applied method in telehealth. The following section provides an outline of the key benefits and challenges of SAF, then reviews a number of studies on the validity and reliability of this method in different diagnostic scenarios.

Studies have shown that the use of digital images in diagnosing different types of skin pathologies can be a useful alternative to face-to-face consultation. For example Buckley, Andelson and Agazio (2009) used this method in nurse-to-nurse consultation on wound treatment; Kim et al. (2003) to assess the states of chronic wounds; Tsai et al. (2004) in diagnosing wounds such as gangrenous necrosis and cellulitis; and Moreno-Ramirez et al. (2007) used it in diagnosing skin cancer. SAF teleconsultation has a number of benefits in comparison to real time teleconsultation. These benefits have been detailed in many research studies. Among these benefits are:

- Cost effectiveness (Massone and Schettini 2012; Whited 2010;Oikonomou, Gwynedd and Trust 2009;Whited et al. 1999; Romero,Garrido and Arpa 2008);
- Economies of time (Warshaw, et al. 2011;Oikonomou 2009; Eedy and Wootton 2001)
- An improvement in the standard of care provided and enhancement of accessibility to healthcare in rural areas (Armstrong et al. 2012a; Fabbrocini et al. 2011; Pathipati and Armstrong 2011; Shannon and Bucker 2010; Eedy and Wootton 2001).

The SAF teleconsultation method is simple and easy to use (Hsueh et al. 2012; Rotger et al 2011; Deshpande et al. 2009; Romero, Garrido and Arpa 2008). It provides training and educational opportunities (Massone and Schettini 2012; Pathipati and Armstrong 2011), and it has also been shown to be reliable and accurate in clinical outlets (Colven et al. 2011; Lasierra et al. 2012: Rios-Yuil 2012). SAF teleconsultation provides better satisfaction and better quality care, and can contribute to a patient centred healthcare model (Witkamp 2009; Romero, Cortina and Vera 2008). It is also fast and convenient with good user satisfaction in relation to face-to-face interaction (Hsueh et al. 2012).

There are many challenges to using store and forward teleconsultations. A few of these challenges are: the initial cost (Gomez et al 2001); legal requirements when applied across different countries (Phillips et al. 2002); acceptance hampered by rigid decision making structures, slow adoption processes and concern for its consequences (Witkamp 2009); lack of patient contact with the dermatologist and delay in obtaining the diagnosis and advice on management of the treatment (Eedy and Wootton 2001). Store and forward teleconsultation, even though it is widely used, requires considerable development and enhancement to fulfil the purpose of the healthcare system.

In the area of tele-dermatology, a recent study in Spain by Orruno et al. (2011) showed that most medical doctors accepted the idea of SAF teleconsultation. However, they reported that some healthcare providers are still reluctant. Rios-Yuil (2012) investigated the use of tele-dermatology with complicated skin diseases in Panama and discovered a high degree of correlation between SAF and more conventional face-to-face methods.

Colven et al. (2011) reported a very strong positive Spearman rank-order correlation in a comparison of face-to-face diagnosis with SAF teleconsultation. They found that the validity and reliability of both methods is similar in terms of accuracy of clinical diagnosis. They also carried out a trend analysis, which demonstrated that the level of agreement increased in direct proportion to the number of samples.

There were also strong results from a study by Smith et al. (2013), which showed how dermatologists, general practitioners and experienced nurses could use digital images in a low cost teleconsultation setup to diagnose simple skin conditions, such as dermatophytic fungal infections (tinea).

A study of healthcare provision in rural areas of Mexico by Romero et al. (2008) showed that simplicity, ease of use, and avoidance of transportation costs were primary reasons for the regular use of SAF teleconsultations between centralised general hospitals and satellite healthcare clinics. In this study, the popular social media site Facebook.com was used to store patients' medical data in order to facilitate telehealth consultations in cases of simple and commonly encountered skin conditions. The same study showed that 75% of patients did not need referral to the general hospital and benefited from the service. They mentioned that the quality of the images captured was adequately suitable for diagnosing common conditions and that only 25% of the patients were asked to see doctors or required to be be re-assessed due to unusual conditions that required deeper investigation.

A study by Armstrong et al. (2012b) focusing on store and forward tele-dermatology in the United States stated that it is a better solution than the traditional face-face method. Healthcare providers of tele-dermatology prefer SAF because of its cost advantages, its enhanced patient throughput, its ease of use, its value for training purposes and its ability to overcome time and distance barriers.

According to Wurm, Campbell and Soyer (2008), real-time teleconsultation provides very satisfactory interaction but it is time-consuming, expensive and requires the availability of all participants at the same time. Whilst the SAF method does not provide interactivity, it is cheaper and easier to perform and the data that it is capable of acquiring has consistently been found to be eminently suitable for investigatory and diagnostic applications.

Considering all of the above opinions, it is clear that SAF provides the most cost effective solution for reducing demands on the time and resources of both patients and health services. It is apparent, therefore, that maximising the effective use of such systems would be a pragmatically worthwhile topic of investigation. This must include the removal of errors of judgment caused by poor capture and transmission of images, the absence of a common language and poor environmental conditions.

2.2.5 Use of Digital Images in Diagnosis

The following section analyses previous studies that have used digital images for diagnostic purposes.

Buckley, Adelson and Agazio (2009) conducted a comparative study to analyse the effect digital images had on remote nurse-to-nurse consultations. In this study, home care nurses assessed 43 adult patients with a total of 89 wounds and completed a wound assessment and recommendation form for each patient. These nurses then submitted a verbal report to the senior wound, ostomy and continence (WOC) nurse who also completed a wound assessment and recommendation form based on the verbal report. The WOC nurse then accessed a variety of digital images of the wounds and, based on their study of these, modified their initial assessment form and recorded the reasoning behind their amendments. The study compared the assessment forms filled out by the home care nurse with those of the WOC nurse based on only the verbal report, and compared the WOC nurse's assessment based on only the verbal report with the new assessment based on both verbal and image based information. The clinically relevant disagreement between both the WOC nurse's assessments was 58% (52/89) with only 29% (26/89) agreement overall. These results show that digital image based visualisation is essential for a maximally accurate assessment. The digital images extended the view of the nurse and added some environmental factors and extra depth to the verbal report. Buckley, Andelson and Agazio (2009) found that taking different types of images for the same wound (such as close up, general, distant and both extremities) provided details and access to deep information that could not be captured in a single image. All these findings are very useful, especially that related to capturing images of the same area from different angles, since these offered a clearer, more detailed, picture that could assist considerably in investigating and diagnosing a pathology.

The purpose of the current study mirrored that of this study as they both evaluated the impact of digital images on medical assessment. The results of both studies were consistent with the conclusion that the use of digital images increased acuity of clinical assessment. However, whereas their study used a one-to-one analysis between cases and healthcare professionals, the current study implemented a many-to-many comparison using a larger sample size in order to increase the reliability of the study. Furthermore, capturing images over a period of time is also important to track the development of a pathology. It is also recommended to keep all these images numbered or coded with dates, and compare them with images that were taken previously, to monitor the degree of severity and the rate of disease progression. Massone et al. (2009) reported using a mobile phone to capture and transmit digital images as part of a telemedical system that utilised internet, wireless and satellite communications technlologies. Recent similar studies conducted by Lin (2012) and Wurm and Soyer (2012) used similar methods. The quality of new generation mobile image capturing equipment is high and quite acceptable. Image processing software and hardware has also improved in parallel with camera and image capture technology. The study mentioned here was related to cancer, but has potential application to other diseases. A concordance of 91% between face-to-face diagnosis and remote diagnosis has been reported by the same researchers for tele-dermoscopy, with transmission of images via email or dedicated web application. In another study by the same authors 83% of diagnostic accuracy was achieved for melanocytic neoplasms, in comparison with diagnoses based on conventional histopathology.

Another recent study by Tran et al. (2011) found that dermatology patients (30 cases were studied) were able to send images of sufficient quality for investigative and diagnostic purposes using mobile phones, in their case with 5-megapixel cameras, The study discovered a 75% agreement mean with face-to-face consultation of the same thirty cases. Biemans et al. (2005) carried out a similar study and detailed the benefits of using a mobile phone application called WoundLog by nurses during follow up home visits to make reports and to capture and transmit digital images to dermatologists. WoundLog was also used to transmit feedback to the nurses..

All the above studies by Lin (2012), Wurm and Soyer (2012), Tran et al. (2011), Biemans et al. (2005), and Massone et al. (2009) supported the idea of using store and forward teleconsultation for follow-up cases after the initial face-to-face diagnosis. This idea was also recommended by Kanthraj (2009), especially in teledermatology. A study of skin tumours by Piccolo et al. (1999) concluded that store and forward teleconsultation of clinical and dermoscopy images provided similar degrees of diagnostic accuracy as face-to-face diagnosis.

In another study by Bergman et al. (2009), it was found that 89% of images (out of 300 samples) could be used for successful diagnosis. The quality of the images taken by the staff was higher than those taken by the patients. However, when the photographs taken by the staff were of low quality, then the patients themselves were asked to retake pictures for evaluation.

A study by Moreno-Ramirez et al. (2009) evaluated the store and forward method, used between a university hospital and 12 primary health care centres in Spain, for the diagnosis and management of skin cancer. This evaluated the effectiveness, accuracy and validity of this method using the following criteria as a standard for the evaluation: diagnostic confidence, quality of images used, clinical information, diagnostic category and management decision. The outcome of this evaluation revealed that the store and forward method had a higher percentage of usage and diagnostic confidence levels than the face-to-face method. Moreno-Ramirez et al. (2009) conducted a survey on the economic consequences of store and forward teleconsultation when used for diagnosing skin cancer. They evaluated the clinical effectiveness, accuracy, reliability and validity of this method by comparing the fixed and variable costs of this method with the conventional method. They concluded that the store and forward method is effective and cheaper in the management of skin cancer.

Edison et al. (2008) conducted an interesting study that compared in-person, real-time and store and forward teleconsultations with respect to diagnostic accuracy and confidence. They used a confidence scale from one to five to test 110 randomly chosen patients and 4 dermatologists. The study shows high agreement between the three methods in diagnostic accuracy and confidence but the in-person method was higher than the other two methods in both accuracy and confidence. Live and store and forward did not differ statistically in the diagnostic accuracy results, but the live teleconsultation showed higher confidence than the store and forward method.

Hypothesis number 3 of this thesis considers the relationship between accuracy and confidence (see chapter 3, section 3.7). Another study by Kim et al. (2003) evaluated the clinical accuracy of store and forward (SAF) methods in assessing the states of chronic wounds, including cellulitis or infection, by using digital images, patient information and wound data that was collected and sent by a nurse who used a laptop and transmitted the images via the internet to a database.

A set-up protocol on a website was established by Tsai et al. (2007) where all participants could have access to check the protocol and submit their images. Both study staff and subjects were trained to capture images using a digital camera. A similar process was done by Krupinski et al. (1999). The photos were taken for all patients and then evaluated as computer images by a panel of dermatologists, who concluded that good quality images compared very favourably with face-to-face consultation as a diagnostic method. Wurm, Campbell and Soyer (2008) and Moreno-Ramirez et al. (2007) both concluded in favour of the effectiveness of using SAF in diagnostic teleconsultation as an alternative to face-to-face diagnoses.

Moreno-Ramirez et al. (2007) further suggested a conclusion that SAF teleconsultation is an efficient, accurate, reliable and valid approach in diagnosing skin cancer and in pigmented lesion clinics. Hersh et al. (2006) reviewed the literature in relation to medical diagnosis and treatment, focusing on three types of services, SAF, home based, and hospital based. The study reviewed 106 references between 2000 and 2004.

The study found that the evidence for the efficiency of these three methods is mixed, showing the benefit of home based services, and suggesting that using teleconsultation such as video conferencing for diagnosis and treatment is the most effective for verbal interactions.

There are limitations found in Buckley, Andelson and Agazio (2009) in that they used only one nurse to view and judge the images. This may be insufficient because there is a risk of bias. It is very difficult to know how representative an individual nurse was in terms of skills, approach and style of working. Training in the operation of digital cameras could improve the quality of the images captured. The diversity of the wounds under examination, ranging from venous stasis ulcers to surgical and traumatic wounds, made it difficult to generalise the results of the study. Using different forms of communication that each encoded different patient information made the experiment lack consistency and this might also have affected the results.

Kim et al. (2003) found that the clinical agreement between telemedicine methods was not high but statistically significantly higher agreement was found in face-to-face consultation. They took face-to-face results as a reference standard for the accuracy of diagnosis and they called it the “gold standard” (or true diagnosis). In this study, it was noted that there was some disagreement between the face-to-face and telemedicine assessments, which led them to think that it may be due to differences in the judgment of physicians. The physicians did indeed agree that they used different criteria to respond to diagnostic questions in the face-to-face consultation, which means that the diagnosis was more subjective.

One of the confusions in Kim et al. (2003) is in the use of different images for the diagnoses by telemedicine and for the diagnoses made face-to-face. This makes it hard to assess how the quality of the digital images (for example in relation to their sharpness and fidelity of colour reproduction) correlated with the accuracy of the diagnostic results.

Wurm, Campbell and Soyer (2008) argued that video conferencing is a good alternative to face-to-face evaluation because it offers the participants the opportunity to interact and ask and answer question directly. They stated that the time required and the cost associated with the face-to-face method is higher than that with SAF methods. They referred to the year 1995 when SAF was first introduced by Perednia and Brown in Oregon (USA). The study also states that the absense of a standard terminology created confusing variation in making dermatological diagnoses, especially when disease categories and classifications overlapped. The use of a common language was examined within the thesis.

According to Wurm, Campbell and Soyer (2008), most of the tele-dermatology studies used face-to-face diagnosis as the reference standard and they investigated the diagnostic agreement between tele-diagnosis and face-to-face diagnosis. Agreement was found to be between 54% and 95% (mean 75%). One significant limitation to this method was highlighted by the same researcher, namely the unfounded assumption that face-to-face diagnosis is always correct.

A study by Piccolo et al. (1999) found a concordance between face-to-face diagnosis and tele-diagnosis of 91%. However, the number of correct (100% match) tele-diagnoses was lower than with face-to-face methods but the results were very similar and, most importantly, the difference was not statistically significant. The research found that there was a similar degree of diagnostic accuracy between face-to-face consultations and teleconsultations using images of patients presenting with skin conditions sent via email. Another interesting result was that JPEG compression, despite being a lossy format, still provided sufficiently high quality images, and also permitted exchange over the telecommunications network at high speed. The quality of the images and the fidelity of their colour representation, in the current study, contributed significantly to the ability to use the images for the diagnosis of cellulitis and conjunctivitis.

Krupinski et al. (1999) also found a high concordance between both methods with 62%, and 83% decision confidence for "definite" and "very definite". Good image sharpness was rated with 93% confidence. However, the viewing time varied from 3 to 167 seconds, which did not correlate in any way to image qualities, such as colour and sharpness. The study concluded that digital photography for SAF provided sufficiently high quality images, and diagnostic concordance rates using such images compared favourably with face-to-face methods. These results supported the results of Kim et al. (2003); Buckley, Andelson and Agazio (2009). However, Warshaw et al. (2009) contradicted them in their experiment comparing the diagnostic accuracy of non-pigmented neoplasm through face-to-face consultations with that of SAF. That study concluded that there was no significant difference between the two methods.

Tsai et al. (2004) found that 3 surgeons performing remote diagnoses of cellulitis cases were in 74% agreement. This implies 26% disagreement, which can be regarded as a form of error. Additionally, the study found that diagnosis was greatly aided by the concurrent online availability of patient case histories and observational data. One of the criticisms of the Tsai et al. (2004) study was their technical assesment of the photographic quality (with regard, for example, to the framing, lighting, and exposure) without regard to the clinical content, which is a very important factor in image assessment. One example of this would be the type of digital compression used, which can affect the clarity of the clinical details in the image.

Most of the previous studies such as Buckley, Andelson and Agazio (2009); Tsai et al. (2007); Wurm, Campbell and Soyer (2008); Moreno-Remirez et al. (2007); and Hersh et al. (2006) focused on using digital imaging in store and forward (SAF) consultation sessions even though there was very poor interaction between the two ends of teleconsultation. They dealt with this challenge by using telephone, email, and web-based conversation to communicate medical and patient information. This could be because real time teleconsultation is more expensive, is technically more complex, is more time-consuming, and requires well trained staff. Most of the previous studies, such as Buckley and Tsai et al. (2007); Wurm, Campbell and Soyer (2008); and Moreno-Remirez et al. (2007), assumed that face-to-face diagnosis was always correct and they took it as a standard control for the accuracy of diagnoses when evaluating other methods such as store and forward SAF. They called this standard a "gold reference" or "true standard" without investigating the reasons behind the disagreement in diagnosis between the dermatologists. Kim et al. (2003) believed that such disagreement could have been due to a a system fault, or originated in individual differences in colour perception. The current study took this limitation into account and dealt with it in two ways: firstly by using expert diagnostic experience to provide accurate results, and secondly by using a colour chart or scale, designed specially for this study and intended for use by dermatologists when diagnosing cellulitis. Some studies such as Kim et al. (2003) showed that the terminology used in describing images is a very important factor because using different terms can easily cause confusion and end with discrepant results from dermatologists with the same levels of medical knowledge and experience. The present study proposed to avoid such confusion, and its concomitant errors, by designing a colour chart as a common reference tool for standardised communication between dermatologists.

2.2.6 Telehealth applications

The study investigated cellulitis and conjunctivitis as particular examples of pathologies where the colour red plays a significant role in their detection and diagnosis. Redness (erythema) of the affected tissues is one of the symptoms which will especially alert doctors to the presence of these conditions.

Cellulitis medical brief

Cellulitis is one of many diseases of the skin. It is defined as a bacterial infection caused by Group A Streptococcus and/or Staphylococcus. It can infect any area of both the lower and the upper parts of the human body such as legs, face, breast, feet, hands, torso, neck and buttocks (Gunderson 2011; Baddour 2000; Fleisher and Ludwig 1980).

Cellulitis appears on the skin as poorly demarcated red, hot, swollen and tender areas of the skin. These can be as a result of breaks in the skin due to surgical wounds or traumatic injury, or may alternatively be caused by ulcers, eczema, psoriasis, tinea infection, hypodermic injection or even from human or animal bites. All these features of cellulitis appear as an opening of the skin (Kroshinsky, Grossman and Fox 2007; Stulberg, Penrod and Blatny 2002; Baddour 2000; Fleisher and Ludwig 1980).

Cellulitis has a wide range of symptoms and patients with this pathology can show some or all of: warmness, poorly demarcated areas of redness, tenderness, swelling and pain.. The patient may also present with associated health conditions such as malaise, fever, chills, cough, headache, pruritus (itching), arthritis and diarrhoea (Figtree et al. 2010; Kroshinsky , Grossman and Fox 2007; Baddour 2000; Torok 2009).

Patients with a medical history of diabetes mellitus, leg ulcers, lymphedema, varicose veins, athletes foot, trauma, lymphatic compromise and alcohol abuse are more likely to develop cellulitis than others (Kroshinsky, Grossman and Fox 2007; Stulberg, Penrod and Blatny 2002; Baddour 2000).

According to Kroshinsky, Grossman and Fox (2007); Stulberg, Penrod and Blanty (2002); Baddour (2000); and Fleisher (1980), doctors initially diagnose cellulitis by conducting a full medical assessment, which includes a physical examination and patient's medical history. The location of the cellulitis on the body is noted, symptoms (such as redness, swelling and tenderness) are recorded and affected skin areas are measured. The affected areas may even be circumscribed with a marker pen in order to monitor the spreading of redness over a period of time. The origin of any visible skin punctures may also be sought.

The above studies also stated that cellulitis may be diagnosed further by taking a sample of the patient's blood or by culturing swab samples taken from an affected area to test for the presence of gram positive streptococcus and staphylococcus bacteria. They added that cellulitis, when discovered early, can usually be treated simply with systemic (usually oral) antibiotics. More serious infections may require parenteral (intravenous) antibiotics, surgical drainage or debridement.

Hedrick (2003) stated that erythema (redness), warmness and tenderness are the most common and key symptoms for the initial diagnosis of cellulitis. The rationale for using cellulitis as an example in the present study is predominantly due to the readily observable and diagnostic nature of these symptoms. These signs are usually subjectively diagnosed by general practitioners or by other medical staff in hospitals.

Cellulitis is not a life threatening pathology in most cases but its reoccurrence can be a drain on the health resources of a country (Gunderson 2011; Figtree 2010; Kroshinky 2007; Stulberg 2002; Fleisher 1980; Baddour 2000; Torok 2009; Kroshinsky, Grossman and Fox 2007). It was reported by Gunderson (2011) that in the USA alone there are over 600,000 cases of hospitalisation and over 9 million cases of patient visits due to cellulitis. This huge number of cases cannot be ignored, hence the focus of this study on an investigation of the potential application of telehealth in classifying degrees of redness as an aid to diagnosis and with a view to improving the provision of healthcare services.

The study investigated the role of the colour red, and its perception by doctors in telehealth, more specifically how they described and classified the colour red during their diagnosis of cellulitis and conjunctivitis. The study tested the importance of using a colour scale during the classification process and the level of accuracy in judging degrees of redness intensity that was possible when using the colour scale. Later chapters will provide more details on the experimental work that was conducted. The following images in figure 2.5 are typical examples of mild and severe cellulitis showing the varying degrees of colour in its development from mild to severe.

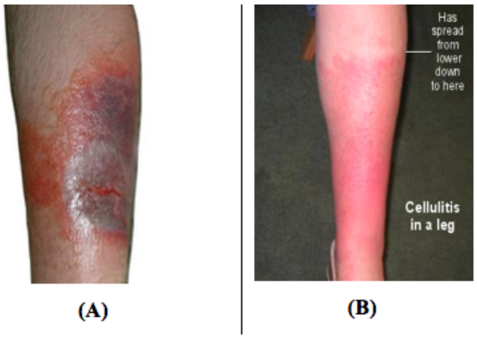


Figure 2.5: Image (A); typical severe cellulitis from clinical resource efficiency support team (CREST) 2005. Image B; typical mild cellulitis from Newson (2015

Conjunctivitis medical brief

This short brief describes conjunctivitis, its key symptoms and classification. Conjunctivitis is a generic term for any inflammation of the conjunctiva of the eye and is seen in a broad spectrum of conditions (Tarabishy and Jeng 2008; Hingorani et al. 2006). D'Aversa et al. (1995) defined conjunctivitis as infection or inflammation of the mucous membrane that covers the anterior segment of the eye. Conjunctivitis, or ”red eye”, results from vasodilation of the blood vessels in the white part (sclera) of the eye (Murphy et al. 2006). A healthy eye that does not have vasodilation is called a ”white eye”. However, it should be mentioned here that, ”red eye” refers to a broad range of different diseases, conjunctivitis being only one of them (Chodosh, Chintakuntlawar and

Robinson 2008; Shapiro and Croasdale 1997). The image in figure 2.6 shows the bulbar conjunctiva white area: the focus of this study, Available from Anatomy atlases. Anatomy of first aid: A case study approach (Bergman, 2015). The most common causes of red eye, among others, include dry eyes, conjunctivitis, keratitis, iritis, acute glaucoma, sub-conjunctival haematoma, foreign bodies, corneal abrasion and blunt or penetrating trauma. Although there are various causes of red eye, conjunctivitis is very common and it can be classified as bacterial or viral (Joseph 2004; D'Aversa et al. 1995). Redness, in conjunctivitis, makes it an important consideration for this research and eminently suitable for use as a case study.

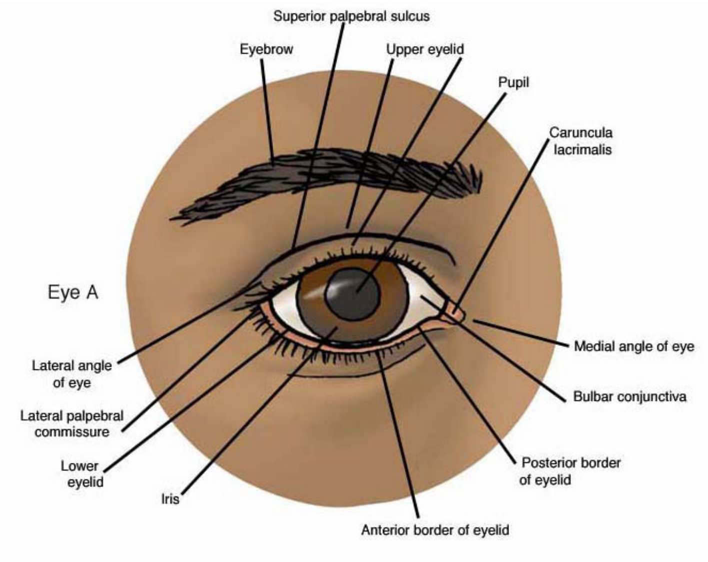


Figure 2.6: The external structure of the eye showing the bulbar conjunctiva white area (Source: Bergman, 2015)

There are three types of conjunctivitis; infective conjunctivitis, conjunctivitis, and non-infectious (allergic) conjunctivitis (Tarabishy and Jeng 2008; Hingorani et al. 2006). Symptoms common to different conjunctivitis pathologies include: watering eyes, acute red eye with sticky discharge, soreness, trauma, scarring, unusual pigmentation, scaling, oil content and texture (Wirbelauer 2006; Shapiro and Croasdale 1997). Additionally, Murphy et al. (2006) identified ocular disease, irritation, inflammation and chemical injury as possible causes of conjunctivitis.

Infective conjunctivitis is often due to viral or bacterial infections. The common symptoms of this variation of conjunctivitis are red and watering eyes and often a sticky coating on the eyelashes especially when waking in the morning. Infective conjunctivitis is reasonably common and is responsible for approximately 35% of eye-related problems noted in GP surgeries the UK. There are 13-14 cases per 1,000 people every year.

Infective conjunctivitis, specifically bacterial, is a common condition in both children and adults and presents with a red eye (Tarabishy and Jeng 2008). Although in most instances bacterial conjunctivitis is self-limiting, antibiotics can aid in accelerating resolution and reducing the risks of further complications. It is essential to differentiate between bacterial conjunctivitis and other forms of conjunctivitis, and more serious vision-threatening conditions, so that patients can be directed to appropriate treatment and, when necessary, further referral to specialist ophthalmologists.

In medical terms, symptoms are subjective experiences of a patient who may or may not have signs of pathology. The symptoms of infective conjunctivitis often start in one eye and, after a couple of days, the infection spreads to the other. The symptoms of infective conjunctivitis are listed below, however there may be individual variations:

- Red eyes - as a result of irritation and the widening of blood vessels in the conjunctiva
- Watering eyes - as a result of the irritant causing glands in the conjunctiva to become overactive and produce mucus and tears
- Sticky coating on the eyelashes - due to the mucus and pus that is produced coagulating on the eye lashes
- Soreness of the eyes
- Enlarged lymph nodes in front of the ear

Close contact with someone who is already infected increases the likelihood of developing infective conjunctivitis; thus it is essential that hands are washed thoroughly after coming into contact with infected individuals (Krachmer, Mannis and Holland 2011).

The risk factors of developing infective conjunctivitis include:

- Extremes of age. i.e. the very young or old, due to weaker immune systems
- Recent upper respiratory tract infections, such as colds
- Diabetes; due to the long term complications of this condition (a weakened immune system)
- Corticosteroid use these medications weaken the immune system
- Blepharitis; This is an inflammation of the rims of the eyelids
- Crowded places

In the majority of cases, a diagnosis of infective conjunctivitis is reached from a patient's history and examination of the eye. Red, swollen eyes covered in sticky discharge are cardinal signs of infective conjunctivitis (Krachmer, Mannis and Holland 2011). Usually no further tests are required for the diagnosis of infective conjunctivitis, however at times a healthcare professional may suggest a swab test to determine the cause of the infection (or if there is doubt concerning the diagnosis) and will subsequently manage the treatment accordingly. Irritant conjunctivitis is due to the introduction of irritants such as chemicals or foreign bodies to the eye. This type of conjunctivitis often resolves with the removal of the irritant. Allergic conjunctivitis occurs as a result of an allergen making contact with the eye. Allergens are substances that cause the host immune system to react (often abnormally) and result in irritation and inflammation (Krachmer, Mannis and Holland 2011).

Diagnosis is initiated with a physical examination of the eye, particularly noting the degree of redness, with consideration of the patient's history. Factors such as a recent history of trauma, pain and discomfort, seasonal or otherwise recurrent incidence and changes in the eyelid may all be significant. (Wirbelauer 2006), (Shapiro and Croasdale 1997). It is important to appreciate that redness of the eyes is also associated with other conditions; for example, contact lenses, as described by Millis (1997). It is therefore very important to correctly identify conjunctivitis as the true aetiology of the redness.

Most of the symptoms of red eye are common and managed in most cases in primary healthcare centres (Wirbelauer, 2006). Treatment of conjunctivitis includes systemic and topical antibiotics in some cases while in other cases hospitalization followed by parenteral antibiotics is required (D'Aversa et al. 1995). Most serious cases are usually associated with pain, and trauma. Chemical burns and penetrating injuries are strongly advised to be referred to eye doctors or ophthalmologists (Wirbelauer, 2006).

All pathologies have features and associated symptoms from the start of the disease to the end of their life cycle. The development stages of the pathologies which form the focus of this research were investigated with a view to understanding how redness functioned as a diagnostic marker. The study then investigated how the degrees of visible redness were presented in digital imagery and how such images could be classified on a colour intensity scale as an aid to diagnosis and treatment. Figure 2.7 shows a sequence of images taken by the researcher tracking the severity of a patient's conjunctivitis as their trreatment progressed.

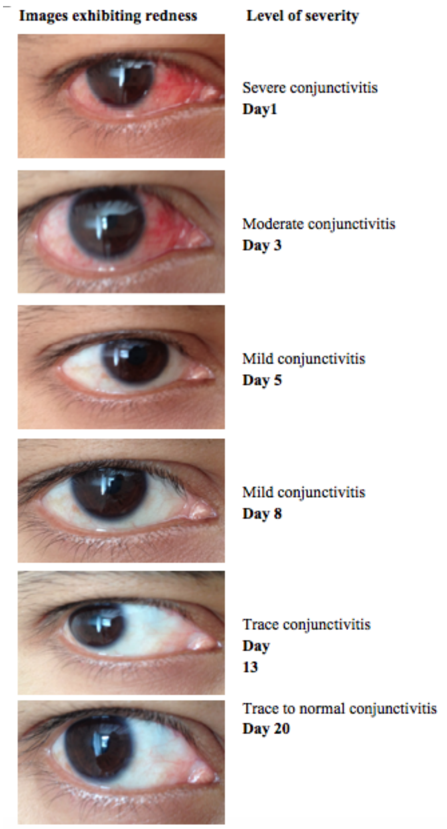


Figure 2.7: conjunctivitis severity levels improving gradually, taken by the researcher (July 2013)

The patient's condition responded to treatment and improved gradually with time. Details of this are provided in later chapters. The images in figure 2.8 are from the same imaging report for the same case. This time the eyes are photographed from different angles as this technique has been shown to contribute to improved colour perception (Brenner, Granzier and Smeets 2007). All images were taken using a digital camera with high resolution (5.0 mega pixel), under the same standardized capturing conditions to ensure clarity and consistency.

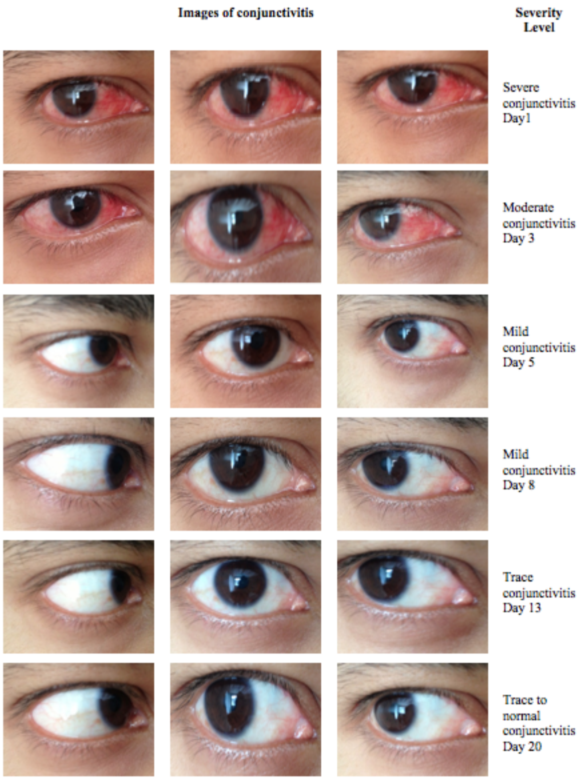

Images of conjunctivitis
Severity Level
Severe conjunctivitis Day1
Moderate conjunctivitis Day 3
Mild conjunctivitis Day 5
Mild conjunctivitis Day 8
Trace conjunctivitis Day 13
Trace to normal conjunctivitis Day 20

Figure 2.8: conjunctivitis severity levels improving gradually, taken by the researcher (July 2013)

2.3 Cognitive Engineering (CE) link

The field of cognitive engineering (CE) involves a number of different disciplines such as psychology, artificial intelligence, linguistics, philosophy, anthropology, engineering and the neurosciences. Today, after cognitive science has become well established and internationally recognised, many of its applications can be found in the area of human computing interaction (HCI), software design, information systems, human factors engineering, healthcare and education (Gersh, Mckneely and Remington 2005). The current study, which focused on using digital images and colour perception in telehealth, was a multi-disciplinary combination that integrated HCI, information systems, human factors, engineering and healthcare. CE (sometimes called cognitive system engineering or CSE) is a comparatively recent discipline that refers to the design of technology, training, and processes intended to manage cognitive complexity in sociotechnical systems. This definition includes identifying, judging, attending, perceiving, remembering, reasoning, deciding, problem solving, and planning (Militello et al. 2010). The current study encompassed all the mentioned cognitive processes (Militello et al. 2010).

Healthcare is heading increasingly towards the general adoption of paperless information systems to avoid errors in transcription, storage, medical history retrieval, and updating of patients' records especially when moving locations. This development in the attitude toward the use of technology is not only in medical records but also in diagnosis and medical consultation. The application of information systems as diagnostic and consultative tools is termed telehealth or teleconsultation or tele-clinic (Lawler et al. 2011). Telehealth is one of the applications of CE in healthcare (Bisantz et al. 2009). Many studies have investigated different areas of telehealth. One of these studies is Enomoto et al. (2006) who studied cardiac care by telehealth nurses in diagnosing cardiac patients through phone interviews. This literature review investigated key and recent studies in the field of telehealth.

Among the main cognitive elements are thinking, memory, language, and the use of knowledge. These faculties are essential for matching colours and digital images, which is the main focus of the thesis. Moreover, using common language or terminologies during human telehealth interaction is also important to filter and extract useful information for the diagnosis. It is important to use HCI effectively and efficiently in order to improve healthcare service, to save costs and to increase human reliability. Otherwise, it may bring more errors rather than preventing them. It is anticipated that, as technology advances, healthcare services will also be enhanced. However, choosing the right technology to suit each case and the appropriate provision of training for healthcare professionals to enable them to use the technology is critical to avoid unexpected errors. Kanthraj (2009) suggested that face-to-face methods are recommended for initial diagnosis; store and forward for follow up cases; videoconferences for medical education and counselling; satellite links for rural areas and mobile for home visits. Additionally, Militello et al. (2010) elaborated that the main purpose of CE is to make systems less complicated, safer, user friendly and free of errors. These were core motivations behind the current study. Furthermore, the study was linked with the fundamentals of cognitive science, because when participants in a telehealth system receive medical information and process it, they use cognitive functions to analyse the information and make decisions in order to provide an accurate diagnosis. The study further investigated human colour perception and its effects on the accuracy of diagnosis when using digital images in telehealth systems. It also looked at the rate of errors that may occur during the interaction between the operators and the system when ranking and classifying redness in medical digital images and the influence of clinical experience and training on the accuracy level of the colour classification. CE supports healthcare's objectives by focusing on either preventing errors or establishing an effective and efficient system. For example, communication between the users of healthcare systems has been the most frequent cause of medical errors (Rogers et al.

2004). The main three areas in the project that were integrated with the above CE were:

- Human errors, classification and telehealth
- The influence on diagnostic accuracy of the confidence level of the system operator
- Human colour perception and factors that may affect perception

The following are briefs on the above key cognitive issues.

2.3.1 Cognitive systems and human errors

The concept of human errors is used loosely with the assumption that everybody has the same understanding. This leads to many interpretations. The term can have at least three overlapping meanings: human errors as cause, human errors as events or actions, and human errors as consequences (Latino, 2007). However, there is no universal definition of human errors and there is no universal human errors classification system. Latino (2007) attempted to produce three generic definitions. These definitions encompass the following:

- Wrong action or decision that affects or has potential for reducing the safety, effectiveness or performance of a system.
- An action that led a system to another new task that is outside it's acceptable limits.
- An action the result of which is not desired by the set rules or an external observer.

Landrigan and Friedman (2007) stated that medical errors are common in hospitals and that they cause as many as 98,000 deaths each year in the United States. This was estimated in 2007 by the US Institute of Medicine (IOM). Byrne (2011) also reported similar concerns in the context of the UK, where clinical errors are a key cause of both morbidity and mortality, which is estimated to cost the health service in the UK over 1 billion every year. Even though it is difficult to define errors and differentiate between them it is very important to do so. Byrne (2011) described user errors, usually with doctors in clinical settings, and emphasised that, despite doctors being trained individuals, they make a considerable number of errors due to the limited available capacity for information processing. Even if sufficient information is available, this does not guarantee its correct use. Schulz et al. (2011) suggested that situational awareness can have three phases: perception, when the subject is aware of the issue, comprehension, when the subject is consciously aware of the meaning of the situation, and projection, when the subject can predict the consequences of the situation. This project focused more on the first phase where the colour perception of doctors and other participants of telehealth systems can be recorded and analysed as they rank and classify the colour red, rather than the actual diagnosis process itself.

According to Reason (2000), classifying errors is very difficult as there is no universally agreed system. However, he defined three stages for an error: planning, storage, and execution. Planning is a process that involves thinking and other cognitive procedures. In the case of this project, this was the mechanism that was used by the participants before performing the tasks. Storage is mainly about remembering and recalling previous experiences or knowledge; in our case it was recalling similar digital images that have similar medical conditions, especially redness degrees. The final stage, execution, refers to actually performing the tasks, in this case describing, grouping and ranking the images. Reason also noted two types of mistakes (failure of expertise and lack of expertise) corresponding closely to the rule-based and knowledge-based levels of performance. Failures of expertise are usually when useful information is applied incorrectly and a lack of expertise is when the user has no pre-knowledge which can be applied to the task.

This project sub-divided the participants into doctors and non-doctors. It is generally assumed that the errors of doctors are due to failures of expertise and the errors of non-doctors are due to the lack of expertise.

This was the basis of two of the hypotheses in the current study (hypotheses number 1 and 4, chapter 3, research hypotheses section).

According to Reason (2000), errors are either constant or variable. Constant errors can be easily identified, recognised and predicted, and as a result, should be controlled or avoided in any system design. However, variable errors are difficult to identify, recognise, and predict because they are not repeated in a systematic way. This makes it difficult for any system design to avoid or control them. For example, if an error occurs intermittently in a system in a multiplicity of states or under varying conditions then it will be difficult to design the system in such a way as to avoid such errors. However, if an error occurs consistently and predictably when a system is in a particular state or under particular conditions then it should be easier to avoid the error by designing the system so as to anticipate the error and prevent it.

In the context of this thesis, variable errors would arise from mishaps by the participants in their answering of the questions posed in the experiments, or, relating this to real life situations, from misdiagnosis by medical staff working within the telehealth field. Constant errors, on the other hand, would be created by faults in the operational system, such as substandard equipment, poor technology, adverse environmental conditions or poor design of the experimental procedures. In response to these considerations, an agile development and design methodology was employed for the study. Reason refers also to the potential for recurrent errors, which may arise from tasks involving routine, repetitive activity combined with distraction. The relevance to the experiments of this thesis lies in the fact that both doctors and non-doctors took part in the experiments. It would be natural to assume that doctors, given their field of work and expertise, might perform better in the tasks of the experiments (Tversky and Kahneman 1974). It is possible, however, that since the experiments might form the type of low-level, repetitive work referred to by Reason, that non-doctor participants, for whom the work would be new and more challenging, might prove to be more focussed, and therefore more accurate, than doctor participants.

The focus in telehealth, as a safety critical system, is on controlling potential risk. This requires critical analysis of all the conditions in which errors may occur and needs to be identified and addressed as part of any design. Figure 2.9, adapted from Reason (2000), shows how errors can occur when the system has latent conditions or weak points. These conditions need to be identified and addressed in order to control any active failures and prevent further errors. The slices of cheese in the figure 2.9 represent defenses against accidents and incidents.

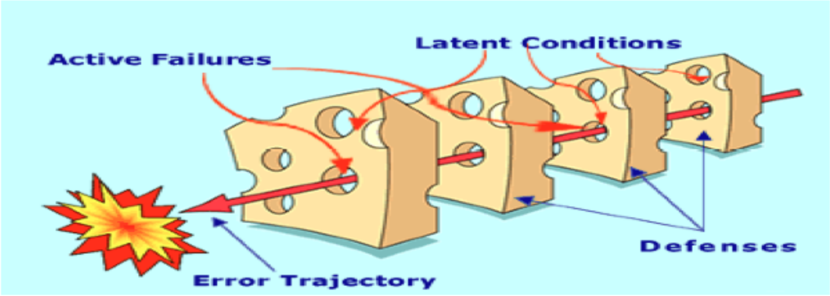


Figure 2.9: Human error systemic model. Adapted from Reason (2000)

Reason (2000) also said: “We cannot change the human condition, but we can change the conditions under which humans work”. The study attempted to understand these conditions and how they could be changed or controlled in the design stage rather than attempting the more complex and unpredictable task of changing the human system. In relation to the above, the study investigated the impact of using colour intensity scales in diagnosing red eyes and explored this by comparing the rate of errors when using and when not using the scales. Furthermore, the study compared the rate of errors when using a scale of three divisions with the rate when using a scale of five divisions.

2.3.2 Human error prediction

In general, human error prediction or control is a very complex task, especially if their forms or behaviours are not clear or are not easy to define when they occur. As mentioned earlier, constant or frequent errors are easier to predict compared to variable errors. However, the accuracy of any error prediction depends on our understanding of errors, their nature and under what conditions they occur. This study used an experimental approach to identify possible errors by doctors or non-doctors when classifying and ranking the colour red.

McIlvaine (2006) stated that it is possible to predict errors from people in their first experience of a task where they learn new skills. He also ranked the progression of the operators of a system into five levels based on their knowledge and experience: novice, advanced beginner, competent, proficient and expert. He stated that the main difference between these levels is not in years of experience but rather in the way that knowledge is applied. This is important for the current study because the focus in any evaluation is usually upon years of experience. However, McIlvaine (2006) believes that this emphasis on counting the years of experience is not warranted. This recognises that inexperienced doctors, because of the way that they apply their knowledge, have the potential to achieve as accurate results as experienced doctors.

Reason (2000) stated that there are three major elements causing errors which must be understood and studied in depth:

1. The nature of the task and its environmental circumstances.
2. The mechanisms leading performance.
3. The nature of the individual who is interacting with the system.

These three elements are critical when analysing any operational system. The level of complexity in performing a task, together with external environmental factors, can affect the accuracy of the outcome. For example, diagnosing digital images showing redness on a flushed skin is more complex than diagnosing digital images showing redness on brown or white skin, and there could arise confusion in differentiating between the normal red and the pathology. Also, diagnosing in a room where there is poor lighting and poor image presentation is more subject to errors than when the environment is well designed and standardised. The type of tools and procedures used when performing a task could also influence the accuracy, for example, a camera with a higher resolution would be expected to show more details when capturing specific medical conditions than a less capable device. Furthermore, the physical circumstances and condition of individuals also play a significant role in any diagnosis. For example, an experienced medical doctor, who is not colour-blind, who does not have any visual impairment and who is not physically tired at the time of diagnosis, is more likely to be accurate in judging colours. This study investigates the effect of medical background, using a colour scale, on the accuracy demonstrated when judging colours during the performance of different tasks, such as describing, grouping, ranking and matching colours.

Reason (2000) stated that understanding this model is enough to forecast the conditions under which an error will occur, and the exact form of the error.

If telehealth errors can be predicted by studying these three elements, then they can be avoided or at least controlled. However, they might still occur in different forms if conditions change. Similarity and frequency are very important concepts in order to control errors. For example, if there is an error that occurs every time a system enters a particular state, then it can be predicted that the error will reliably occur each time that state is entered. This type of knowledge about the errors and the conditions under which they arise makes the control process easier. Potential slips and lapses in telehealth systems might be detected and controlled by studying these important factors and conditions. However, mistakes of this category are harder to investigate because they are cognitive in nature and their conditions are not easy to define.

Latino (2007) divides errors into two categories: errors of commission, when an inappropriate action results in something that was not intended, and errors of omission, when a lack of action leads to undesirable consequences. Reason (2000) and Norman (1983) stated that human errors can be classified into two types: mistakes and slips. Mistakes occur through conscious deliberation. They are very conceptual and cognitive in nature. Slips are unintentional and happen by "accident". They also differentiated between mistakes and slips by looking at the developmental stages at which the errors typically occurred. In this classification, mistakes occur at the cognitive and conceptual planning stage during which goals are identified and the means to achieve them are decided. Slips occur at the execution stage, when actions are undertaken which do not conform to those intended or anticipated during the planning stage. Reason (2000) added another stage between the planning and execution phases. He named this middle stage the "storage" phase. This comes before running the intended actions since these do not typically start immediately. The cognitive nature of the mistakes makes them a greater danger, more complex, less understood, and harder to detect. There are different stages in telehealth: diagnosing protocol, or in this case teleconsultation, and tele-diagnosis, in this case using digital images that show red eye or cellulitis. There are possible human errors at each stage in the proposed system, (Reason 2000; Norman 1983), which need to be classified into mistakes and slips. These should be analysed and studied separately because they are different cognitively.

2.3.3 System responses to error

Traditionally, the focus, when an error occurred, used to be on the individual who made the error and who was held to be responsible for the failure (and the possible legal, or other, consequences). This was a universal attitude. However, the focus nowadays is more often on the design of a system which can avoid or minimise errors (Dekker 2006).

A complex system is more likely to contribute to a higher rate of human errors. For example a system which has 10 steps in which the participants are performing the task with expected 99% accuracy will result in an overall accuracy for the system of 90.4%. One method of re-designing systems to reduce the likelihood of errors is to reduce the number of steps in the system. Since fewer potential opportunities for errors exist in the system the likelihood of errors is proportionally reduced. This understanding of minimising medical errors by reducing the complexity of the healthcare system is stated and detailed by Landrigan and Friedman (2007).

Latino (2007) described three strategies to reduce human errors. Each one of them was designed based on the level of risk and its consequences. These three strategies are:

- Exclusion strategy. This strategy means that the system will be designed in such a way that it does not allow the error to occur at all. This strategy is reserved for situations in which human errors may have catastrophic consequences.

- Prevention strategy, which has a procedure to prevent errors that could happen because the system is not designed to stop it. This strategy is indicated for errors that are not of such criticality as would warrant the application of an exclusion strategy.

- Fail safe strategy, which is designed to lessen the consequences of human error rather than to prevent them altogether.

Reason (2000) listed the following six possible reactions that a system might produce in response to to an operator's error.

1. **Gagging**: Gagging means that a system knows already a list of errors which refer to the user's unrealisable intentions. When the user tries to perform an unacceptable action or command, the system automatically enters a "gag" state that prevents the command from being processed.

2. **Warnings**: In this case the system does not present a block to any inappropriate action from the user but it issues an error message that informs the user of a potential dangerous situation and then the user must respond and decide about the next action.

3. **"Do nothing"**: This is the simplest technique. A system does nothing when any illegal command is processed and fails to respond. In this case the user needs to think about what went wrong. Sufficient and appropriate feedback is very important to the user.

4. **Self-correct**: In this case, when an error is detected, the system tries to give solutions and guess the most appropriate action.

   The best case is when the estimation is correct; the worst case when it is not correct but still legal and safe. This would then require the user to undo the action and make another choice.

5. **"Let's talk about it"**: This is when the system switches to an automated dialogue which allows the user to interact directly with the system in order to locate the error.

6. **"Teach me"**: The system asks the user to explain something that is unknown or unclear to the system.

   An interactive dialogue is then initiated which continues until the system accepts the new or undefined terms. In other words, the system will discover what was in the user's mind.

For telehealth, all of the above possible responses should be studied carefully in order to find out which of these options can be practically implemented for which applications or in which circumstances.

2.3.4 Users and individual differences in system design

Any design is expected to consider the characteristics of the users in the context of the whole system, and not just the interface design. This was detailed by Pfautz and Roth (2006), Preece, Sharp and Rogers (2015), and Hartson and Hix (1989) in addition to Preece and Benyon (1993).

Neglecting this principle leads to poor use of technology which can lead, in the case of telehealth, to errors in using and learning the system, as well as increasing the workload for the users.

2.4 The Use of a Confidence Scale

The following section looks at the reasoning behind the use of confidence scales within the experimental methods of this thesis and at details of their application. The current study used a confidence scale from 0 (no confidence) to 9 (absolute confidence) that was used in conjunction with the experiments conducted within the study. It was also decided to use both doctors and non-doctors within the scope of the experiments in order to provide a comparison of experience versus accuracy.

2.4.1 The concept of human confidence

Confidence generally refers to firm trust with a high degree of certainty in something or somebody. This includes beliefs, knowledge, perceptions, predictions, judgments, and decisions. The term confidence also refers to feelings of certainty or a sense of self-reliance about our abilities in different situations Hawkins (1991).

Furthermore, Kepecs and Mainen (2012) have also defined confidence as "The degree of belief in the truth of a proposition or the reliability of a piece of information (memory, observation and prediction) or an estimate by the decision-maker of the probability that a decision taken is correct". In the light of this definition, the current study focused on the certainty that the users have in their answers in relation to colour and diagnosis.

2.4.2 The concept of overconfidence and underconfidence

Kepecs and Mainen (2012) reported that overconfidence occurs when a person believes that their judgment is more accurate than it actually is; while underconfidence occurs when a person believes that their judgment is less accurate than it is. They stated that overconfidence is the measure of calibration, which is the discrepancy between confidence and objective reality. Klayman et al. (1999) and Mann (1993) reported that the overconfidence of a person denotes a level of confidence that exceeds the level of accuracy in their performance when doing a task.

They found that there is a difference between confidence and accuracy due to overconfidence in most cases and underconfidence in some cases. They stated that overconfidence increases directly in proportion to the information that a person retrieves to support their initial impressions, Overconfidence also appears to be innately linked with difficult tasks and underconfidence usually with easy tasks.

These conclusions were supported by Kebbell et al. (2010), who added that overconfidence is more likely to occur if the participants are unable to relate to the given tasks, also it is commonly observed that people tend to be overconfident in their decisions most of the time (Lichtenstein and Fischhoff 1977). The current study investigated the level of confidence with both easy and difficult tasks.

Kahneman (2011) discussed the concept of overconfidence proposing that intuitive decisions are routinely based on memories of past experiences. It would be reasonable, given their experience, to assume that doctors would necessarily be more accurate in their evaluation of the redness associated with infection than non-doctors. Kahneman (2011) points out, however, that although the past experience of doctors may lead to greater confidence, this may not necessarily equate to greater accuracy, due to the heuristic which he calls the "illusion of validity". This may result from an individual having less experience than he believes himself to have, or from experience of one type of pathology being erroneously applied to a different pathology. This is an important consideration within this study which employs images of both conjunctivitis and cellulitis. Kahneman (2011) refers to rule based and case based thinking. Fast thinking, which he refers to as "system 1" and slow thinking, which he terms "system 2". System 1 tends to be intuition based and relies heavily on recollections of past experiences whilst system 2 involves a slower and more analytical approach to decision making. As a result, it is possible that when viewing and evaluating the qualities of visual stimuli, doctors may well rely on system 1 thinking, using their past experiences as a baseline for evaluation, whereas non-doctors, having no such experience on which to rely, may well use the more carefully judged system 2 thought process. Thus, applying Kahneman's logic, it is a possibility that non-doctors may prove to be as accurate as, or even more accurate than, the doctors in the context of the experiments.

2.4.3 The relationship between confidence and accuracy:

Studies such as Ames et al. (2010) and Ordinot, Walters and Van Koppen (2009) argued that confidence is not a reliable predictor of accuracy. The latter, and others such as Mengelkamp and Bannert (2010) and Kebbell, Evans and Johnson (2010), suggested that it can, nonetheless, generally be a useful indicator towards accuracy.

Kahneman and Klein (2009) made a clear statement that confidence is no measure of accuracy. They postulate that instinctive judgement is really only valid when applied by individuals with considerable experience working within a familiar environment. In the majority of cases, in their opinion, a false impression of an individual's own ability may well give rise to a level of overconfidence. In such cases it is probable that instinctive judgements are likely to be less accurate than the individual supposes them to be. Kahneman and Tversky (1996) discussed the use of confidence scales (which they refer to as "probability scales") used in conjunction with accuracy assessments. They highlighted that in such a case, levels of confidence which exceed levels of accuracy may be due to either overconfidence or a failure to use the scale as instructed. They further stated that the degree of confidence may well be governed by the perceived simplicity of the task and the knowledge or previous experience of the individual taking part. In the case of this study, and in the context of telehealth, doctors are expected to produce a higher level of accuracy than the non-doctors due to their medical knowledge, their clinical experience, and their familiarity with the diagnosis tasks and the environment. Non-doctors are expected to have less accuracy for the mentioned reasons, especially if they rely on their instincts when judging colours. However, non-doctors will usually produce more accurate results if they concentrate when doing a task, while doctors might feel overconfident about the same task and underestimate its level of difficulty.

Confidence judgment has been of interest to psychologists and researchers for a long time and a series of studies have supported this theory. Peirce and Jastrow (1885) is one of the key early studies that stated the positive relationship between confidence in performing a task and its accuracy in an experiment where subjects were asked to report and rate their confidence in their ability to lift a weight. Adams (1957) reported their work and conducted further experiments based on their finding and used a scale of 10 categories from 0% to 100% to measure confidence. Little (1961) reported scales of 4, 5, and 100 discrete points. He found that the mean of the confidence of a user across all answers of an experiment does not have a positive relationship with the accuracy. He called this "deviation certainty". Little (1961) explored the confidence means of participants in each task of a set of experiments in order to compare it with the confidence means of other samples of users when performing the same tasks. The study also compared the means of confidence for the entire sample in different tasks, which showed whether confidence is consistent. Mathematical values were applied to the assessments of degrees of confidence and of degrees of accuracy in order to facilitate a statistical comparison of the relationship between the two. This study has applied the same methods to analyse confidence data.

Engelke, Maeder and Zepernick (2012) tested the relationship between an observer's confidence and their judgements regarding the quality of images. They developed predictive quality models to evaluate the quality of 80 digital images collected from a public image quality database using the peak signal-to-noise ratio (PSNR). They used the models to predict the mean observer confidence to confirm the outcome of the wider used opinion scores. They concluded that using a limited confidence interval and standard errors leads to a better assessment of the image quality than using wider confidence intervals. Also, they found that there is a strong association between the quality perception and confidence, which results in high accuracy predictive models. The current study used a confidence interval of 0-9. This interval was divided into three parts: no or poor confidence (0-3), medium confidence (4-6), and high confidence (7-9). A study in the area of criminal justice by Li Lee et al. (2012) found that confidence is a strong indication of the reliability of an eyewitness statement. Their findings were supported by another similar study by Ordinot et al. (2009) who found that most information recalled with high confidence was accurate.

A clinical study by McNiel, Sandberg and Binder (1998) found a positive relationship between the confidence that a sample of 78 physicians had in predicting the level of violence by 317 patients during their first week of admission in the hospital and the accuracy of their prediction. Another medical study by Dempsey and Burr (2009) investigated the relationship between confidence of 203 radiation therapists (RTs) and the level of responsibility that they were willing to accept in relation to providing treatments without being supervised. For their experiment, they developed a series of six clinical planning scenarios with three increasing levels of difficulty. They found that the level of confidence and willingness to accept responsibility depended on the knowledge and experience of the therapist as well as the level of difficulty of the tasks. Clear protocols and procedures as well as structured education and training are all vital for confidence improvement.

This conclusion was also supported by another medical study by Panduragan et al. (2011), which highlighted that confidence is a key for making decisions if the health-care professional is knowledgeable and competent to solve a problem. The study used cluster-sampling methods to measure the confidence of 189 nursing students in the standard of accomplishment of their tasks. They used a two-category scale to measure confidence (confidence or no confidence). They found that a lack of knowledge reduces the level of confidence. The investigations of Dempsey and Burr (2009) and Panduragan et al. (2011) were very relevant to this current study because they investigated the relationship between confidence and years of experience. Their study found that all radiation therapists (RTs) were confident when completing all the procedures regardless of the challenges, with the exception of newly qualified RTs in their first year of practice, who lacked confidence when the tasks were difficult.

Individual differences and self-belief are also key factors that impact on the level of human confidence. A cognitive study by Stankov et al. (2012) found that confidence can be influenced by human behaviour and trait. They tested the ability of 15-year-old students of mathematics and English and found that students with strong confidence had the potential for higher academic achievement. It can be concluded that the level of accuracy can be predicted by the level of confidence. However, another academic study by Sanders and Sanders (2003), that used an academic confidence scale to compare the academic performance and expectations with the level of confidence for two groups of students, found that confidence is only responsible to a small extent for the differences in accuracy. Unlike other studies, they used only 4 levels in a confidence scale. They found that the first level means very confident and the last level means not at all confident but they did not state any descriptions or percentages for the other two levels, which may have affected the results.

Edison et al. (2008) conducted an interesting comparative study of diagnostic accuracy and confidence between in-person, live real-time and store and forward teleconsultations. They used a confidence scale from 1 to 5, with 1 denoting no confidence and 5 total confidence. 110 patients and 4 dermatologists participated in the study randomly. The study showed high agreement between the three methods in diagnostic accuracy and confidence but the in-person method was higher than the other two methods in both accuracy and confidence. Live real-time teleconsultations and store and forward teleconsultations did not show statistical differences in the diagnostic accuracy results but live teleconsultations showed higher confidence than those using the store and forward method.

Krug (2007) reviewed four popular methods for examining the relationship between confidence and accuracy. First, the calibration curve, which plots a participant's confidence against their percentage of accuracy. An overconfident person has a high confidence level at low accuracy, while the opposite is the case for an underconfident person. A high confidence level with high accuracy means a well calibrated participant. This method was used by Krug (2007), Brewer et al (2002), Olsson, Juslin and Winman (1998). The second method reviewed was the use of an over-underconfidence value that is used to complement results derived from calibration curves. The scale ranged from -1 (underconfident) to 1 (overconfident) with 0 being a perfect calibration score. This method was used by Bornstein and Zickafoose (1999). The third method measured confidence on a percentage scale from 0% to 100%, but recorded accuracy measures at two levels: correct or incorrect. This method was used by Olsson and Juslin (2002). The fourth method reviewed, the gamma statistic, is used to measure resolution. Resolution is that which distinguishes accurate from inaccurate identifications of phenomena. This method was used by Brewer, Keast and Rishworth (2002) and Olsson (2002). Good resolution results from participants assigning high confidence levels to accurate identifications and low confidence levels to inaccurate ones.

2.4.4 Time relationship with confidence and accuracy

Time is an important variable explored by researchers when investigating accuracy and confidence relationships. Kepecs and Mainen (2012) stated that there was a positive relationship between the confidence of test participants and the time that they spent in making a decision. This, however, contradicts an earlier study by Vickers and Packer (1982) who found that there is an inverse relationship between confidence rating and response time. These two opposing views are examined and addressed in the experimental work presented in the current thesis (hypothesis number 3, chapter 3).

Wright and Ayton (1989) found that participants make fewer errors when they are given more time. In addition, Pleskac and Busemeyer (2010) concluded that time has only a moderate effect on accuracy of judgment but confidence does correlate positively. This moderate effect of time on accuracy was also validated by Lee et al. (2000) and Brewer et al. (2002). This positive relationship between time and accuracy level is very important when designing experiments and systems that aim at high levels of accuracy. The current study explored the relationship between three related factors: the accuracy of the doctors or telehealth operators in their red colour classification, their confidence and the time that they spent in judging the images.

2.5 Understanding Colour and Digital Images

This section provides a brief on key relevant aspects of colour and digital images in some of the pertinent previous studies.

Figure 2.10 demonstrates that the use of flash illumination makes a large difference between the images. The two images in Figure 2.11 show that different types of camera can make a difference to the quality and colour reproduction in images (Patricoski and Ferguson 2009)

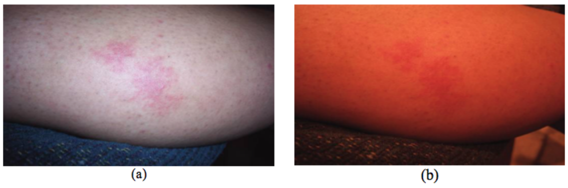


Figure 2.10: (a) shot with Kodak V1003 with flash and incandescent, (b) shot with Kodak V1003 without flash under incandescent lighting. The resulting skin tone has an orange cast.

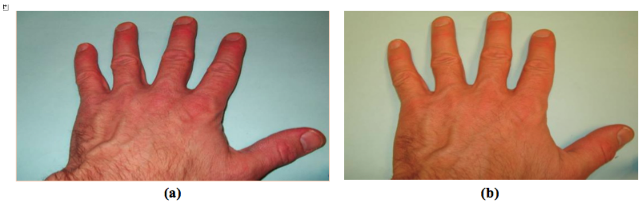


Figure 2.11: (a) An image made with a with Fuji F40 camera using flash demonstrates excellent (natural) skin colour, (b) An image of the same subject made with a Sony DSC-W200 camera using flash demonstrates a yellow cast to the skin colour

Images are initially taken in analogue format and then digitised. The range of colour values which an image can represent is referred to as pixel depth. The number of pixels (dots) in an image determines the resolution, in other words the degree of sharpness and clarity, of the image. Finally the image is converted back into analogue format for viewing.

Voipio and Lamminen (2002) supported the view that colour perception depends on the human eye without any physical variable representation. They mentioned factors that affect colour interpretation such as illumination, reflection and light absorption and emission by an object's surface. They also stated that cone cells in the retina of the eye detect three basic colour wavelengths: long wavelength (red), medium wavelength (green), and short wavelength (blue). Mather (2009) stated that different colours are interpreted by the eye and brain based on variations of these wavelengths.

Theroux (1995) stated that colours can be represented as standard primary colours; red, green and blue. Other colours can be formed by the addition or subtraction of these primary colours. Colours can be matched by mixing different proportions of the primary colours to produce secondary colours examples are blue + red = magenta, red + green = yellow and green + blue = cyan. The mentioned details were also explored by Gilchrist and Nobbs (2001).

Perednia (1991) detailed that any object regardless of whether it is visible or invisible, large or small, static or changing, can be imaged as long as some of its properties such as reflectance of sound waves, a change in elevation, or emission of light can be measured and recorded. For example the heart, a mountain, and a star all constitute valid imaging objects, even though the first is hidden by the chest wall, the second is massive, and the third is billions of miles away, because certain properties they possess can be described, measured and recorded.

In physics, colour is represented as an emission of a combination of wavelengths within the visible region of the electromagnetic spectrum (Nassau 2001). The visible region of the spectrum is that part which is visible to the human eye.

Colour representation in computer systems has a four-component representation; hue, saturation, intensity and the attenuation of the system. This representation of colour in computer systems allows flexibility, error detection and correction (Foley 1996; Foley and Van Dam 1982).

Diagnosis can be affected by an individual's interpretation of colour based on their perception of colour characteristics:

- Hue, which is the unique name of the colour and is an attribute associated with the dominant wavelength in a mixture of light waves (Gonzalez and Woods 2008).
- Saturation, which is the purity of the colour or how much neutral/white colour is present. Saturation also refers to the relative purity, or the amount of white light mixed with a hue. (Gonzalez and Woods 2008).
- Intensity, which refers to the light in the colour or its brightness. Lightness and brightness are often interchangeable terms. Lightness is a scale from black to white, whilst brightness is the intensity as perceived by the human visual system (Gonzalez and Woods 2008).

In colour, connecting the mind to the physical world, leading scholars from cognitive psychology, philosophy, neurophysiology and computational vision provide an overview of contemporary developments in our understanding of colour and of the relationship between the "mental" and the "physical". Human perception of colour is in terms of its attributes hue, saturation and intensity (Mausfeld and Heyer 2003). In the psychology of colour, design and human perception: The "hue" of the colour is the unique name of the colour; the "saturation" of a colour is the purity of the colour in terms of the amount of "white" present in the colour; and the intensity of a colour refers to the light in the colour and its brightness or value (Mather 2009).

In medicine, different colours sometimes indicate different types of pathological or physiological states. Medical doctors monitor the medical conditions of a patient's responses to treatment and follow the different stages of pathologies. At every stage, a pathology might exhibit specific qualities of colour. Interpretation or perception of the colours affects the way that medical staff capture, transmit, store, display and classify images. Most relevant to the present study, the perception and interpretation by medical personnel of the redness presented in cases of cellulitis and conjunctivitis could have a significant effect on the rate of errors when using telehealth.

2.5.1 Colour transmission

In the field of telecommunications, colour can be transmitted by different means. Each of these techniques can be more useful if used in specific ways and conditions, considering the medical needs and technological requirements in telehealth systems. in the case of this study, in order to maximise the benefits and ensure the accuracy of the diagnostic process and also to avoid any possible errors that may affect the clarity of the colour in the image, which may lead to diagnostic errors.

Transferring a colour is part of transmitting the image that has the colour. There are many techniques that can be used in colour transmission. These techniques are varied and each of them has advantages and disadvantages. A technique may be best suited for a particular kind of colour transmission, depending on the image size, quality, and mode of transmission and so on. The following section explains some of these techniques:

Transmission without compression

This technique can be used when the size of medical image is small or if the mode of transmission is effective such as a network with very high speed broadband connection and large bandwidth size capable of transferring large-sized image files easily. This method is considered as the best option for transferring images for secure image quality, colours and clarity. The received images and colours would have the same or almost the same quality as the original. This technique, however, can be very difficult to be used practically in most of the current telehealth applications because its technological requirements are very expensive to acquire and maintain. Furthermore, the transmission may take a longer time because the sizes of the files that carry the images are too large (Zukoski, Boult and Iyriboz 2006).

Transmission with compression

Image compression makes things easier by reducing the size of data making up images and colours.

Uncompressed images have larger sizes because they typically contain unnecessary repetitive and predictable data. Compressed images and colours are lighter and easier to process, store, and transmit because they use only the relevant image data while the other remaining common data is reintroduced when the image is decompressed at the other end of the system (Zukoski and Boult and Iyriboz 2006).

Uncompressed images have the disadvantages mentioned earlier which could affect the overall telehealth care for patients particularly in emergency or critical situations where these images would take too long to be processed, saved, and transmitted.

Image sizes are usually large and it becomes a challenge to store, process, transmit and receive the images but image compression techniques can be employed to reduce the size of the image without any loss in the information contained in the image. Two types of image compression techniques can be employed: lossless and lossy compression: Lossless compression allows all the information to be preserved without any loss to the original information, when this information is stored, processed, transmitted and received the same information can be retrieved without any loss to this information. The lossless compression method application to image compression keeps most of the original colour and image data in the file produced, and it reduces image size to 5-25% of the original. When reviewed, lossless compression files are identical to the original image because the 5-25% of the data that was previously removed. Examples of lossy file format include Joint Photographic Experts Group JPEG (JPF), Tagged Image File Format (TIF), and Portable Network Graphics (PNG).

In lossy compression techniques an average of the original image information is used for compression and this compressed information is stored, processed, transmitted and received, but not all the image information is received due to the compression technique employed. Either of these techniques is acceptable depending on the application or the quality of information required. Whatever the compression techniques used, there are errors involved in the equipment used to compress the image information because the equipment employed is not 100 percent efficient and has an acceptable tolerance level which is usually specified in the equipment performance characteristics.

The lossy compression loses or discards larger amounts of unnecessary data in the image to produce lighter size presentations of the original images. In this technique, the image can be reduced to 80% of the original. When decompressed, they do not look identical to the original because large amounts of data in the image have been lost, which may include some of the colour. The most common file formats for lossy compression are JPEG and GIF. Lossy and lossless compression can be used in medical imaging such as teledermatology as both can give clear enough image data for diagnosis. However, lossless compression gives a clearer and identical image to the original when viewed which makes it more acceptable than the lossy compression for usage in medical imaging (Zukoski , Boult and Iyriboz 2006).

2.5.2 Image quality and colour resolution

Digital images have properties such as image resolution, image pixel dimension, Image bit depth, image dynamic range, image file size, image compression and file formats which affect image quality. An understanding of the image properties makes it easier to have a clearer appreciation of image composition and the effect of image properties on image quality. Colour resolution usually refers to the number of different colour pixels, which can be represented in an image, which is stored, viewed, and or printed out. The higher colour resolution has more colour detail and the lower colour resolution has less colour detail. The current study focused on image resolution as an indication for the level of image quality because the amount of medical details in the image is critical for the accurate diagnosis. Maglogiannis et al. (2001) explored the following key factors that need to be considered when capturing and transmitting images and colours in teleconsultation

- Image resolution: Image resolution is determined by the number of pixels per inch. Increasing the number of pixels per inch increases the resolution and vice versa. Any limitations in image capturing equipment that result in low image resolution may affect the accurate representation of colours. However, most of the current generation of digital cameras and mobile phone cameras are capable of producing images with sufficiently high resolution.

- Noise or particles on skin: The GP or patients should avoid any noise which may affect image colours when captured and transmitted.

- Ambient illumination: The lighting and temperature of the environment around the medical image can easily affect the brightness of the image, which may mislead the diagnosis. In order to minimise these effects, care should be taken wherever possible to ensure that lighting conditions and room temperature are maintained at a consistent level. Images may be captured in indoor, outdoor or in controlled environments. Images captured indoors without special preparation may have poor lighting; those captured outdoors may be excessively illuminated. Ideally, images should be captured under controlled conditions where illumination and other ambient conditions can be optimised to obviate representational errors.

- Reflections: The manner in which light is reflected by some subjects may affect the representational fidelty of colours in an image.

- The size of the image: This is determined by a combination of an image's data content, resolution, bit-depth, compression techniques and metadata.

- Bit depth: Is the number of bits used to encode colour information in a pixel. A colour image, depending on the file format in which it is encoded, may have a bit depth of 8 to 24 bits or more. Colours are encoded by mixing red, green and blue components. A 16 bit colour image, for example, will use 4 bits each to encode the proportion of red, green and blue (and 4 bits to encode transparency) in each pixel. In the case of greyscale images, an increased bit depth will increase the tonal range of the image. A greyscale image is represented by having 2 to 8 bits or more for its representation. Bitonal images are those which have 1 bit which can encode two possible values: 0 or 1.

- Dynamic range: Is the ability of the image to produce tonal information which ranges between the lightest and darkest part of the image. An increase in the dynamic range will allow more shade representation.
- Compression: as mentioned earlier, image sizes can be quite large and become a challenge for storage, processing and transmission but modern compression techniques are able to reduce the image file size while preserving most of the fidelity of the representation.

There are different cameras and manufacturers with different performance standards, efficiency and qualities as capturing equipment. The capturing equipment can have errors arising from camera focusing and white balance. These types of errors can make the image blurred and the image colour may contain impurities. Ensuring the camera is well focused, whether by skillful operation of the device or by using a camera's auto-focus feature, reduces and eliminates focusing errors. Performing white balancing before image capturing ensures that the impurities arising from colour degradation will be reduced and eliminated.

Digital images are fundamentally represented as binary codes. for digital equipment to be able to read and interpret information stored in the image as digits or bits of pixel which can then be compressed and stored in the image capturing equipment for the purpose of viewing, processing or transmission and reception. The images are not readable by humans while in the digital form and need to be reproduced in analogue format for display. Equipment errors can be minimised by ensuring that all equipment employed for image capture, storage, processing, transmission and reception is of good quality. It should also be ensured that such equipment has been set-up appropriately for the task, has an acceptable level of efficiency for the task and has an error detection and correction mechanism built into it, or that it possesses the facility to have such a mechanism integrated with or added to it.

These colour representations allow display of colours in a web application with the aid of browsers. It is very difficult to decide what is ideal for setting up all the condition and factors which may affect the quality of the transmitted images. However, standardization of all these conditions is a key to ensuring consistent work and controlling for errors such as disordering or mismatching. McNeill et al. (2002) listed and explored the following factors as a good examples of standardization: using the same lighting in the room consistently with all work; placing a featureless background behind the patient when capturing images, to avoid reflections and shadows cast by flash illumination; using a black background to observe the flash shadow and get clean images; standardising the distance between the camera and the skin when taking images. If any of the aforementioned factors (for instance the distance, background or lighting) are changed then the quality of colour representation in an image may also change accordingly, even if taking images from the same patient and at the same time. The telehealth team in Aberdeen Royal Infirmary uses the aforementioned guidelines when conducting store and forward and live teleconsultations.

There are many subjective methods that consider the opinion of image observers in assessing the quality of digital images. Subjective opinion is relevant because human beings are the ultimate receivers of the images in most cases. Engelke, Maeder and Zepernick 2009; Ouni et al. (2008) explored these various methods. Among these methods are; single stimulus continuous quality (SSCQ), single stimulus impairment scale (SSIS), double stimulus impairment scale method (DSIS or EBU), double-stimulus continuous-quality scale method (DSCQS) and the mean opinion score (MOS). The latter was reported as the most reliable subjective method for evaluating the quality of digital images and videos (Engelke, Maeder and Zepernick 2009; Ouni et al. 2008).

However, it is time consuming (due to the number of images that are required to be reviewed), costly, and inconvenient (Ouni et al. 2008). There are also objective methods used for evaluating image quality such as full-reference (FR) metrics, no-reference metrics, and reduced-reference (RR) metrics. Each application area has its own method and criteria based on the required information and image conditions such as lighting, capturing device, transmission, processing, and display.
Ouni et al. (2008) assessed distortions in images as artifacts of the equipment used in acquisition, digitising, processing, restoration, compression, storage, transmission and reproduction. Their investigation suggested that the image processing techniques introduced distortions, which need to be known and analysed for their effects on image quality.

They explored both subjective and objective image quality assessment. Subjective evaluation, in spite of its drawbacks of being tedious, expensive and requiring manual execution, is still considered an effective method in image quality assessment. Their objective method was employing the specification of the image properties metrics as a mathematical model, human visual system or a no reference model in its assessment. The result of the experiment suggested that the subjective evaluation method is a reliable method for image quality assessment.

Engelke, Maeder and Zepernick (2009) experimentally investigated the reliability of user estimation of image quality and suggested the addition of observer confidence in assessing image quality. Their study analysed two methods of confidence measurement; a confidence score or a score of the time taken by a participant to provide a quality response. They conducted an experiment using a set of 80 grey scale JPEG formatted images portraying a range of severity levels. These images were presented to participants for quality rating, including an appraisal of their level of confidence in their ratings on a 5-point scale. The result from this experiment suggested that rating very good or very bad images was very easy but images between the two extremes became difficult to rate. It also suggested that the accuracy of the image assessment by participants depended on their response time. They concluded that there is strong interrelation between quality scores, confidence score and response times.

Burningham, Pizlo and Allebach (2002) investigated using image quality metrics including sharpness, graininess, tone scale, and colour retentions to analyse the quality of images produced on electronic devices. It observed that imperfections in images could be attributed to mechanical and electrical sources during the image capturing, storage, transmission, reception and reproduction processes. The results of this study suggested the effectiveness of image metrics as an effective tool in measuring image quality.

Chandler (2013) investigated the challenges presented by image quality assessment. Their study explored the actual image processing equipment and the changes this could make to an image during capture, storage, processing and transmission. The alterations made to processed images such as online shared photos, multimedia and streaming video can have an impact on the assessment of an image. They used an image quality assessment algorithm that took its input from an image's physical attributes, which was then compared to a stored reference image in an image quality database. The results from this experiment yielded a satisfactory outcome when comparing a distorted image relative to a reference image.

2.6 Human Colour Perception and Diagnostic Accuracy

The following section provides a brief review of human colour perception in key previous studies. Also treated are the factors that may influence perception and that may affect the accuracy of performance in related tasks that involve colour; in the case of this study, test subjecs' perception of the colour red during the diagnostic procedure for cellulitis and red eye.

2.6.1 Human colour perception in key previous studies

Human colour perception starts with the image receptors located in the retina of the human eye, which is the surface at the rear of the eye. There are two types of image receptors: cones and rods. Cones function under bright light and the rods under low light. They added that visual information is passed via the optic nerve from the retina to the brain area called the visual cortex, where the visual processing is completed.

A study by King (2005) examined colour vision and perception from an evolutionary and anthropological perspective. The eye employs photoreceptors, which convert light into signals for processing by the eye. The converted light can be represented in terms of its wavelength, which is visible to the eye with the aid of colour photoreceptors or cones. The cones are able to tune to different colour wavelength with the aid of photo pigments, which are a type of protein responsible for tuning the eye to perceive different colour wavelengths. There is a limitation to the visible light which the normal eye can absorb and perceive. Colour vision perception in animals can be classified based on the number of cones available for vision. Humans have three types of photoreceptors, which enable humans to distinguish up to 1,000,000 colours and are known to be trichromatic. The three photoreceptors in humans understand the following colours: blue, green and red, which are the primary colours from which other colours are formed. The study concluded that there is a relationship between human evolution and human colour perception.

A study by Lotto and Purves (2004) investigated human colour perception and defects in the human colour vision system. They stated that humans perceive colours by means of a visual system that sends visual stimuli to the brain for recognition and interpretation. The accurate perception of a colour means that this colour is recognised and interpreted accurately. However, if any of these two colour perception stages is not satisfied then an error may occur in the colour perception. This is the reason why it is possible for a particular colour to be interpreted by the brain and perceived differently by different people. A defect in the colour visual system occurs when there is a problem in the ability to distinguish one wavelength from another. Biological imperfections in the human eye can cause problems with colour perception. Some human eyes have conditions such as colour blindness, which affects colour perception and makes the colour perceived by the eye different from the actual colour of the object. Colour perception can also be explained based on its own physical characteristics in the form of electromagnetic radiation with wavelengths varying from 400 to 700nm. The physical characteristics of a colour and its electromagnetic radiation are critical for its recognition and interpretation (perception). The human visual system can identify the colour of an object by its own physical characteristics such as hue, intensity, saturation and brightness. The study of Lotto and Purves (2004) concluded that for the effective recognition and interpretation of an object's colour by the human eye, the colour of the object would have to be identified by its physical characteristics. Biologically the human eye should be perfect for reception and perception. However, any form of defect in the visual system can affect human colour perception. A very common example of how humans perceive colour is the popular perception of a dress colour that was posted online (Figure 2-11). The dress appeared gold and white to some but blue and black to others. As discussed by Winkler et al (2015), the perception of colour for humans could be daunting, as the eye depends on both the reflectance of object as well as the spectrum of the illuminating light source. This example shows how human vision varies in estimating colour of objects.

As seen in (Figure 2.12, a) the stripes on the dress that numerous observers perceived as white or blue becomes unambiguous shades of yellow (Figure 2.12, b). Lafer et al (2015) carried out a survey with 1,401 subjects that were asked to identify the actual colour of the dress, 57% of the participants identified the dress as blue and black, 30% as white and gold, 11% as blue and brown and 2% were not sure.

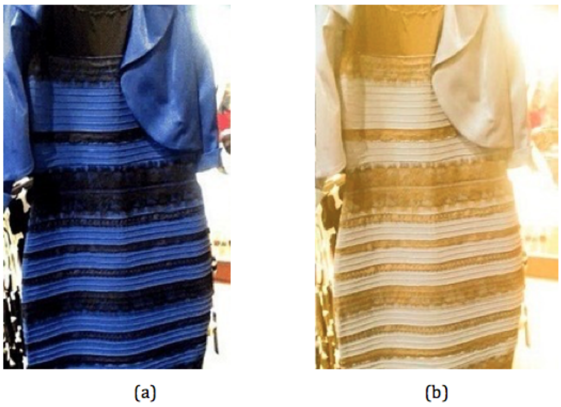


Figure 2.12: (a) The picture of a dress perceived as white or blue, (b) The picture of same dress in (a) perceived as unambiguous shades of yellow

In digital photography, four elements are required: The object to be photographed, appropriate lighting, a suitable location and the digital camera. The camera should be mounted on a tripod if available, the camera's white balance adjusted or auto white balance chosen, the camera's zoom should be adjusted and the image clearly focused.

A study by Bloj and Hedrich (2012) investigated the physiology of human colour perception. They stated that human eye is sensitive to electromagnetic radiation of wavelengths ranging from 380nm to 780nm. This radiation is absorbed by the retina and this information is transferred to the brain for interpretation. The processing of visual information in the brain can be very complex because it involves some 30 regions of the cerebral cortex. The human visual system is termed trichromatic because there are three types of cone, each type responsible for the identification of one of the additive primary colours red, green or blue. The three primary colours can be mixed to produce different colours. This is the same principle as is used in colour visual display systems such as televisions and digital cameras. This study also found that the colour perception of an object could be affected by its light source, reflection and location.

A guide to understanding colour communication by X-Rite, acknowledged as a global leader in colour science and technology (A Guide to Understanding Colour Communication report 2013) suggested using colour-measuring instruments to differentiate between colours and to assign numeric values to colours for the purpose of identification. Spectrophotometers, colorimeters, and goniometers are examples of the instruments suggested by the study for effective colour differentiation and identification in scientific applications. This study concluded that for proper identification of colour differences, having colour measuring instruments would be effective and efficient for different applications.

Green and Martin (1990) investigated the measurement and perception of skin colour in skin cancer. The skin colour changes in cancer are major features in the diagnosis of melanoma and non-melanoma. This study employed a skin cancer subjective grading method for the classification of skin colour changes in cancer patients. This method recognised fair, medium, olive and black using the reflectance of light at a wavelength of 650nm at six different sites.

This experiment suggested that females statistically showed a paler skin at all sites than males. This concluded that the physical characteristics of the skin are a key factor for colour perception. Therefore, employing colour-measuring scales in colour classification may help in the description, evaluation, diagnosis and treatment of skin cancer patients.

Another interesting study by Brenner, Granzier and Smeets (2007) investigated whether certain parts of an image such as the centre, its edges or where it is facing is important in the classification of image colours. The study employed classification image techniques for analysing the colour of an image. The result of this experiment suggested that a combination of eye movement and retinal adaptation can contribute to image colour vision and restricting eye movements can contribute to a large difference in colour matching tasks. A conclusion from this experiment suggested that the part of the image viewed, angles, and direction can contribute to the perception of image colour by the observer, with emphasis on the importance of eye movement to colour vision, as they reflect the area of interest in the image.

Tai, Yang and Liao (2010) stated that three basic elements are required for effective colour generation. These elements are light source, object, and the human visual system or human eye. Also they mentioned that the human brain records the information of the colour from the lighting reflection in the object to the retina inside the eye. Colour characteristics such as intensity and brightness are very important for effectively increasing colour quality. The study suggested that the variation of colour mixture and matching are associated with its identification and requires a colour management system which can effectively simplify the process of colour perception. A similar study by Beretta and Moroney (2008) investigated several factors which may affect the perception of colours such as reflections, the light source, colour properties of objects and colour sensitivity of observers, colour of illuminating source and wavelength. An experiment in the study, conducted to measure the colour of an object by comparing this to a reference coloured object in a three dimensional space, suggested that the colour of an object depends on the light sources, colour properties of the object, colour sensitivity of the observer and the interpretation of the perceived colour by the human brain.

Spalding (1999) explored colour vision deficiencies in the medical profession and the effect of these defects on the skills and decisions made by the affected medical doctors. It suggested that about 8% of male medical practitioners are affected by colour vision deficiencies. This figure could be significant in its effect on ability to make accurate diagnoses based on colour, and should perhaps be addressed by screening and testing against the required standards of allowable colour vision for effective medical practice. People affected by colour deficiency may be unaware of its extent and effect on making colour related decisions. This study explained that the prevalence of congenital colour-vision deficiency in the medical profession is about the same as in the general population. However, due to the sensitivity of colour in the identification, development and evaluation of pathologies in the medical profession, there is a need for a standardised colour visual system for medical practitioners.

2.6.2 Sex differences with colour perception and vision

There is a clear contradiction in previous studies about the impact of sex differences on human colour perception and vision. The following are key studies on this topic.

Abramov et al. (2012) found that there is a marked sex-related difference in colour vision. Jain et al. (2010) reported that women have better colour vision than men with high statistical difference. They added that women gave more correct answers and took less time than men especially in identifying red and green. He concluded that women can see a greater range of colours when compared to men. Early literatures showed similar results such as McGuinness and Brabyn (1984) who demonstrated a sex-related difference in visuo-spatial ability, with men outperforming women by one standard deviation on a number of different visuo-spatial tasks. Another early study by Nicholas (1884) reported that males had better hue discrimination. A contrary result was reported by Murray et al. (2012) and Hennon (1910) who found females have better hue discrimination. Kuehni (2001) found that there is difference between the two sexes. Pardo et al. (2007) found a significant difference in colour matching. Rodriguez-Carmona et al. (2008) found a substantial difference between sexes in red-green discrimination. Some literature argues that the difference in colour perception could be because of the character of the sexes but not the sex itself. Chaudhari and Shaw (2012) stated that females tend to match colours better than men. The authors stated that the reason behind this could be cognitive, in the divergent patterns of socialization for males and females. Jameson, Highnote and Wasserman (2001) argued that the difference in the ability to identify different shades of colours between sexes could be because of the genes influencing retinal photo-pigments. Women with more than four photo-pigment genes were found to perceive colour more clearly than others.

Male and female doctors and non-doctor participants in this study were treated similarly during the experiments, and in the analysis of the results, because the main focus of the study was the clinical background and not sex-related differences. However, the study includes an additional section presenting the results separately.

There are other factors that may affect colour perception such as geographic area, ethnic, language, and culture, religions, tradition, fashion, sex, and age (Tai , Yang and Liao 2010; De Bortoli and Maroto 2001). The human eye is capable of perceiving 7 million different colours. This study also suggested a language barrier to colour description because some languages do not have names or descriptive terms for some similar colours and this limitation can affect human perception of colours. The study also mentioned that non-cultural factors such as psychological factors (visual effects and contrast between colours), the shape of the object, health conditions such as schizophrenia (which can cause abnormal colour perception) and colour blindness may all have effects on human colour perception (De Bortoli and Maroto, 2001).

From the above studies, it is clear that a multiplicity of factors can potentially affect human colour perception during the performance of tasks that involve colours. They can be summarised into factors related to the colour itself, such as its wavelength, and its physical characteristics or properties; factors related to the object that has the colour such as its physical characteristics, light source, light reflection and light location; factors related to biological imperfection or colour deficiency in human visual system (human eye) such as poor colour memory, eye fatigue, colour blindness, the inability to distinguish one object from the other, and colour sensitivity and lastly, factors related to the use of digital images such as the area of interest in the image, the angle and direction of the image when captured and viewed. To avoid colour related errors due to these factors, there is a need for a standardised colour visual system especially in healthcare and telehealth systems in order to make an efficient diagnosis and treatment of pathologies.

All the above factors need to be considered when designing the diagnostic protocols of teleconsultation systems.

2.7 The colour red and image matching in healthcare and telehealth

The following section discusses the importance of the colour red and the use of a red colour scale in healthcare in general and specifically in telehealth. The section also explores key methods in image processing such as the use of image descriptors, image matching, image databases, isolating the area of interest in digital images, and colour feature extraction.

Images are made available for medical analysis using capturing equipment which is widely available as standalone cameras or integrated as part of a mobile digital system such as a mobile phone with integrated camera.

Digital systems find usage in medical imaging research because digital images are easily stored and can be easily transported over communication networks. Images in digital format are suitable for medical analysis.

2.7.1 The importance of the colour red in healthcare and telehealth

Punchard, Whelan and Adcork (2004) stated that the red colour is very significant in medicine because it is considered by doctors as one of the key symptoms or reference points when diagnosing many medical conditions. From a pathophysiological perspective the colour red or redness (rubor) is one of five cardinal signs of inflammation, along with the other signs: swelling, heat, pain and loss of function.

This classical description of inflammation accounts for the visual transformations that are perceived during several disease states. McNeil et al. (2002) elaborated that colour in images provides critical information for medical experts when diagnosing pathologies remotely.

It is acknowledged that inflammation is a complex process which involves a major immune response to infection (although not all infections cause inflammation) and tissue destruction. There is an extensive range of causes of inflammation such as bacterial or viral infections, or trauma. The course of inflammation can also range from acute inflammation, for example as a result of S. aureus infections of the skin, to chronic inflammation which can result in atherosclerosis, a build-up of fatty plaques on arterial walls. The increased redness that is perceived is due to additional erythrocytes (red blood cells) passing through the site of inflammation. Since rubor is a cardinal sign of inflammation it would be justifiable to state that redness is likely to play a significant role in the diagnosis of clinical diseases such as psoriasis, candidiasis, phototoxic rash or cellulitis. Most of these pathologies are dermatological in nature and involve rubor as a sign. Redness is also a key symptom in other medical conditions such as conjunctivitis. The importance of rubor varies depending on the particular condition or illness.

2.7.2 Standardisation of teleconsultation

This section focuses on the minimum quality that must be ensured when using digital images in healthcare in general and specifically in telehealth; on the use of image quality scaling or classification in previous studies and how they are related to the current study; and on the use of digital image databases or atlases All these several factors are also considered in relation to the current study.

A large number of studies used email for transmitting images (SAF method) as stated by Buckley, Andelson and Agazio (2009) and Moreno-Remirez et al. (2007). However, other studies such as Tsai et al. (2004) and Kim et al. (2003) used the internet to send images to a database where a dermatologist could access and view the images in order to provide diagnostic reports and medical advice.

There are many types of devices and methods that may be used to capture, transmit, receive, display and store digital images. This is understandable with the development of technology and systems. However, it is imporant that, at both ends of a telehealth link, the technical requirements for ensuring a minimum acceptable image quality are recognised and supported.

Some studies such as Tsai et al. (2004) have used a standard protocol in their diagnostic process. In the first study for example, they used the same type of camera and same type of display monitor (HP 1740 flat screen 17 inch with display resolution set to 1,024x768 using Microsoft Picture Manager. However, the quality of the images that were taken by the study staff was higher than the images that were taken by the patients, which perhaps was because of the level of training or experience.

Different photographic techniques and different images are applicable to different types of people and for different types of pathologies. For example, in the case of cellulitis, an expert needs to see particular details or symptoms that are important for diagnosing the pathology such as redness intensity and area of spread. Ciocca, Marini and Schettini (2009) supported this, saying that an image that is perfect for one application can be inappropriate for another.

According to an experiment conducted by Bittorf et al. (1997), the minimum resolution required for digital images used in dermatology is 768x512x24. This is the threshold resolution level for what can be accepted for dermatology. Current technology is more advanced than the technology of the time of the mentioned authors, as resolution can reach, in some cases, over 4,368x2,912x24. However, the study stated that resolution is the main factor that affects the diagnostic rating of the images. The quality also can be affected by the image content and it is recommended to focus on the affected part of the body during image capture. The resolution can also be affected by the distance between the image and the observer when viewing the image (Bittor Schuler and Diepgen 1997). In their study, eight selected images were assessed by six experienced dermatologists using a descriptive scale to rate the images' quality. The participants were also asked to do image matching by stating the level of similarity between the images in relation to their quality.

Rimner et al. (2010) evaluated the quality of digital images sent by 46 patients via e-mail for use in tele-dermatology diagnosis. The images were collected from the patients, stored in a database and then evaluated against a set of criteria which were drawn up for assessing image quality. The results showed that images sent by patients, given specified minimum resolution requirements of 800 x 600 pixels, were of sufficient quality for use in teledermatology diagnosis. This recent study supported the validity of tele-dermatology that uses skin images captured and sent by patients. Also it demonstrated the accuracy when setting a minimum resolution requirement. The current study considered these two findings in the experimental work.

The American Telemedicine Association (ATA) established technical and clinical specifications as guidelines for the minimum requirements for best practice in telehealth such as: 75 ppi (pixels per inch) as a minimum resolution for a digital device, 1:500 or higher contrast ratio (CR), and a monitor with a minimum of 0.19-dot pitch to be used (Krupinski et al. 2008), also they recommended some ethical considerations such as identifying all persons, including medical staff, in the examination room to all participants prior to the teleconsultation. As regards colour of skin pathology, Krupinski et al. (2008) stated that the recorded appearance of a colour may be influenced by lighting and background conditions. Krupinsky et al. also referred to standard protocols for using digital imaging in teleconsultations related to skin infections They recommended that any digital camera used for capturing images should have good colour acquisition and white point adjustment. In use, it should be well focussed and must be kept as steady as possible to avoid camera motion artefacts. They also recommended capturing both close-up and general images of the affected part of a patient's skin. The guidelines promulgated by the ATA highlighted the importance of colour clarity in digital images. They advised that, to obtain accurate colour, the acquisition device should allow for at least 24 bits of colour and automatically apply compression (e.g. JPEG, setting). The white balance and the lighting in the frame should be white or close to white and the subject should be illuminated to at least 150 ft/candles. Backgrounds should be plain and non-reflective. The guidelines are very useful as general recommendations for training of new staff in the field but they are very general and do not cover the different medical conditions and pathologies that might require specific technical and clinical arrangements in order to ensure that the required information is available in the images. On the other hand, in some cases, e.g. in tele-endoscopy, lower technical specifications are required for diagnostic accuracy (Ganeshalingam et al. 2010). This study considers ATA guidelines in the initial setting before looking into the specific requirements. Also a telehealth link needs to be standardised at both ends in such a way as to ensure the quality of the images and medical data represented in them. The current study used the same technical standard in all experiments which involved capturing images of cellulitis cases.

Kim et al. (2003) and Stanberry (2000) stated that further study is needed to test whether differences in disgnoses can be traced to technology or system failure or whether it is due to the absence of clear standards of diagnostic criteria between dermatologists in face-to-face methods. This supports the current study in suggesting the need for a colour standardised chart or scale to aid in diagnosing cellulitis based on a solid theoretical foundation.

A study by McNeil et al. (2002) investigated technical calibration and standardization between the two ends (patients or GP and expert consultation room) for image quality assurance in teledermatology. They used simple standardised systems to check lighting and the colour of the computer monitor using a standard colour display system, called colour cathode ray tubes (CRT) and a physical colour checker chart. They used these tools to match the set up between the two ends before the diagnostic process started. They added that the colour temperature should be, at both ends, between 3200 and 3500 Kelvin and the background of the rooms should be painted with a matt light blue latex paint. They found high image quality and high accuracy of diagnosis in their experiment when they compared the results of diagnosing 309 live dermatology patients by one of three dermatologists with the results of diagnosing 309 digital images of the same cases by one of the same three dermatologists. They used a subjective confidence rating as indication for the accuracy of the result and found it to be high and to have a positive correlation with the accuracy. It was noticed by Krupinski et al. (2008) that, despite the ATA guidelines on minimum technical specifications for telehealth image acquisition, every study tends to use its own technical set up and local technological standard. The present study focused on consistency, which meant using the same standard during all the stages of the protocol in digital imaging. This implied, for example, using the same camera to capture all the images used in the study. Such technological control during a teleconsultation in very important for avoiding errors.

Massone et al. (2009) reported on the utility of mobile phones in capturing and transferring digital images in a telemedical system that communicated via internet and satellite links. Recent similar studies conducted by Wurm and Soyer (2012), supported these findings. The quality of current generation mobile image capturing capability is high and quite acceptable. Since the time (2009) of author's reporting about the impact of this technology, image processing software and the capabilities of mobile phone cameras have improved dramatically. This is very useful, especially in follow up cases. The study mentioned here was related to cancer but is applicable to any disease. A 91% concordance between face-to-face diagnosis and remote diagnosis was reported by the same researchers for teledermoscopy, with images sent via email or by uploading to a dedicated web application. In another study by them, 83% of diagnostic accuracy was achieved for telehealth-based diagnoses of melanocytic neoplasms in comparison with conventional histopathological diagnoses. Another recent comparison study by Tran et al. (2011) evidenced that current generation mobile phones used by patients, in their case with 5-megapixel camera, can produce high quality images. In Tran et al.'s study 30 patients sent self-captured images to dermatologists by store and forward teleconsultation. Tran et al.'s investigation revealed a 75% agreement mean with face-to-face consultations of the same thirty cases.

Another recent study by Sikka et al. (2012) confirmed the benefits of using mobile phone digital images in teleconsultations between patients and medical experts. They found in their experiment that the quality of 87% of mobile phone images out of 94 cases were accepted by doctors for diagnosis purposes. However, they concluded that even though most current generation mobile phones are capable of producing good quallity images, the quality is variable and can lead to incorrect evaluation. Berndt et al. (2012) also conducted experiments in Germany about the effectiveness of smartphones in transmitting medical images self-captured by skin disease patients to their doctors for follow up teleconsultation. They used a web-based portal to administrate the medical data and patient records. They found that using such technology has high potential in teledermatology for both patients and healthcare providers. Cavalcanti and Scharcanski (2011) used standard cameras and standard illumination in their experiment. The current study used the same method. However, in real life, patients from different backgrounds and locations use different cameras or mobile phone cameras in different capturing environments. The received images do not present consistency in quality. Crucial medical information may get lost in some cases as a result of poor quality or low clarity. The current study tested the use of an image quality scale to assess images before using them in teleconsultation for diagnosis.

2.7.3 Using colour scales in medicine

Diagnosis can be subjective (i.e. based on clinical experience, knowledge and perception) or objective (i.e. when there are standardised clinical tests such as laboratory). The benefit of using a diagnostic scale is that it provides doctors with an objective measurement anda reference with which to assess clinical cases. This can be coupled with their clinical experience to reduce the number of errors (mainly of misdiagnosis). There are several pathologies that are diagnosed in a more subjective manner such as; lupus vulgaris, toxic erythema, TENS, cellulitis, psoriasis, candidiasis, phototoxic rash and conjunctivitis. For this thesis, cellulitis and conjunctivitis were chosen as common examples. The diagnosis of pathologies with redness such as cellulitis and conjunctivitis is predominately subjective. It is subjective because of differences in clinical experience and doctor perception. Thus, there should be a drive to make such diagnosis more objective. The following section provides a brief of key previous studies showing the importance of using colour scales in medicine in order to provide more objective diagnostic outcomes.

Previous studies by Kahn et al. (1975) and Terry et al. (1995) have suggested that using medical images to standardise diagnostic assessment can improve accuracy of record keeping and also minimise subjectivity when compared to verbal descriptions with an arbitrary scale for severity.

Historically, McMonnies and Chapman-Davies (1987) were the first to formulate a photographic scale assessing ocular responses to contact lens wear. This scale laid the foundations for future scales and demonstrated that usage of their scale exhibited high levels of inter and intra observer reliability. Many scales often use five steps to represent differing scales of severity and often these reference images are selected based on clinical experience or subjective judgement. It is this subjectivity that is often criticised as it can result in unevenly spaced reference points, which do not represent the full range of possibilities. In the same way, dermatological tests require the application of the practicioner's subjective assessment, as opposed to objective tests, such as haemoglobin count, to which a measured value can be applied. Several studies have investigated the effects of incremental steps between the reference images of a grading scale on observer reliability, between other observers for a repeated task and for the same observer during varying conditions. Rheingans (2000) mentioned five considerations when designing any colour scale. These are goals, nature of data, audience, overall visualization, and cultural connotations. These considerations are relevant for the current study, for example: the diagnostic goal, colour red intensity, and clinical background of the operators.

Schulze-Wollgast, Tominski and Schumann. (2005) mentioned three factors that influence the choice of colour scale when presenting data. These factors are: type of data, visualisation goal, and general context which includes user colour perception, output device and user preferences. Based on these criteria they developed a method for choosing appropriate colour scales automatically. They used the colour scale to represent data, which includes any type of data. In contrast the current study used the red colour scale to represent the degrees of intensity in an image. Furthermore they used the scale to visualise data quantitatively while the current study used the red colour scale for interpretation and diagnostic purposes. The aforementioned factors also were also emphasised by another study by Al-Hindawe (1996) who explored the challenges when designing scales. They focused on designing a semantic differential scale (SDS) which is used for measuring social attitudes especially in linguistics and sociology. SDS uses a number of divisions presenting the measured quality. The number four in the scale represents a neutral position when users are not sure about their answers. The study designed SDS to measure human personality features such as honesty. They explored five, six, seven, and nine point scales and suggested that a seven point scale is the best for such an area as it is not too big or too small and it includes a neutral choice. They added that the nature of data and subjects are important for making such decisions, which can be critical considerations for easier analysis and interpretation of results. A similar study by Heise (1970) investigated the use of SDS in social psychology research for the analysis of people's reactions to stimuli, words and concepts using a sematic differential scale. Heise's experiment asked people of different cultures about their attitudes to certain issues. The study developed a scale for the measurement of such subjective attitudes. The result suggested a cost-effective, easily set-up and reliable differential scaling system. Flavell and Heath (1992) investigated the impact of the number of divisions in a perceptually uniform colour scale. They found that the rate of errors increases with an increase in the number of divisions. The haemoglobin red colour scale is a typical example of such diagnostic methods. This was explored by

Anand and Sexena (2009) and Critchley and Bates (2005). It is a linear scale with six shades of red representing haemoglobin levels between four and fourteen. These levels in the haemoglobin colour scale refer to red intensity where a higher the intensity corresponds to a higher blood count.

Using colour scale for cellulitis diagnosis

Vu et al. (2003) developed a clinical scale and severity index for preseptal cellulitis in children. They created a single scale to describe different levels of severity when diagnosing children aged between one and sixteen with cellulitis. Vu claimed that when they started their attempt, there was no standard objective clinical method available to rate the severity level of preseptal cellulitis and its response to treatment over time. However, there were many non-standardised methods used. The key contribution by Vu incorporates the main symptoms of cellulitis in one single scale which includes local features such as location, erythema (redness), tenderness and pain. It also includes the systemic features such as interaction and fever. Each feature was divided into three levels. For example, erythema was divided into minimal, red and ecchymotic.

The scale was developed by a group of four experts. The validity and reliability of the scale were evaluated after using the scale in assessing seventeen case studies of cellulitis. They reported another experiment of 104 photographs showing cellulitis assessed by eight medical experts, ophthalmologists, emergency doctors and general practitioners who were not involved in the care of those children. They found high accuracy compared to the results of a global score, which is widely used for measuring impressions.

In a third experiment conducted by the same researchers, they took instant photographs for all the case studies and asked three ophthalmologists to rank them using their severity index. They found that there was a moderate correlation between their results and the results of other experiments when using the index and global score (Vu et al. 2003). They explained that the moderate results were because the experts could not recognise all the medical features such as fever, pain, and tenderness.

The development of redness needs to be measured and monitored over time in relation to its size and intensity. Other medical conditions are well-defined and can be diagnosed by GPs and nurses in face-to-face consultation, or by the specialist during the real time teleconsultation. Also in some cases it can be reported by the patient or carers. Vu et al.(2003) stated that their scale can be used as an objective and accurate tool to diagnose children with this condition and also can be applied to other medical conditions. They added that this tool can help any healthcare professional or new practitioner with basic clinical experience, as it is easy to follow and was tested with medical doctors and professionals of varying degrees of experience and not only experts.

Using a scale for diagnosing red eye (conjunctivitis)

Transmitting and diagnosing medical images has had a long history in tele-ophthalmology. In 1987 retinal vessels were monitored during space flights and the images were transmitted electronically for consultation by NASA (Li 1999). Since then, store and forward teleconsultation has been used for diagnosis in the USA and Australia. Tele-ophthalmology became a reality and started to play a critical role in the lives of patients and healthcare professionals (Li 1999).

Eye redness scales have been used in clinical eye diagnosis. The red eye grading scale usually has four or five levels (Murphy et al. 2006). Recording and measuring redness is an important part of the diagnosis (Li, 1999). Ophthalmologists in the eye unit at ARI use a red eye scale of five levels, called a grading guide, shown in the figure 2.13.

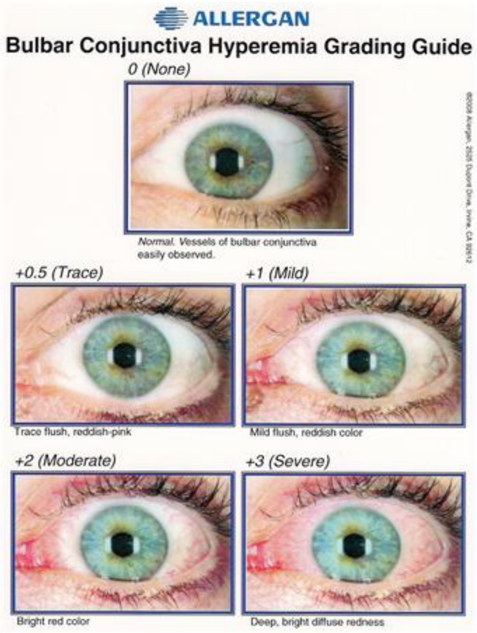

ALLERGAN
Bulbar Conjunctiva Hyperemia Grading Guide
0 (None)
Normal. Vessels of bulbar conjunctiva easily observed.
©2008 Allergan, 2525 Dupont Drive, Irvine, CA 92612
+0.5 (Trace)
Trace flush, reddish-pink
+1 (Mild)
Mild flush, reddish color
+2 (Moderate)
Bright red color
+3 (Severe)
Deep, bright diffuse redness

Figure 2.13: Red eye grading guide designed by Allergan and used by the eye unit at ARI

Allergan is a global health company based in the USA which provides eye-care products and other speciality pharmaceuticals (http://www.allergan.com, accessed on 8th of Oct.2013). ARI utilises a chart produced by Allergan as a grading guide for bulbar conjunctiva hyperaemia. The Allergan grading guide uses five images (i.e. it is a discrete scale) to aid the health professional to assess the severity of the conjunctivitis. None (0) is an image of a “normal” eye and is used as reference for the subsequent images. This first image illustrates how the vessels of the bulbar conjunctiva are easily observed. The second image is an example of trace hyperaemia which is reddish-pink in colour. The third image is of mild hyperaemia and the image indicates a more reddish colour with mild flush. The fourth image illustrates moderate hyperaemia and the eye is described as being bright red. The fifth image is of severe hyperaemia and the description, deep bright diffuse redness is used. This scale was used by those participating in the experiments of this thesis.

Doctors in hospitals work in shifts and monitor the development of redness in a particular case by comparing the pathological symptoms each day. When doctors change shifts the new incoming doctors look into the records and see the notes of their colleagues, but these notes do not have common language or reference points but rather personal and subjective observations that differ from one doctor to another based on their knowledge and clinical experience.

A study by Schulze, Jones and Simpson (2007) investigated the development of a bulbar redness scale using digital conjunctival hyperemia photographs obtained via a photo-slit lamp of controlled exposure. The system used a scale with a range from 0 to 100 to compare the chromaticity of spectrophotometer images. This scale was found to perform well with high statistical results. This grading scale system can be used for patient monitoring and classification of images for measuring the severity of pathologies. Other scales of different reference points were also investigated and compared by the study. They stated that the level of detail required from the observers decides the number of divisions that should be in the scale when it is developed. They found that using a 5-step interval scale appears to be the most reliable for clinical assessments of bulbar redness. A similar study by Schulze, Hutchings and Simpson (2009) investigated the usage of a psychophysical scaling method for the analysis and estimation of perceived redness against reference images of different validated bulbar redness scales of four, five and six reference levels. The reference images used in this study were cropped and separated into 3 categories based on their colour information, greyscale and digital information expressed in binary digits. The images were analysed using the physical attributes of the image on a scale ranging from 0 to 100 points. All the three tested scales were found to be effective for the analysis of redness. Fieguth and Simpson (2002) conducted an experiment to examine the relationship between physical image characteristics and the clinical grading of images for conjunctival redness. Thirty images were graded by 72 clinicians using a 100-point scale with three reference images at 25, 50, and 75 points. The outcome of this experiment showed clear inconsistent grading with a range of differences in the judgments of participating doctors. The study also developed another automated method to measure the redness of the bulbar of the eye as well as the number of red vessels. They used edge detection and colour feature extraction techniques. However, their edge detection was done manually using the Canny edge detection operator, which made their method not fully automated and involved human intervention. They used lossless compression methods in order to ensure that the important clinical information in the images was retained. However, understanding the information that must be captured in an image leads to a more effective compression method. They found that using automated methods to measure the redness of the eye had similar results to the clinical grading but with more accuracy and less variability.

Unlike the Fieguth and Simpson (2002) study, the current study used descriptive as well as numeric colour scales for clarity and accuracy. Also the current study involved non-health workers in the experiments, and not only clinicians, in order to compare the results as well as testing the applicability of the system to a wider community of non-doctors. The current study, unlike the Fieguth and Simpson (2002) study, where only one task was involved, investigated different tasks including describing, grouping, ranking, comparing and matching redness in digital images with and without using colour images.

A study by Abelson (2010) explored the importance of the colour red characteristics for doctors when diagnosing eye pathologies. The hue or depth and intensity of the colour red as well as the location of the vasodilation are key features for diagnosing eye pathologies including conjunctivitis, which was the focus of this study. This information was also supported by Rodriguez et al. (2013). Abelson elaborated that the feature of the colour red progress and variations in the level of intensity over the time it takes the appearance of mild red shade in the case of dry eye, mild diffuse pink in allergic conjunctivitis as shown in the figure 2.14 below, and different features of redness in bacterial conjunctivitis, developed over time.

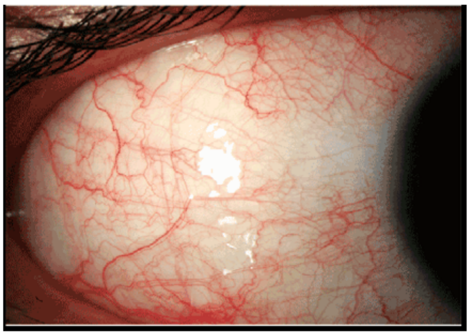


Figure 2.14: Mild diffuse pink in allergic conjunctivitis, by Abelson (2010)

Using a diagnostic scale provides doctors with an objective measurement and references to assess clinical cases. This can be coupled with their clinical experience to reduce the number of errors (mainly misdiagnoses). Cellulitis and conjunctivitis were selected as the two main examples because the former does not have a standardised diagnostic scale and hence is more subjectively diagnosed, whilst for the later there are several diagnostic scales and thus a more objective diagnosis can be made.

2.8 Auto-classification Techniques for Images

Reason (2006) stated the importance of automation not only to save time and resources but also to avoid operators' errors and difficulties such as high workload and high stress. However, he also mentioned the critical role of skilled and well trained operators who are needed in the event of a system failure or emergency. Such high level of skills and experience can be obtained by training and by practice during regular testing, monitoring and reviewing of the performance of automated systems (Reason 2006).

Automation is commonly used in many real life applications. Pandey, Deen and Pandey (2013) have listed, in their comparative survey, 19 different studies of automated image classification based on colour content. Machine learning (ML) is one of the techniques that is used in the auto-classification of images. The following section provides a brief about ML and its key role.

Successful image classification depends on several factors: high quality images, the design of the classification system and the skill of the analyst. They stressed that one of the most critical aspects is identification of the classifier as there is a lack of guidance as to the selection of the most appropriate algorithms. This raises a particular challenge to the creation of an auto-evaluation system for use in a telehealth context and demonstrates that considerable extra research needs to be carried out in this field (Lu and Weng, 2007)

2.8.1 Machine Learning

The use of ML as an image classification technique has spread rapidly in the last decade in computer science and other fields (Domingos 2012; Mitchell 2006). ML can play a key role in various applications such as natural language processing, computer vision, speech recognition, robot control, web search, spam filters, credit scoring, fraud detection, stock trading and drug design (Domingos, 2012). It also provides potential solutions in the area of image recognition and classification, which applies to the study of this thesis. Samuel (1959) stated that "ML enables computers to learn with experience and without being clearly programmed". Mitchell (1997) added that experience improves computers and produces more accurate outcome when there are performance measures for specific tasks. Furthermore, ML provides self-monitoring and self-diagnosis and self-repair systems (Mitchell, 2006).

ML provides solutions for many real-world problems of a complex nature which cannot be solved by numerical means alone, such as predicting rocket engine explosions, future traffic patterns from past traffic patterns, cancer patterns, and the market value of a house. ML has been applied to each of these areas with great success. Furthermore, ML improves the efficacy of systems and the design of machines (Kotsiantis, Zaharakis and Pintelas 2007). The algorithms that are used in ML are able to calculate how important tasks can be performed by generalising from previous examples which make it more feasible and cost-effective. However, the accuracy of the output depends on many factors such as the quality and amount of data and the chosen algorithm (Domingos, 2012). Computing science and statistics play a key role in shaping and developing ML by providing new ideas for viewing and progressing human learning (Murphy 2012; Mitchell 2006).

Supervised (or predictive) ML and unsupervised (or descriptive or knowledge discovery) ML are the main commonly used classifiers in ML. In supervised ML, an accurate conclusion can be reached by a predictor function which compares new data with a pre-defined set of training examples and uses techniques such as the nearest neighbour matching algorithm in order to build brief but accurate models of the distribution of class labels in relation to predictor features (Kotsiantis, Zaharakis and Pintelas 2007; Mitchell 2006). In unsupervised ML, a data set is used to discover unknown but useful patterns or relationships or classes of items intuitively. In other words, output is expected without any input (Murphy 2012).

Reinforcement learning (RL) is another type of ML but it is uncommon. This technique uses signals of reward or punishment when certain acts or behaviours are conducted (Murphy 2012). RL as well as unsupervised ML are out of the scope of this study. A supervised K-nearest neighbour classifier was applied in the experimental work of this thesis, training the model with labelled images, and testing using new data with the k-fold cross validation method. Rtsch (2004) discussed several options for the construction of classification algorithms such as linear discriminant analysis, decision trees, neural networks, and K-nearest neighbour classification. The latter of these was the option taken for constructing the algorithm used in the auto-evaluation system of this study. This classification technique was chosen because it is the simplest method, it is intuitive and produces low classification errors.

A further consideration in the construction of an image classification algorithm is the colour space. It is common in image capture and transmission to use the RGB colour space. This however can have disadvantages in classification. An alternative is to use the HSV colour space which is more intuitive (and therefore more appropriate to learn and compare system), and is unaffected by changes in illumination and camera direction (Sergyn 2007). There exists a standard formula for the translation of RGB colour space into HSV. It was decided, therefore, to use HSV in the construction of the algorithm for the auto-evaluation system. There are many challenges facing ML, for example: keeping the balance between data privacy and data exploitation, transferring the learning experience from one task to a different but related one and building never ending learners (Mitchell, 2006). Other challenges can fall loosely into the categories of correct algorithm construction, data storage requirements and the quality of data gathered (Domingos 2012).

2.8.2 The use of image databases

The method used by the following key previous studies in using image banks or databases is closely related to the research problem with regard to image comparison and retrieval pre-diagnosis. A related study by Tsai (2007) investigated the performance of image classifiers. They created a classifier by training it with texture features (related to scenery) that were extracted from a set of training images and then encoded as feature vectors. The texture features were each associated with a class label. The trained classifier was then able to classify unlabelled new images represented by their texture features as shown in the figure 2.15.

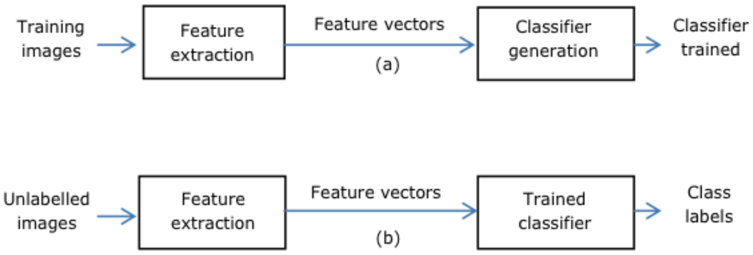


Figure 2.15: (a) Training stage and (b) Classification stage. Tsai (2007)

Another study by Chang et al. (2004) also used an image feature extraction method from a database that stored labelled JPEG compressed images. These original images were degraded to create a series of images that were ranked in a scale based on characteristics of the extracted features. The query image was compared to partially-decoded labelled images retrieved from the database according to the features detected in the query images, and ranked based on the similarity of the features detected in the query image to the features in the retrieved images. The current study proposed a similar method in building a database that stored images of red eyes that were ranked in a scale based on the intensity of their redness. New unlabelled images received by the telehealth system would be ranked and labelled based on the similarity of their features to images stored in the database.

Reason (2006) highlighted information retrieval based on common criteria or features. He named it similarity-matching (called conditions). Brilakis and Sobelman (2005) also showed how content-based methods of image search can expedite automatic retrieval and indexing of images from image databases. Such methods employ pattern recognition algorithms to retrieve similar images.

Nakazato, Manola and Huang (2003) proposed an interface for image retrieval from image databases that used the concept of "image groups" to retrieve similar images in groups instead of single images.

A study by Haeghen et.al (2000) discussed the use of digital imaging systems in dermatology by focusing on colour image acquisition and calibration. An experiment was conducted using dermatology images tested against a standard colour imaging system via RGB. This experiment used a digital camera for image capturing and a computer based imaging system for storing the image characteristics of a set of reference images. The results suggested the effectiveness of using an imaging system for the development and classification of images in dermatology.

Cheng and Chen (2003) examined the extraction of colours, textures and regions from images and stored these features in a searchable index. This index was used in extracting images from an image database based on chosen colour, texture and shape features. Query images were then matched against this set of extraced images. This approach was found to be effective and can be used for image classification and retrieval in telehealth.

Views differ as to the principal challenges to the transmission of images for telehealth purposes. Dixon, Hook and McGowan (2008) regard security, image resolution and technical support as the main challenges.

Kuntalp and Akar (2004) however, stated that, in their opinion, the two main challenges in receiving images by any telehealth system are: the size of the images and supporting different image file formats.

The first problem can be resolved with large sized storage devices and databases where these images can be stored. The second problem can be resolved by a system that has the capability to recognise different types of image file formats automatically and change them to a standard compatible image format.

2.8.3 Quarantining the region of interest and extracting colour features

Processing of medical images to isolate important areas of interest is essential in medical imaging because the medically significant information in images is often subtle and the automated processing and extraction of the salient features from them is complex. This isolation will eliminate irrelevant data and reduce noisy data, however achieving this presents challenges.

Wolffsohn (2004) showed that using edge detection and colour extraction as examples of objective image analysis is more repeatable than using subjective matching by eye using colour scales. Many other studies have explored region of interest extraction in image processing such as Rodriguez et al. (2013), Bista et al. (2013), Shao-zhen and Xian-dong (2010), Zhang and Xiao (2008) and Cheng and Chen (2003).

Rodriguez et al. (2013) used colour feature extraction and image edge detection techniques in order to diagnose dry eye redness. They used a scale ranging from 0 to 4 (the Ora Calibra Dry Eye Redness Scale [OCDER], an established clinical scale for dry eye diagnosis) in order to develop and test an automated computer redness grading method for objective diagnosis. They defined the degree of redness by intensity, size of the location and the prominence of horizontal conjunctival vessels and suggested that the prominence of vessels increases with increase in redness intensity. Figure 2.16 shows the region of interest (ROI) extraction. This involved three stages. Firstly, the region of interest was highlighted, as shown in the top image. Secondly, the degree of redness was determined, by intensity, size, location and prominence of horizontal conjunctival vessels. Thirdly, measurements were made of the prominence of the vessels.

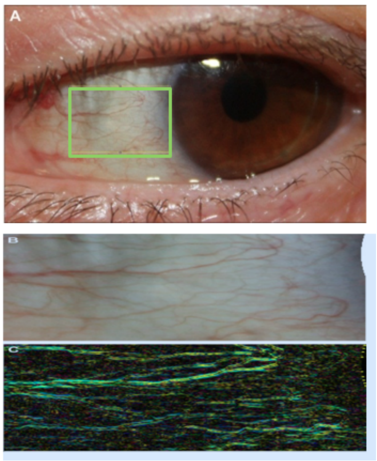


Figure 2.16: Region of interest selection by Rodriguez et al. 2013: (A) Dry eye image. (B) Region of interest (ROI) showing pathological redness. (C) Vascular structure in the ROI showing the prominence of graded horizontal vessels

Zhang and Xiao (2008) also used image feature extraction and feasibly proposed a fast contrast-based method for extracting regions of interest in biomedical images, in their case of MRI scans of brain tumors. Cheng and Chen (2003) used a method that extracted features, derived dissimilarity metrics, and then classified query images based on those metrics. In the study, image-based features extracted included the colour histogram, which was directly extracted from a quantized RGB image.

Peterson and Wolffsohn (2009) stated that the combination of redness intensity, vessel edge detection and prominence measurement provides more accurate diagnosis. However, a study by Maenpaa (2004) discussed the classification of images by extracting their colours and texture information.

The study compared between extracting colour and texture together or separately. The study found that they should be treated individually as they are perceived and processed separately by the human visual system.

2.8.4 Image Descriptors and Colour Histograms

This section briefly describes colour histograms as a well-established example of an image descriptor.

Histograms generally have been a widely used method in image analysis and processing. However, colour histograms specifically have been used and explored in many studies of image processing in recent years (Paschos and Petrou 2003). From the date of the previous study, it can be understood that colour histograms are a relatively new method in the field.

A colour histogram is a colour descriptor, which represents the colour intensity distribution in an image. As with any colour descriptor, creating a colour histogram requires two processes: a feature extraction algorithm and a matching function. Feature extraction does image mapping in the feature space.

The matching function returns a measure of the similarity of two images based on their colour features. It compares between a query image and a training image retrieved from a database. It counts similar pixels that have the same colour features and stores them in order to describe the total number of pixels for each specific colour independently. There are two types of colour histogram: a global colour histogram (GCH) and a local colour histogram (LCH). The following section describes these two types separately.

1. Global colour histogram (GCH):

GCH is the most common and best known type of histogram. GCH depicts the overall distribution of colour intensity in an image as a whole without specifying which part (blocks or pixels) of the image has this olour. It is used to detect similar images based on their colour features. It is a general comparison between the total colour intensity of an image with the total colour intensity of another image.

- **Feature extraction algorithm of GCH**: First, the colour-space is discretized into n colours. The size of n will usually be determined by the number of colours that an image format supports. The number of pixels of each colour is then counted and stored.
- **Matching function of GCH**: The most common matching function (although there are many others) for this method is Euclidean distance. To compare two images A and B the following equation is used:

$$D(A,B)=\sum_{i=0}^{n}\sum_{j=0}^{n}\sum_{k=0}^{n}(A(i,j,k)-B(i,j,k))^2 D(A,B)=\sum_{i=0}^{n}\sum_{j=0}^{n}\sum_{k=0}^{n}(A(i,j,k)-B(i,j,k))^2$$

Here, $A(i,j,k)$ = the number of pixels that have colour (i,j,k) in image A. Where i= red component, j= green component, k= blue component . D = Sum of Euclidean distances. The larger the distance value, the less similar the two compared images are.

2. Local colour histogram (LCH)

LCH provides details on the distribution of the colours in discrete regions of an image. LCH is similar to GCH but it compares colours in two images region by region. Each region in one image will be compared with the same corresponding region in the other. LCH gives the position of the colour of the image while the GCH gives the total colour intensity of the image. A difference value is obtained by computing a GCH for each region. The total distance between any two images is the sum of all GCH distances between them. The larger the distance between the two images, the less similar they are.

- **Feature extraction algorithm of LCH**: The image is split into regions. Then a GCH is computed for each pair of regions.
- **Matching function of LCH**: Images are matched by comparing the sum of the

GCH values for all the regions in the images. In the context of red eye diagnosis, GCH can be used for comparison purposes between received query images and training images in the database. LCH can be used to provide more details on the location of the colour intensity in the image if it is needed. Sometimes LCH is more efficient than GCH. However, if the image is rotated, the output might be different.

Paschos and Petrou (2003) have used colour histograms in colour texture classification. Colour histograms can have drawbacks, for example, when two images with the same histogram have different colour distributions. This can be a limitation in content based information retrieval (CBIR) systems that have a large number of database images. In such cases, a collection of images can end up having the same colour histogram resulting in use of the wrong database image to match a query image. To mitigate this, the concept of combined histograms is employed, which makes use of statistical information about the combined colour components in order to determine whether two images are similar or different in terms of the colour distribution (Pass and Zabin 1999).

2.8.5 Colour moment

Colour moments are image descriptor measures that compare images based on their colour similarity. In this case, the colours in an image are separated into colour channels. For each colour channel, the mean, variance and skew are computed thus producing nine moment values The combination of these moments is a good descriptor to differentiate between the colour distribution of an image (Stricker and Orengo 1995; Keen 2005; Singh and Hemachandran 2012).

Matching function Let the colour channel at the image pixel be defined by . The three colour moments then can be defined as:

Mean:

$$E_i = \sum_{j=1}^{N} \frac{1}{N} p_{ij}$$

The mean can be understood as the average colour value in the image.

Standard deviation:

$$V_i = \sqrt{\left(\frac{1}{N}\sum_{j=1}^{N}(p_{ij} - E_i)^2\right)}$$

Standard deviation is the square root of the variance of the distribution.

Skew:

$$S_i = \frac{E_i - M_i}{V_i}$$

Such that V_i is the standard deviation, E_i the mean and M_i is the median. Skew can be understood as a measure of the degree of the asymmetry in the distribution. A function of the similarity between two image distributions is defined as the sum of the weighted differences between the moments of the two distributions. In other words, to compare two images (a, b) it is necessary to sum up all the computed CH using following equation: Such that is the standard deviation, the mean and is the median. Skew can be understood as a measure of the degree of the asymmetry in the distribution.

A function of the similarity between two image distributions is defined as the sum of the weighted differences between the moments of the two distributions. In other words, to compare two images (a, b) it is necessary to sum up all the computed CH using following equation:

$$D(a,b) = \sum_{i=0}^{r} w_{i1}\left(\left|a_{e_i} - b_{e_i}\right|\right) + w_{i2}\left(\left|a_{V_i} - b_{V_i}\right|\right) + w_{i3}\left(\left|a_{s_i} - b_{s_i}\right|\right)$$

Here, N = number of colour channels
E_i = Mean of colour i
V_i = Standard deviation of colour i
S_i= Skew of colour i
D = Similarity distance
W_{ij}= User specified weights (j= 1, 2, 3) for each colour i

This project focused on the redness occurring in cellulitis and conjunctivitis as medical symptoms. This red eye is different from the so-called red eye produced by a camera flash in digital photography as this affects only the iris. Studies in detection and correction of red eyes in digital photography can be found in (Lepisto, Launiainen and Kunttu 2009; Mufit 2008). The project proposed no human interference during the auto-diagnosis procedure in order to avoid any possible subjective assessment. Rodriguez et al. (2013) stated that automated image analysis has the advantage of objective image judgement because it eliminates the variables of grader fatigue and innate comparative bias which may occur when grading a series of images.

2.8.6 Auto-evaluation system

The foregoing literature review makes it abundantly clear that key components in the further development of telemedicine are image quality, classification of redness intensity and accuracy of the transmission of images. The main aim of this thesis and its associated experiments was to improve understanding of how these elements relate to and affect human perception of colour in images. As was mentioned in the introductory chapter, a side issue of the study was to aid in the development of an auto-evaluation system to assist in the diagnosis of pathologies which present colour changes as a key symptom, the aim being to minimise human input (together with its associated errors) from the early stages of the diagnostic process. Fig 2.17 below demonstrates how the thesis would relate to key components in the formulation of an image evaluation protocol.

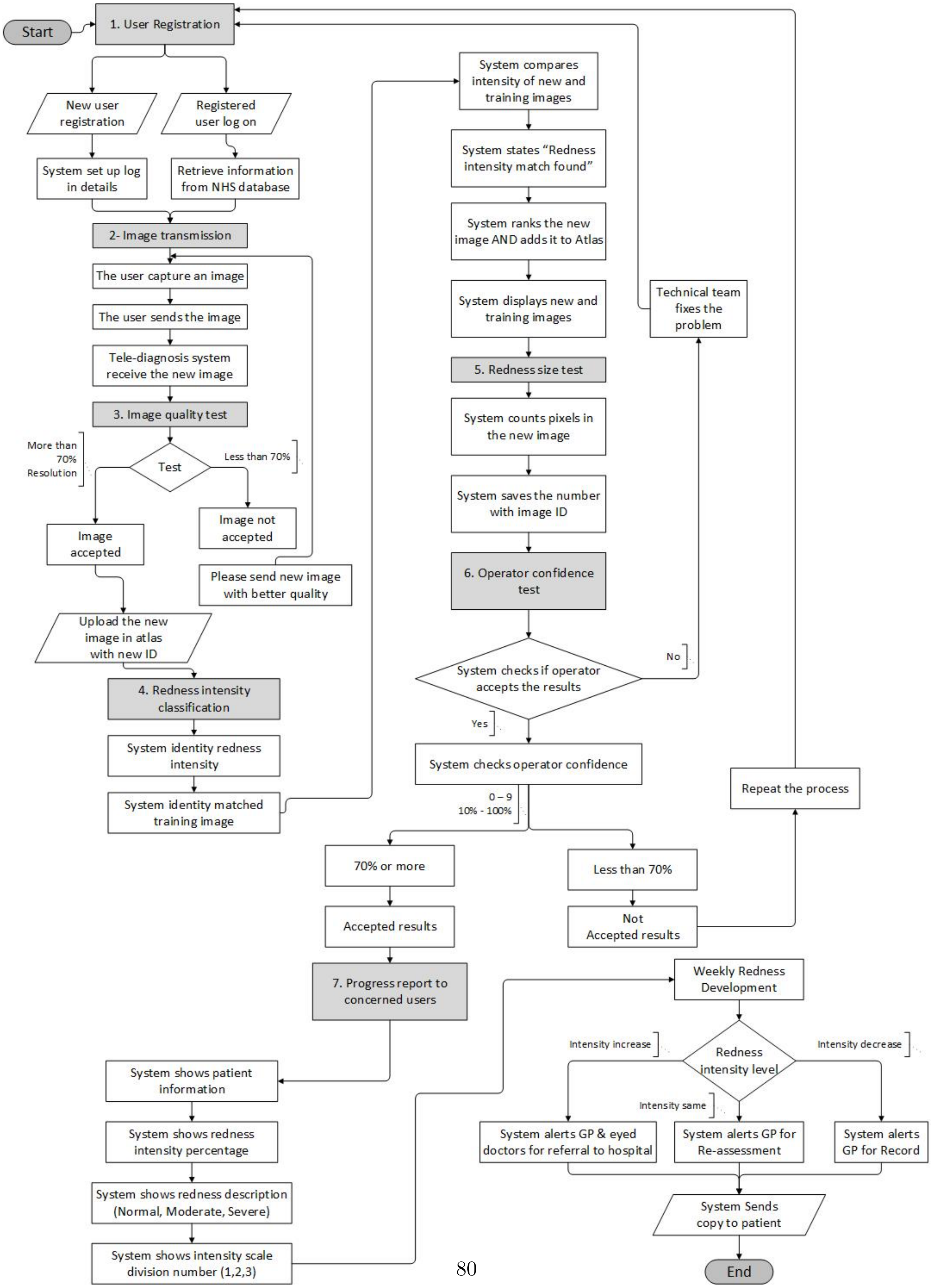

Start
1. User Registration
New user registration
Registered user log on
System set up log in details
Retrieve information from NHS database
2- Image transmission
The user capture an image
The user sends the image
Tele-diagnosis system receive the new image
3. Image quality test
More than 70% Resolution
Test
Less than 70%
Image accepted
Image not accepted
Please send new image with better quality
Upload the new image in atlas with new ID
4. Redness intensity classification
System identity redness intensity
System identity matched training image
System compares intensity of new and training images
System states "Redness intensity match found"
System ranks the new image AND adds it to Atlas
System displays new and training images
5. Redness size test
System counts pixels in the new image
System saves the number with image ID
6. Operator confidence test
System checks if operator accepts the results
No
Technical team fixes the problem
Yes
System checks operator confidence
0 – 9
10% - 100%
70% or more
Less than 70%
Accepted results
Not Accepted results
Repeat the process
7. Progress report to concerned users
System shows patient information
System shows redness intensity percentage
System shows redness description (Normal, Moderate, Severe)
System shows intensity scale division number (1,2,3)
Weekly Redness Development
Intensity increase
Redness intensity level
Intensity decrease
Intensity same
System alerts GP & eyed doctors for referral to hospital
System alerts GP for Re-assessment
System alerts GP for Record
System Sends copy to patient
End

Figure 2.17: Red eye auto-evaluation system

2.9 Critical Discussion and Conclusion

This section explores the knowledge gap in the previous studies and the contributions of the current study toward each challenge.

Using telehealth opens an opportunity to many people in remote areas to access healthcare. Telehealth can also save resources for healthcare providers. Most of the previous studies supported the use of digital images in telehealth and demonstrated their effectiveness as an alternative to traditional face-to-face diagnoses.

Even though there are many technical and logistical differences between real-time and SAF teleconsultations in telehealth, the main purpose of both is still the same, which is the provision of healthcare services with the best possible quality patient care, especially in remote areas. The idea of this study can be applied to both methods even though the study focuses only on SAF teleconsultation.

Most of the previous studies referred to traditional face-to-face diagnoses as the gold standard against which to measure their results. However, this assumption is not always correct as doctors disagree with each other in diagnosing different types of pathologies especially if the diagnosis is done by individual doctors and not by a medical team. This in fact is the case for most diagnoses.

The two examples of pathologies proposed by the study are currently diagnosed face-to-face through direct physical observation. The current study proposed using telehealth in diagnosing these two pathologies. This idea can be applied to other pathologies where redness is involved. The study did not investigate the diagnosis itself but rather the description and rating of redness. This is one of the primary cognitive tasks that doctors carry out at the start of any typical diagnostic process when colour is involved.

There are similarities and differences between cellulitis and conjunctivitis in relation to redness classification and diagnostic process. In both pathologies, the colour red is one of the key symptoms and is usually assessed and described by doctors as part of the diagnosis. However, there is a well-established common language between doctors when treating red eyes symptoms. They recognise five different severity levels of the pathology at the developmental stage (normal, trace, mild, moderate, and severe) and they use a red colour guide with standard sample images that represent each level (Schulze Jones and Simpson 2007; Fieguth and Simpson 2002). Cellulitis does not have any of these methods, which makes diagnosing this condition more subjective and variable from doctor to doctor. A study (Vu et al. 2003) attempted to design a scale for all symptoms of cellulitis including location, erythema, tenderness and pain. Erythema (or redness) was divided into three levels on this scale (minimal, red and ecchymotic). The current study applies the same concept as Vu et al. 2003 with more focus on colour characteristics, mainly the intensity of redness which is recognised to be critical for classifying degrees of redness in digital images (Abelson 2010). The other key difference between the two pathologies is the clear white background in conjunctivitis cases in the bulbar conjunctiva (white area) of the human eye. This common colour background for all humans makes the diagnosis easier, unlike the case of cellulitis where the background varies based on the original colour of the skin which differs between individuals. For reasons of convenence during the analysis, the current study compares the results of cellulitis experiment sets with the results of conjunctivitis experiment sets.

The absence of a common language or of a colour scale for describing and rating the colour red as a key symptom during diagnosis leaves the door open for individual differences between system operators. This leads to differences in opinions due to misunderstandings and the use of subjective descriptions. This also justifies the need for a colour scale as basis for a common language between doctors and system operators.

Most of the studies such as (McIlvaine 2006; Reason 2000; Norman 1983) stated that human errors can be predicted if they repeatedly occur at certain times and in particular system states. Error prediction is the first step in the control of these errors and in designing a system that can obviate or minimise them. The current study counted the errors and circumstances of their occurrences in the conducted experiments.

From the literature, it is clear that the relationship between the confidence and accuracy of system operators vary between strong positive, moderate, poor and no relationship. There is also a conflict of opinion about the relationship between the time spent in performing a task and the accuracy, with most showing a positive relationship. The current study used a confidence scale to compare a participant's confidence level with their accuracy. The time taken for each task in each experiment was also recorded, as well as the total time taken for the experiment in order to investigate the relationship of time with accuracy.

The current study also tested theories described by studies such as Krug (2007) about overconfidence. This is usually concomitant with low accuracy while performing difficult tasks (Klayman et al. 1999; Mann 1993). In the same way, the current study investigated the concept of underconfidence, which is usually concomitant with high accuracy in easy tasks, and also tested the relationship between confidence and experience (Panduragan 2011; Dempsey and Burr 2009).

Even though Bittorf et al. (1997) stated specific minimum levels of image resolution needed for dermatology, the medical information required in digital images that are suitable for diagnostic purposes varies from one pathology to another. This requires further experimental research. None of the previous studies used any scale to measure the image quality received by the observer. However, some of these studies used the resolution to be the key measure that determined image quality. The current study designed and used an image quality scale based on the resolution of the image and also used a confidence scale when evaluating image quality. This was similar to an experiment conducted by Engelke Maeder and Zepernick (2009). According to that study, the level of accuracy and confidence was high in the extreme case of same or different images. Confidence was lower in cases in which the difference between images was not so immediately distinguishable. The current study investigated the level of accuracy and confidence when rating images that had different degrees of similarity.

The previous studies argued about the number of divisions that a scale should have (3, 5, 10, or 100). The current study used different scales based on the required information and data provided. For example for the image quality scale, the study used a scale of 100 divisions which represented all possible percentages from 0-100 of image resolution. Another example was the red colour intensity or redness severity levels scale where the study compared between five levels of numeric descriptive scale with standard sample images and with a three level scale of the same nature, and lastly used a scale of ten divisions to measure the confidence of participants. The previous studies confirmed two different critical issues in relation to human colour perception. Firstly, the colour cognitive mechanism, which is the same for all humans and functions in the manner detailed in the literature earlier in this chapter. This theory confirms that human colour perception is a generic process in all humans. This means that all operators of a telehealth system will perceive colour in the same way. Thus, the system can be designed in a way that suits human faculties and avoids or minimises potential errors arising from these. Secondly, there are different integrated factors that affect human colour perception in relation to colour, objects, image, and visual disorders of the eye, as detailed earlier in the chapter. These factors differ from case to case and apply to all humans to differing extents. These considerations imply that both doctors and non-doctors in telehealth are subject to errors within and between their populations. The current study used image quality scales, confidence scales, colour red intensity scales, and colour blindness tests to improve the current subjective diagnostic protocols. The study also used technical and technological standardisation during the experiments in order to avoid errors. The current study drew the following two important conclusions. Firstly, there is a need for improving the current subjective traditional diagnostic protocol. In traditional face-to-face diagnosis, clinicians see the affected part of a patient directly and relate their symptoms and characteristics to conditions seen in previous cases. In doing so the doctor applies their knowledge, clinical experience, and their ability to recall and compare the past with the current case. The accuracy of a diagnosis can thus be affected by a diagnostician's knowledge, experience, level of confidence and fatigue. The current system therefore investigated the impact of adding an image quality scale, a red colour scale and a confidence scale as supportive tools to make diagnosis more objective and to minimise errors that might potentially occur as a consequence of one of these diagnostic challenges. This is done by comparing the results of experiments conducted using the above scales and that of experiments conducted without using any scale. After this, the level of agreement between participants was measured for consistency.

As part of the thesis it was decided to create an elementary form of auto-evaluation system, purely to test whether, in its most basic form, the principle could be applied to, and tested against, the experimental results. Any future development of an operational system would need to take into account the following factors:

- **Isolating the region of interest**: Conventionally, doctors mark an area of interest with a marker pen. In automated systems, other techniques such as edge detection must be used.
- **Colour feature extraction**: In this case, extracting information about the intensity of the colour red in the area of interest of a query image.
- **Colour feature measurement**: In this case, assigning a numeric value for the aggregated redness intensity in an image. This may be expressed as a dimensionless quantity or as a percentage. This needs to be done initially to a database of training images and then applied in turn to query images.
- **Image comparison**: Making comparisons between the intensity values in a query image and those in a training image.
- **Image matching**: Matching a query image with training images featuring the same redness values.
- **Query image classification**: Ranking a query image against training images.

Chapter 3

Research Methodology and Methods

3.1 Introduction

The following is a brief description of the methodology used in the thesis including data collection, experimental design, research hypotheses and ethical issues.

3.2 Data Collection

This research used an experimental approach to collect data in order to investigate human colour perception and related errors. The experiments investigated the use of colour scales in classifying medical digital images showing cellulitis or conjunctivitis. This study is the first to investigate the application of the SAF teleconsultation method in the clinical investigation of these two pathologies. Other methods of data collection, such as interviews or questionnaires, were not used in the study because experiments allowed the procedures to be planned in detail. The experimental design involved individual participants, participant groups (specifically doctors and non-doctors), or a mix of both.

The following procedure was applied for acquiring the digital images:

- Acquire the photographic equipment: Equipment varies widely with regard to type, features and quality.

- Select, or prepare, the location of the photographic subject: The subject may be located indoors or outdoors and within a controlled or uncontrolled environment.
- Set up the subject to be photographed so as to maximise the informational quality of the image.
- Photograph the subject after adjusting the digital equipment for focus and white or colour balancing. Adjustments may be performed manually or by using automatic self-adjusting functionality built into the equipment.
- Preview image quality: Images used for medical analysis need to be as faithful to the original as possible to avoid errors in the medical diagnosis.
- Write digital images onto media for storage or transmit them for further processing and display.

The nature of the data collected in this study was a combination of both qualitative and quantitative data. The qualitative data describes the subjective interpretation of colour in an image whereas the quantitative data represents the measured number of errors in a specific task.

3.3 Experimental Design

The main approach taken in this research was experimental in nature. This is common in the framework of human factors, in particular in the fields of cognitive science and HCI. Of the studies discussed in chapter 2, Campos et al. (2012), Lopez et al. (2011), Lasierra et al. (2012), Rios-Yuil (2012), Colven et al. (2011), and Smith et al. (2013) all used various types of experimental designs. The research problem that this study investigated was defined and detailed in the introductory chapter to this thesis. The principal research objectives and questions were also specified in that same chapter. The study applied the empirical hypothetico-deductive reasoning model in its methodology. Popularly known as the "scientific method", this approach uses formal hypothesis testing as the core element of the research (Popper 2002). The proposed experimental procedure was first tested by a descriptive and cluster analysis of a pilot study that involved 37 participants. Following the pilot, twenty-one experiment sets were designed and conducted in order to confirm or refute the hypotheses (Hayes 2000; Dawson 2005; Oates 2005). The experimental design isolates causes and manipulates the variables under controlled conditions in order to test out specific factors. This approach tends to minimise the complexities of human experience, as stated by Hayes (2000), and the experimental design took account of this. The previous chapter provided a high-level specification for a future auto-evaluation telehealth system that could be adopted by further studies. In order to facilitate the classification of images by such a system, Machine Learning (ML) techniques were investigated and used. Two sets of images were used, one for testing and one for training purposes. The software for the classification process was coded in Python using IPython Notebook.

3.4 Machine Learning

For differentiating between healthy eye images (normal eye with no redness) and images of eyes with conjunctivitis (moderate or severe), ML, with the help of the OpenCV library, was used. OpenCV is an open source computer vision and machine learning software library originally developed by Intel Research. The library includes some 2500 optimised computer vision and machine learning algorithms, both classic and state-of-the-art. The procedure for downloading OpenCV is explained in appendix 2.

In order to use ML, the images were represented in the form of feature vectors. Because the most distinct feature that differentiates a healthy eye image from an image of an eye with conjunctivitis is the colour (redness) of the eye, a technique using colour histograms was used for feature representation.

Colour histograms represent the distribution of pixel intensities in an image. By default, OpenCV supports images in BGR (blue-green-red), where three channels are used to represent the three colours blue, red and green. This is opposed to the more usual RGB (red-green-blue) format. However, for classification purposes, it is assumed that images of eyes with conjunctivitis will contain a higher concentration of red. Thus, for more accurate classification, the images were represented using the red channel only.

Due to the difficulty of isolating the area of interest, there was a danger that the results could be affected by contextual colour if analysing the red channel only. For this reason, additional information was added to the algorithm to permit analysis by HSV. The results of this system were also analysed as a comparison with the red-only analysis method. A supervised ML technique was applied by dividing the image data into training and testing sets, with 60% used for training and 40% for testing. The images were assigned to different classes corresponding to the severity of infection in the eye. Accordingly, the problem was modelled as a classification task where the aim was to predict class labels for each of the testing images. The accuracy was calculated by comparing the labels predicted by the algorithm to the actual class labels of the images. For classification, the 1-nearest neighbour algorithm was used. This is described in the literature review. Given a testing image, the nearest neighbour algorithm works by comparing the testing image to training images. A classification decision is made based on the class labels of the training images that are most similar to the testing image. The algorithm works by reading in the entire set of training images, computing their colour histograms and storing them in the system's database. The algorithm compares the colour histograms of testing images with those in its database of training images and the testing image is then assigned the same classification label as that of the closest match. The accuracy of this assignment is then computed by comparing the class labels assigned by the algorithm to the actual class labels of the testing images. The source code used for image classification and the results of the auto evaluation can be viewed in appendix 3 and were also uploaded to the GitHub public open source software repository (https://github.com/Ibrahim-Alwawi/image-classification).

The embryonic version of an auto evaluation system was developed by firstly creating two learning folders, one each for cellulitis and conjunctivitis. As the performance of the auto-classification system was to be judged against that of the human participants, it was necessary to use the same images as those utilised in the experiments. This meant that the number of unique training images was limited to 78 for cellulitis and 29 for conjunctivitis. This was far fewer than was ideal as it limited the size of the knowledge base and therefore the potential accuracy of the system. Testing images of cellulitis and conjunctivitis pathology were then submitted to additional folders, and the system was asked to find, for each testing image, its closest match from the appropriate learning folder. The tests were run with evaluation alternatives of two, three and five levels, equating to binary either/or testing, or to the 3 point and 5 point scales offered to the participants. Further examination and explanation is included in chapter 8 and a specification for the auto-evaluation system can be found in appendix 1.

The planning of the experiments in this study was influenced by the agile approach used in software development (e.g. Highsmith 2002) as shown in figure 3-1. Projects undertaken following the agile methodology are conducted in small steps with frequent detailed reviews and re-planning to assess results and to decide on further work and adjustments. In this work, a pilot experiment using cellulitis images tested the ability of participants to describe, group, rank, and match the colour red. However, the cellulitis experiments were more specific and precise. After reviewing these experiments, the methodology of the conjunctivitis experiments was developed. The agile approach has the advantage of allowing the experimental cycle to evolve through constant review and redesign, thereby permitting the experiments to be adapted to the complexities of circumstances (Abrahamsson et al. 2002; Lyons 2004; Misra, Kumar and Kumar 2006).

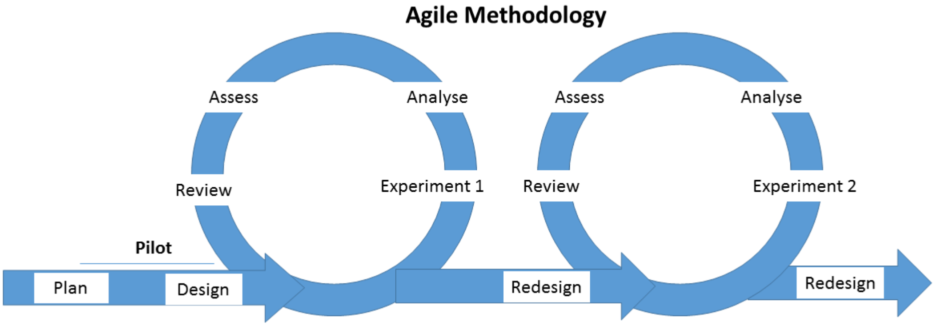


Figure 3.1: Agile Methodology stages-graphic designed by researcher

Independent and Dependent Variables

As regards control groups or factors for this research project, it is essential to highlight that the study compared the colour perception of doctors and non-doctors in assessing redness in a diagnostic or medical context and not the diagnosis itself. For this purpose, the research considered the number of errors (the accuracy) in each task of the experiments as the main measure for colour perception. The study also investigated the use of a red colour scale in assessing and classifying redness by providing it as a guide in some tasks and removing it when repeating the same tasks in a random order.

A number of independent variables were considered in consultation with Aberdeen Royal Infirmary (ARI), the partner in the study. These included:

1. The types of medical conditions that might be investigated (cellulitis and conjunctivitis).
2. The levels of expertise of the participants (medical doctors and non-doctors).
3. The use of a redness intensity colour scale in telehealth (chapter 2, section 2.6.3.2).
4. The colour characteristics (such as intensity and resolution) of the used images.
5. The use of an image quality scale in telehealth. This considers the resolution of an image in terms of its quality; classified as low, medium or high, as shown in figure 3.2.

The following images shown in figure 3.2 were designed, proposed and used by the study as an image quality scale with sample images to guide the participants in assessing the quality of the images before answering the questions.

As mentioned previously, Corel PHOTOPAINT X5 and Adobe Photoshop 11 were used. The raw image files from the camera were opened in Corel PHOTOPAINT then exported as JPEG format image files. When exporting the images there is a 'Quality' option which can be set from 0% to 100%. For the purpose of this experiment, 0%, 5%, and 10% were classified as low quality; 30%, 40%, and 50% JPEG as medium quality; and 80%, 90% and 100% were high quality.

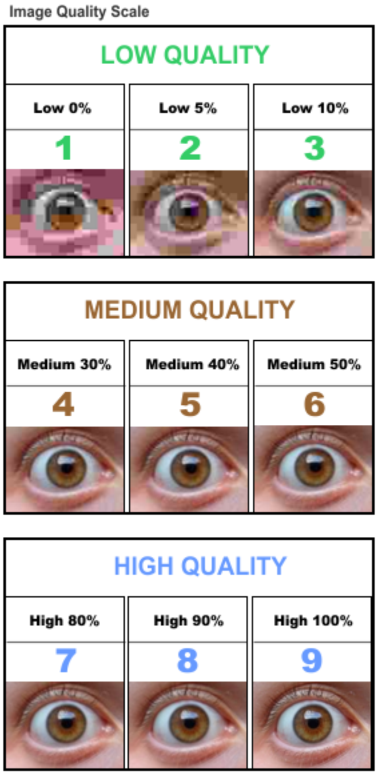


Figure 3.2: Red eye image quality scale used in the study

The dependent variables in the study included the number of accurate answers, the number of errors made, the time spent in performing the experimental tasks (the 'run-time'), subjective qualitative assessment of the degree of redness in images and classifications of the ranking and grouping of images. It further considered the mean of the opinions of observers in assessing image quality, using resolution as the key characteristic

Number of experimental images

Studies have hitherto used different numbers of experimental images. For example, Buckley, Andelson and Agazio (2009) had 43 patients with 89 wounds. Tsai et al. (2007) had 300 digital images from 20 patients. These images comprised 100 front facial photographs, 100 of the right side of the head and 100 of the left. 180 images were captured by study staff and 120 by subjects. Kim et al. (2003) had 70 patients who were enrolled and assessed in person; the data was collected during 430 visits (with a maximum of 6 visits per patient).

For the experiments involved in this study, a range of images were used as follows:

- Pilot experiments used 8 images of red roses and 12 images featuring plain areas of red.
- Cellulitis experiments used 197 images, many of which were duplicated and modified.
- Image matching experiment used 50 images, 25 of which were modified.
- Red eye experiments used 410 images, many of which were duplicated and modified.

The number of images used in the experiments associated with this study are held to be sufficient to provide a high degree of confidence in the results.

Confidence in the current study

The study used a scale of 10 categories (from 0 to 9) to measure the confidence of the users in the correctness of their answers. Zero (0) on the scale represents a complete absence of confidence while nine (9) represents absolute or full confidence. These values correspond to 0% and 1005 respectively on the confidence scale used by Peirce and Jastrow (1885) and Adams (1957). This study used a 0-9 scale instead of percentages in the experiment to simplify the results. Percentages, however, were calculated and used subsequently in the analyses. The study looked into the overall confidence values, but also investigated differences between doctors and non-doctors in their confidence level. Confidence levels were gauged when using or not using colour scales, when using scales of three or five reference points, and when carrying out easy or difficult tasks.

The present study explored the relationship between confidence of the participants and accuracy of their answers. This relationship is called calibration (Pulford 1996). All the above-mentioned studies confirmed this relationship in different ways. However, confidence is largely a subjective measure since there are individual differences due to human cognitive behaviour such as overconfidence, under-confidence, level of education, or experience.

It can be concluded that confidence can be an indication of the accuracy but is not always a strongly correlated factor due to the aforementioned reasons. This conclusion also applied to the time spent when doing a task. The study proposed a common language, in this context a red colour scale, to be used in telehealth as a reference point in diagnosis in order to minimise errors. Confidence scaling was explored and used in the experimental work as a supportive indication of the accuracy level of using the red colour scale in order to compare the results.

3.5 The Experiment Sets in the Study

The following figures show the four experimental groups and samples in order:

1. Pilot experiments:

The first group was the pilot. This was conducted as two experiments repeating the same tasks with the same participants but adding more information and providing guidelines about the colour red in the second experiment (figure 3.3). The pilots were used as a system for testing the effectiveness of the experimental methodology and to determine whether adding further guidance would improve the accuracy of the results.

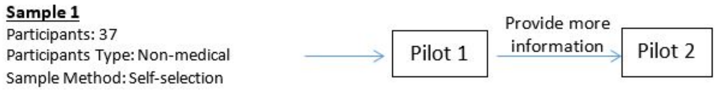


Figure 3.3: Pilot experiments

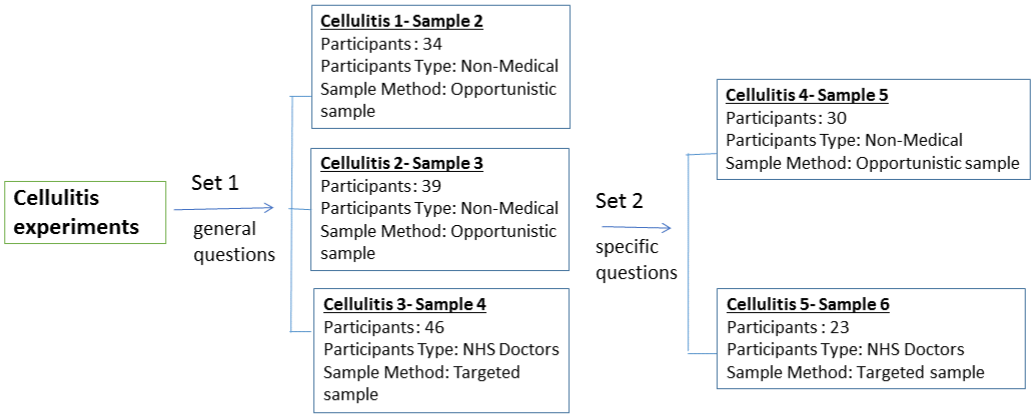

Cellulitis experiments
Set 1
general questions
Cellulitis 1- Sample 2
Participants: 34
Participants Type: Non-Medical
Sample Method: Opportunistic sample
Cellulitis 2- Sample 3
Participants: 39
Participants Type: Non-Medical
Sample Method: Opportunistic sample
Cellulitis 3- Sample 4
Participants: 46
Participants Type: NHS Doctors
Sample Method: Targeted sample
Set 2
specific questions
Cellulitis 4- Sample 5
Participants: 30
Participants Type: Non-Medical
Sample Method: Opportunistic sample
Cellulitis 5- Sample 6
Participants : 23
Participants Type: NHS Doctors
Sample Method: Targeted sample

Figure 3.4: Cellulitis experiments

2. Cellulitis experiments

The second group of experiments considered the skin infection cellulitis. This comprised five experiments in two sets, and used medical images which were collected from ARI. The difference between the two sets, apart from the samples, was that the questions were more precise and specific in the second set.

3. Image matching experiment

The third experiment group was an image matching investigation. This was a repeat of experiment 5 in the cellulitis experiments that was undertaken in response to the high rate of errors committed by participants in answers to specific questions during that experiment. It was evident that these questions had been designed in a way that made them difficult to answer due to the degree of degradation in colour red between the images being too small to be recognised. The repeated experiment was intended to further investigate these results with more understanding of the similarities and differences between the images and their colour characteristics.

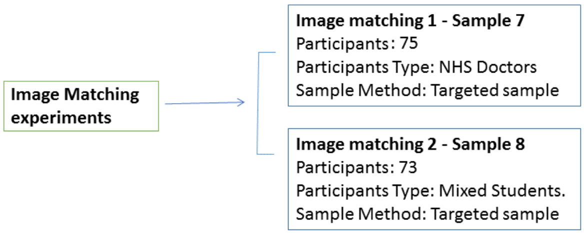


Image Matching experiments
Image matching 1 - Sample 7
Participants: 75
Participants Type: NHS Doctors
Sample Method: Targeted sample
Image matching 2 - Sample 8
Participants: 73
Participants Type: Mixed Students.
Sample Method: Targeted sample

Figure 3.5: Image matching experiment

4. Red eye / conjunctivitis experiments

The fourth experiment group was the red eye or conjunctivitis series. Five experiments were conducted using two participant samples who were provided with images showing various levels of conjunctivitis severity. The experiments also introduced the red colour scale into the process of colour judgement. Due to the rate of errors observed in experiments 1 and 2 (which were considerably higher than in experiments 3, 4, and 5), it was decided to repeat the first two experiments using two new participant samples in order to confirm the findings.

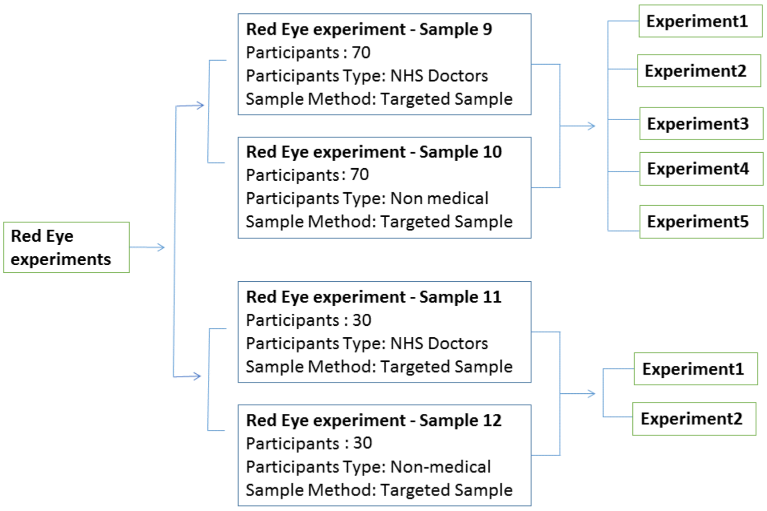

Red Eye experiments
Red Eye experiment - Sample 9
Participants : 70
Participants Type: NHS Doctors
Sample Method: Targeted Sample
Red Eye experiment - Sample 10
Participants : 70
Participants Type: Non medical
Sample Method: Targeted Sample
Experiment1
Experiment2
Experiment3
Experiment4
Experiment5
Red Eye experiment - Sample 11
Participants : 30
Participants Type: NHS Doctors
Sample Method: Targeted Sample
Red Eye experiment - Sample 12
Participants : 30
Participants Type: Non-medical
Sample Method: Targeted Sample
Experiment1
Experiment2

Figure 3.6: Red eye experiment

All experiment sets were standardised with the same controlled conditions, instructions, tools, and environmental conditions for all participants. For example, all computers used in the pilot, cellulitis, and red eye experiments had the same technical specifications. The distance between the display screen and the participants was kept constant.

The illumination was ambient daylight from nearby windows backed-up by fluorescent strip lighting and these lighting conditions were also kept the same in every instance. i.e.. Chapters 4, 5, 6, and 7 report in detail on these experiments.

3.6 Population and Sampling Method

Collis and Hussey (2009) defined the term population as "a precisely defined body of people or objects under consideration for statistical purposes". The present study used the following two populations:

3.6.1 Medical doctors

This included all certified medical doctors who have had experience of or familiarity with cellulitis (for the first set of experiments) and with conjunctivitis (in the second set of experiments). However, doctors who did not have experience with these two pathologies were also included for comparison purposes. The study contacted NHS doctors directly and invited them to participate.

3.6.2 Non-medical people

This study refers to these individuals as "non-doctors". They included any person who was not a qualified medical doctor. Nurses and other medically trained professionals were also excluded in order to avoid confusion. Participants who were vulnerable such as children, elderly persons, or convicts were also excluded. To ensure that the sample represented the population as closely as possible, RGU students from all the different schools and departments were invited to participate in an online experiment.

3.6.3 Sampling method and frame

It is assumed that doctors who use telehealth systems have a level of knowledge about colour even though they are not usually taught this specifically during their medical training. As a result of this, they may use many different and disparate terminologies when describing colour. The research investigated the way that they described and stratified the colour red in digital images that showed cellulitis or red eye. Also the study focused on the level of accuracy in the performance of doctors and non-doctors who used the system and their confidence in their answers. The study identified the cognitive faculties involved in the perception of the colour red, and how to improve the performance of medical and non-medical users of telehealth systems, in order to get the maximum benefits and most accurate results.

Individual differences play a key role in recognising, recalling and matching visual images and colours. This led the research to assume that individual differences between users of telehealth systems might lead to different types of errors during the diagnosis process. The study investigated the difference in accuracy between doctors and non-doctors in using telehealth when diagnosing cellulitis or red eyes using digital images. This type of investigation should show if there is any difference between the doctor and non-doctor groups on the one hand and the difference between the users within each group on the other hand.

Sampling techniques are vital to obtain a reasonable amount of data from only a sample rather than the entire population. According to Saunders et al. (2012), a sample frame provides a list of the members of the population from which the sample is drawn. The samples selected for the experimental work of this study were either NHS medical doctors or non-medical persons who were all undergraduate or postgraduate students from RGU.

Table 3.1: Sample frame of the study for primary data

Experiment Name	Participants type	Participants group	Number of Participants in each group	Total Participants in each group	Type of selection method
Pilot 1& 2	Non-medical & RGU students	1	37	37	Self-selection volunteers
Cellulitis 1 - 5	Non-medical MSc students	2	34	68	Opportunistic sample
	Non-medical undergraduate	3	39	117	Opportunistic sample
	NHS doctors	4	46	184	Targeted sample
	Non-medical MSc students	5	30	150	Opportunistic sample
	NHS doctors	6	23	138	Targeted sample
Image matching 1 & 2	NHS doctors	7	75	525	Targeted sample
	Non-medical Mix students	8	73	584	Targeted sample
Red eye 1-5	NHS doctors	9	70	630	Targeted sample
Red eye 1- 5	Non-medical RGU mix students	10	70	700	Targeted sample
Red eye 1& 2	NHS doctors	11	30	330	Targeted sample
Red eye 1& 2	Non-medical RGU mix students	12	30	360	Targeted sample
				Total = 3823	

A total of 12 groups of participants took part in this study, each group containing varying numbers of individuals. The total number of experiments that were conducted was 23 as shown in table 3.1. Some participants volunteered to take part in several different experiments, which made the process easier and the results more accurate, as finding identical or similar samples can be challenging. In both pilot experiments (1 and 2), and in the cellulitis experiments (3, 4, and 6), 37 students were selected from the undergraduate and postgraduate student population of the School of Computing Science and Digital Media at RGU where the researcher does his academic teaching training. More than 60% of the students in these three classes volunteered to participate in the experiment. In experiments 9, 15-19, and 22-23, the samples were targeted. Again, all were RGU students who volunteered to take part in response to an email appeal for participants. In experiments 5, 7, 8, 10-14, and 20-21, the samples were also targeted samples. All participants were NHS doctors from ARI, who again volunteered in response to an email. The total number of participants in all experiments was 3823 as detailed in table 3-1 above.

3.7 Experimental Report

All 23 experiments are reported in this thesis using the same structured approach. Chapter four reports the two pilots; chapter five reports the five cellulitis experiments; chapter six reports the two image matching experiments; and chapter seven reports the ten red eye experiments. Each experiment is reported according to the following protocol;

- Introduction and objectives of the experiment.
- Method, including a description of the attributes of the participants. This includes personal characteristics such as age or technical background.
- Materials and tools used in the experiment. This includes items such as images for each question, instruction sheets, answer worksheets, consent forms; design and variables of the experiment.
- Experimental procedure, which gives details of how the evaluation was carried out and administered.
- Summarised results.
- Discussion of the results in the context of the research hypotheses, questions and objectives.

3.8 Research Hypotheses

Hypothesis 1: The impact of clinical experience on the accuracy of colour perception
There is difference between doctors within their group, non-doctors within their group, and between the two groups in the level of accuracy in relation to colour perception due to medical background and clinical experience. (The null hypothesis of which is: There is no significant difference between the participants)

Hypothesis 2: The impact of a numeric descriptive the accuracy of colour perception
There is difference between the accuracy of colour perception when using numeric descriptive colour scale with or without standard pictures with three or five divisions. (The null hypothesis of which is: There is no significant difference between the two groups)

Hypothesis 3: The relationship between accuracy, confidence and time
There is a relationship between the accuracy and confidence as well as the time spent in teleconsultation when using digital images that are showing cellulitis or red eyes. (The null hypothesis of which is: There is no relationship between the three variables)

Hypothesis 4: The impact of a digital image scale on accuracy, confidence and time
There are changes to the accuracy, confidence or time in teleconsultation when using digital images showing cellulitis or red eyes, owing to the use of the proposed scales. (The null hypothesis of which is: There are no significant changes between the variables)

Hypothesis 5: The difference between machine learning and human models in colour classification accuracy
There is difference between the accuracy of a machine learning model compared to a human system in colour classification. (The null hypothesis of which is; there is no significant difference between the two systems)

3.9 Overview of Ethical Issues

The following section details the ethical issues affecting the conduct of the study. These include such matters as obtaining ethical approval, the rights of participants, data confidentiality, the risks involved and protocols for the dissemination of results.

3.9.1 Ethical codes and approvals

This research carefully applied all of the following ethical codes and policies:

- The Code of Ethics and Conduct of Robert Gordon University (RGU) which is available at www.rgu.ac.uk/research-ethics-policy, accessed on 3 October 2013.

- The ethical policy and guidelines of The British Psychological Society (BPS) which is available at http://www.bps.org.uk/what-we-do/ethics-standards/ethics-standards, accessed on 3 October 2013.

- Research ethical review and guidance of the National Health Service NHS which is available at http://www.nres.nhs.uk/applications/guidance/research-guidance/?1654606_entryid62=83668, accessed on 3 October 2013.

The project did not include vulnerable people as participants. There were two ethical approvals required for the practical work of the study. The first one was the academic ethical approval from RGU. The second was approval from Grampian NHS for obtaining medical data and images. The academic ethical approval was obtained and granted by the ethics committee at RGU, see appendix 4: RGU ethical approval form. Dr James Ferguson (Director A&E at Aberdeen Royal Infirmary (ARI) and external advisor to this project) obtained the medical ethical approval from the appropriate local ethical committee(s) at ARI for the cellulitis experiments. The medical images used were provided by ARI. This project undertook to abide by the rules and stipulations set by that committee for using the medical images in the experiments. The fourteen red eye experiments were reviewed and approved by the North of Scotland Research Ethics Service (NRES) committees as well as NHS Grampian Research and Development R&D, see appendix 5: NHS ethical approval letter.

Due to the time constraints, the images that were used in the red eye experiments were collected from the web and then verified by three NHS experts who also participated in the five experiments. Their agreed answers were considered and taken as the standard for the experimental design system and were used as control references for the answers provided by participants. The three experts were provided with the questions of the experiments separately and any disagreements were removed from the question sets that were finally used in the experiments.

Studies have hitherto used different numbers of medical staff to evaluate images. For example, Kim et al. (2003) had 6 physicians and rotated the role in pairs in both face-to-face and SAF methods, Tsai et al. (2004) and Krupinski et al. (1999) had 3 dermatologists and Tsai et al. (2007) had 2. Buckley, Adelson and Agazio (2009) had one nurse at each end of a tele-link. It is important for the evaluation to be done by more than one dermatologist in order to be more reliable and avoid any risk of bias in the results. Following are some of the key ethical issues that were applied during the practical work.

3.9.2 Rights of participants and data confidentiality

The research recognised all the following rights of the participants:

- They were fully informed about the experiments before taking part by means of a participant information sheet.
- Upon being accepted as volunteers, they were given as much time as they required to consider whether to actually take part or not.
- They were given the option of changing their minds about participation.
- A consent form was required to be signed by participants, either electronically or in hard copy, before they started the experiments.
- The experiments were conducted at times which were convenient for the participants.
- They were free to withdraw at any time during the participation with no obligation to give a reason.
- All records of their participation in the research were kept confidential and anonymous. The names and personal details of the participants were not recorded. Participants were instead identified using unique serial number IDs.

3.9.3 Images archive security

The images obtained in this research were encrypted and password protected to prevent the images from being accessed by those not authorised to do so. The first step was to store the images on a secure hard drive and then to degrade and encrypt the images in a secure password system before they were used in the experiments. Thus, at no time were these images available to view via a network or any other electronic sharing system that could be accessed by an unauthorised third party. For more protection, images taken were deleted from the original devices immediately after transferring and storing them on a secured computer hard drive in order to be manipulated (degraded in colour, quality and intensity).

These original images transformed for the experiments were password protected to ensure that accessibility to them was restricted to authorised personnel only. All data was stored anonymously in accordance with standard NHS data protection procedures. Images and numerical data were encrypted and stored securely subject to Robert Gordon University procedures. The data will not be kept longer than necessary for this research and will be archived as is acceptable to both university and NHS protocols. The secure computer's hard drive (which was specifically assigned for this research project) will be stored in the archives at RGU.

3.9.4 Risk and ethical impact

The experiments were planned to be safe and risk free for all participants. The clinical data and images used in the cellulitis experiments were gathered directly by medical experts. All patient contact and interaction was also carried out by healthcare professionals in the A&E department at ARI. If participants were sensitive to flashing lights then they were advised not to take part in the study. This warning was written clearly on the patient information sheet.

3.9.5 Feedback of the study

Ethically, the participants should also be given full feedback on conclusion of the study and the findings shared with them. (Hayes, 2000). Accordingly, a summary of the study and its results will be sent to them by email after the work is fully approved and published. The results are published as a part of this PhD thesis. The name and personal information of the participants will not be identified or disclosed in any report or publication. For further information or clarification about the study, the participants were provided with the contact details of the investigator.

3.10 The Data Analysis

Given the nature of the data, relevant non-parametric statistical tests were chosen for analysis. The data was analysed to identify factors that may influence colour perception and accuracy in judging images. For example, Spearman's Rho (ρ) analyses were carried out to investigate the correlation between variables such as the confidence level of the participants and the accuracy in their answers. The specific tests are explored in detail in the relevant chapters of this study.

3.11 Validity and Reliability of the Methodology

The methodology used in the study was deemed to be valid and reliable for the reasons listed below:

- Three medical experts verified the images that were used.
- The colour red scale used by the study was already in use as a support to the traditional diagnostic protocol used by the NHS.
- The confidence scale used in the study was verified by previous studies, as mentioned in chapter two.
- Pilot experiments were employed to check the methodology and experimental protocols for the rest of the experiment sets.
- The familiarity with the images used in the pilot experiments by non-doctor participants and with doctor participants in other experiments.
- Repeating the questions and tasks in a matched guise formalism confirmed the results and proved their reliability.

Chapter 4

Pilot Experiment

4.1 Introduction

The pilot was the first experiment that was conducted in the study. Non-medical images were used of general day-to-day items that exhibited varying degrees of redness. Only non-doctor participants were involved as there was no need for any comparative analysis between the performance of doctors and non-doctors at this stage. The main aim of this experiment was to test the methodology and experimental design of the study, and to answer research question 1 regarding consistency between participants. It was intended that this pilot experiment would provide a framework for the development of the experimental procedure and design-methodology for the next series of experiments. It was also intended to provide interpretation of stimuli in terms of its basic information. Therefore, the experiment provides data on the way humans describe, classify, and rank colours as a generic perceptual and cognitive skill. The experiment consisted of 6 tasks, three of which were called pilot 1 and the other three tasks which were called pilot 2. The three tasks in pilot 1 were called, for identification purposes, (Task 1.A, Task 2.A and Task 3.A). Similarly, in pilot 2 the tasks were identified as (Task 1.B, Task 2.B and Task 3.B). All tasks for the two pilot experiments can be viewed in appendices 6.1 to 6.6.

- Task 1.A: To describe, without guidance, the colours in the images shown in figure 4.1.
- Task 1.B: As above but with guidance (additional information on key colour attributes such as hue, saturation and intensity).
- Task 2.A: To group the images shown in figure 4.2 based on their colour

- Task 2.B: As in task 2A above but with guidance.
- Task 3.A: To rank the images shown in figure 4.2 based on degree of redness.
- Task 3.B: As in task 3A above but with guidance.

The participants were not given any information about colour characteristics in task 1.A and they were expected to answer the questions concerning colour perception based on their own experiences and to group the images based on unique qualities. No information was given concerning the number or specification of each group in task 2.A, so that it would not influence their judgement. In task, 3.A they were expected to rank these images on colour intensity on a scale from 1 to 12, also using their own judgement. However, in tasks 1.B, 2.B and 3.B information was provided to the participants about colour characteristics (specifically about hue, saturation, and intensity) in order to investigate the effect on the results of the added information.

4.2 Participants

As mentioned in chapter three (section 3.5.3) and table 3.1, a total of 37 participants, who were all self-selected volunteers, took part in the experiment. According to Hayes (2000), at least 30 participants are required to ensure that the results are statistically significant. The participants were from varying backgrounds and age groups (but were mainly in their 20s and 30s and were postgraduate students or oil and gas engineers).

All participants were male, but this does not affect the results following previous studies (chapter 2 section 5.2) which showed no consensus regarding colour related activities due to sex. 20 of the participants had normal eyesight and the remaining 17 were corrected to normal with prescription lenses.

4.3 Design and Variables

All participants were given the same instructions and worked within the same standardised environment with the same controlled conditions. The independent variables were the tasks of the experiment, colour characteristics and the environmental conditions in which the study was conducted. The dependent variables were the participants' responses to the experiment in describing, grouping and ranking the images. The numbers of accurate answers and of errors committed were also regarded as dependent variables. This experiment focused on the three colour characteristics referred to in chapter 2 section 4.

4.4 Materials

The following list contains the equipment and tools required for the experiment:

Images

In total 32 images were used in the two experiments. 8 images were used in tasks 1A and 1B, 12 images in tasks 2A and 2B, and 12 in tasks 3A and 3B. There were two different groups of images used in the experiment as shown and described below with some sample images. The images in figure 4.1 are the 8 images that were used in task 1 of this experiment. A red rose was chosen for this pilot experiment because of its familiarity and because it was less likely to cause offence to the participants. Images were not chosen preferentially and no technical measurements (for example of levels of saturation) were applied in choosing the images. The aim was to investigate the common words used in describing these images as opposed to identifying the differences in image quality and colour characteristics.

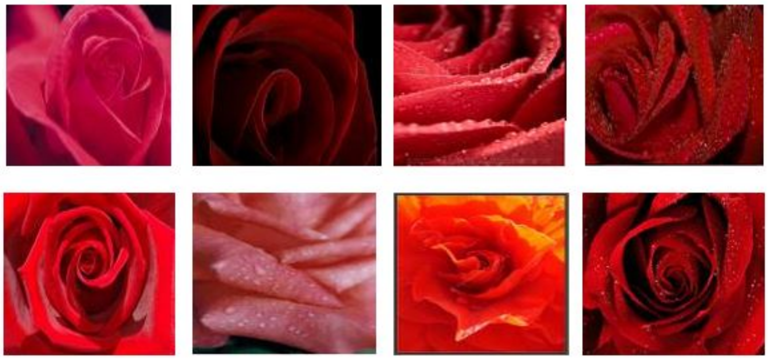


Figure 4.1: Non-medical images used in the colour description task 1 of pilot study

Tasks 2 and 3 employed 12 images with the aim of identifying image groupings and ratings. These 12 images were created and designed using Adobe Photoshop image editing software. Colour chips of pure redness were created to represent varying values of intensity. These were used to assess the ability of participants in differentiating between the images based on colour characteristics. Some chips were identical to test the participants' ability to discern differences (for example, chips 1 and 6 and 10 and 12 have identical colour characteristics as shown in figure 4.2). Figure 4.2 shows the 12 colour chips that were used in tasks 2 and 3.

Table 4.1 describes the saturation and intensity values of the 12 chips (figure 4.2) that were used in tasks 2 and 3 of the pilot experiment. This table was used as a reference to evaluate the participants' answers.

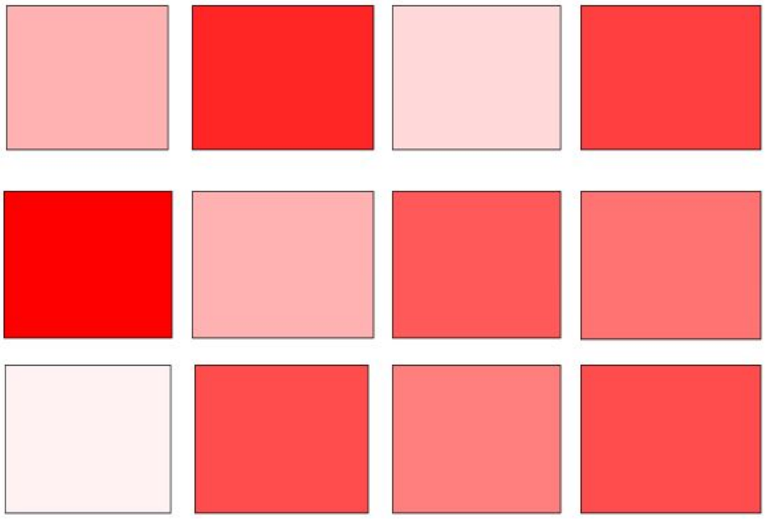


Figure 4.2: Colour chips used in colour description tasks 2 and 3 of pilot study

Table 4.1: Colour characteristics in tasks 2 and 3 of the pilot experiment

Image	Intensity	Saturation
Image 1	204	30%
Image 2	110	85%
Image 3	230	15%
Image 4	128	75%
Image 5	85	100%
Image 6	204	30%
Image 7	144	65%
Image 8	162	55%
Image 9	246	5%
Image 10	136	70%
Image 11	170	50%
Image 12	136	70%

In table 4.1 above, chips1 and 6, and images 10 and 12 are identical in the listed colour characteristics. his can be seen in figure 4.2 above.

Consent form

The participants were asked to formally consent to take part in the study (see appendix 7.1 for details) and were requested to confirm that they had read and understood all the information that had been provided to them about the study. Participants were also required to declare that they understood that their participation was voluntary and that they were free to withdraw at any time without giving any reason. Participants were at liberty to ask questions at any time, and all such queries that were raised were satisfactorily addressed by the researcher.

Colour blindness test sheet

The Ishihara colour blindness tests were used to test participants for colour blindness. A total of 15 plates were used. The participants were asked to write down the numbers that they could recognise on the plates on an answer sheet provided. Plates 13 and 14 did not actually feature a number (only a parti-coloured background) in order to test the reliability of the participant's answers to the Ishihara plates.

Instruction and answers sheets

In order to conduct the pilot experiment the participants were provided with forms which displayed the images shown in figure 4.1 and were asked to answer the question printed on the same form (task1). For tasks 2 and 3 they were provided with the colour chart shown in figure 4.2 together with questions and an answer sheet. See a sample sheet in appendix 6. The full instruction and answer sheets may also be accessed via the Dropbox link at http://bit.ly/1l5WyTy.

4.5 Procedures and Tasks

All ethical issues mentioned in chapter three above were addressed. Before commencement of the experiments, the participants were fully informed of their right to withdraw at any time and their formal consent was recorded. Each of the experiments required about 45 minutes to complete. All tasks were explained in both verbal and written forms to the participants.

All participants were screened for colour blindness before participating in the experiment. Participants with reference numbers; 11, 23, 24, 26 and 36 were identified as being colour blind and had their data removed from the results. The results from the remaining 31 were deemed to be acceptable since a population size of 31 was held to be a sufficient number for statistical analysis (Hayes 2000). Instructions, materials and environmental conditions were standardised for all participants.

For both pilot 1 and pilot 2, the order in which the sequence of tasks comprising each experiment was carried out was randomised using a Latin square counterbalance measures design. This procedure ensured that the performance of the participants would not be influenced by the order in which the tasks were undertaken, as this could otherwise have affected the results. This created a total of 36 possible combinations of different orders for the tasks. One possibility was chosen randomly for each group, hence giving a total of 7 actual possibilities that were implemented.

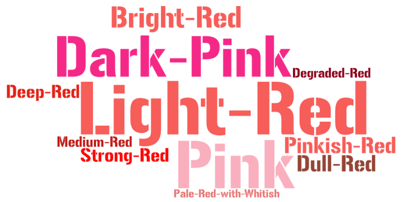


Figure 4.3: Frequency of colour descriptions-Image 1

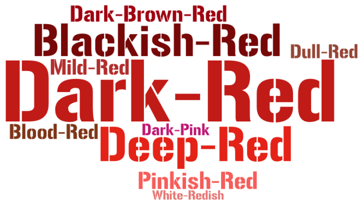


Figure 4.4: Frequency of colour descriptions-Image 2

4.6 Results and Analysis

4.6.1 Results and analysis of task 1.A (colour description)

In this task, the participants were asked only to describe the colour of an image in their own words with no guidance. The following word cloud figures 1 to 8 for the eight images used in task1.A, illustrate the degree of consistency among the descriptions given by the participants. The size of each word in a cloud indicates the frequency of that word in the responses of the participants of a particular experiment run. A word cloud therefore provides an overall visual picture of the frequency distribution. The actual numerical frequency values of all the word cloud images can be viewed in appendix 2.

In image 4.3, the four most frequently reported colours were, in descending order, light red, dark pink, pink, and bright red. However, the colour of the rose in the actual image was dark pink which matched with the second-highest frequency value.

In figure 4.4, dark red, blackish red, deep red, and dark brown red, in that order, were the most frequently reported colours. In this case, the actual colour of the image matched with the most frequently reported colour.

In figure 4.5, light red, red, pale red and dark pinkish red, in that order, were the four most frequently reported colours. The colour of the actual image was pale red which matched with the third most frequently reported colour. The most frequently reported colour, light red, was, however, still close to the actual colour.

In figure 4.6, dark red, deep red, and dull red, in that order, were the four most frequently reported colours. The actual image was indeed dark red, so the majority of participants perceived and described this accurately.

In figure 4.7, bright red, light red, very dark pink and pure red, in that order, were the four most frequently reported colours. The actual image was bright red.

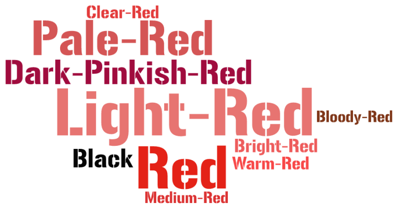


Figure 4.5: Frequency of colour descriptions-Image 3

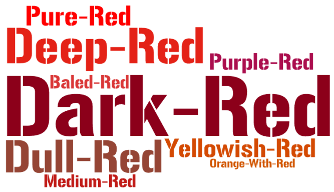


Figure 4.6: Frequency of colour descriptions-Image 4

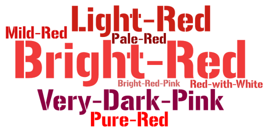


Figure 4.7: Frequency of colour descriptions-Image 5

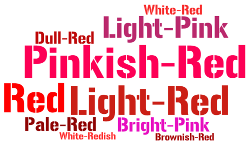


Figure 4.8: Frequency of colour descriptions-Image 6

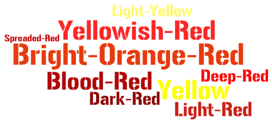


Figure 4.9: Frequency of colour descriptions-Image 7

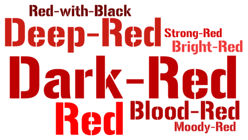


Figure 4.10: Frequency of colour descriptions-Image 8

In image 6, pinkish red, red, light red, and light pink, in that order, were the four most frequently reported colours. The first three of these colours constituted a more general and less specific description of the image. The fourth most commonly reported colour, light pink, was the actual colour of the image and thus the most accurate answer.

In figure 4.9, bright orange red, yellow, yellowish red, and blood red, in that order, were the four most frequently reported colours. Bright orange red was correct only for the central portion of the image. The third most frequently reported colour, yellowish red, was the actual overall colour of the image.

In figure 4.10, dark red, red, deep red, and blood red were the most frequently reported colours. The actual image was dark red and so the majority of the participants described this image perfectly.

As shown in figure 4.3 to 4.10, even though there was variation in the participants' descriptions, the results show that there is usually a strong degree of agreement between participants as to the general colour shown in the images. What is equally clear is that their colour descriptions lack a commonality of language.

This, in practical terms, could easily lead to misunderstandings when relating colour descriptions to medical colleagues and may lead to misdiagnoses in real situations.

4.6.2 Results and analysis of task 1B (intensity and saturation values)

In task 1.B, the participants were asked to assess values for the intensity and saturation of colour in images. The results of this activity are explained below.

The intensity data analysis

Table 4.2 provides information about the intensity scale used in the study to evaluate the accuracy level of the participants when assessing the intensity of colour in a given image. Meanwhile, the results shown in table 4.3 illustrate the accuracy of the answers and how closely they match with the correct intensity values. The levels of intensity used for comparison with the answers were established from the metadata for each image in Adobe Photoshop image processing software.

The parameters set as an acceptable result were the Photoshop value +/- 10 percentage points. The results showed that the averages of the answers of all participants fell within the acceptable level of Photoshop values +/- 10 percentage points. The standard deviation, however, indicates that there were large variations between the values assigned by individuals.

Table 4.2: Intensity scale used by the study

Levels of Intensity	Number of images	Images of the category
High (more than 60%)	3	image1, image 2, image 7
Medium (around 50%)	4	image 3, image 4, image 6, image 8
Low (less than 40%)	1	image 5

Table 4.3: The accuracy of participants' results and Photoshop (intensity)

Images	Accurate answer in colour scale	Accepted answer in colour scale	Average of users' answers	STDV of users' answers	Total Accuracy from 31 users	
					No.	%
Image 1	70	60-80	66.7	19.4	21	67.7
Image 2	20	30-Oct	20.2	17.2	21	67.7
Image 3	60	50-70	58.2	22.6	14	45.1
Image 4	50	40-60	54.9	16.9	16	51.6
Image 5	80	70-90	73	15.5	12	38.7
Image 6	70	60-80	60.4	22.1	17	54.8
Image 7	80	70-90	72.2	12.3	22	70.9
Image 8	40	30-50	49	25.4	13	41.9

The saturation data analysis

The following table 4.4 provides information about the saturation scale used in the study to evaluate the accuracy of the participants when judging colour saturation in a given image.

The results shown in table 4.5 illustrate the accuracy of the answers and how closely they match with the correct saturation values. The saturation values used for comparison with the answers were established from the metadata for each image in Adobe Photoshop. The parameters set as an acceptable result were the Photoshop values +/- 10 percentage points.

As can be seen from table 4.5, the average values of the participants' answers for saturation were considerably less accurate than those seen in table 4.2 for intensity. Again, calculation of the standard deviation shows that there was a wide scatter of opinions among the participants. The consistent inaccuracies of the answers in task 1.B. indicate that the participants experienced difficulty in assessing the values of intensity and, particularly, saturation. This made a strong case for arguing that any attempt at standardisation by means of a word-guided scale was likely to meet with little success.

Table 4.4: Saturation scale used by the study

Level of Saturation	Images of the category	Ratio of images	Percentage
High (more than 60%)	image1, image 6	2 out of 8	25
Medium (around 50%)	image 3, image 5, image 8	3 out of 8	37.5
Low (less than 40%)	image 2, image 4, image 7	3 out of 8	37.5

Table 4.5: The accuracy of participants' results and Photoshop (intensity)

Images	Accurate answer in colour scale	Accepted answer in colour scale	Average of users' answers	STDV of users' answers	Total Accuracy from 31 users	
					No.	%
Image 1	70	60-80	69.4	16.2	23	74.1
Image 2	90	80-100	72	29.3	11	35.4
Image 3	60	50-70	61.9	24.2	14	45.1
Image 4	90	80-100	65.9	22.4	11	35.4
Image 5	80	70-90	66.9	22.3	13	41.9
Image 6	20	30-Oct	33	23.5	21	67.7
Image 7	40	30-50	69.8	13.6	3	9.6
Image 8	80	70-90	52.6	33.7	14	45.1

The next series of experiments explored the use of an image-guided scale.

Clustering analysis task 2A and task 2B

In task 2, the participants were asked to arrange 12 colour chips into clusters based on perceived similarities in the colour red. In task 2A, the participants were asked to group the colours intuitively, using their own knowledge and experience of the colour red. In task 2B, the users were given additional guidance on colour relating to attributes such as hue, saturation and intensity. The raw data was analysed using cluster analysis which is a statistical technique for finding relatively homogeneous clusters of cases based on measured characteristics, in this case colour characteristics. The results were presented in the form of a dendrogram (figure 4.4) which illustrates how the chips were clustered and how these clusters were themselves linked to form larger clusters. The study used this cluster analysis to investigate the consistency between the groupings assigned by the participants. IBM SPSS statistics software was used to calculate the "distances" or similarities between the images in terms of the specified colour characteristics. Figure 4.4 shows a dendrogram of the clustering in task 2.A of the pilot experiment. The figure shows an overlay of the ranking obtained from the colour intensity metrics in the Adobe Photoshop file metadata for each chip. The clusters are linked based on the average of the similarities among the cluster elements.

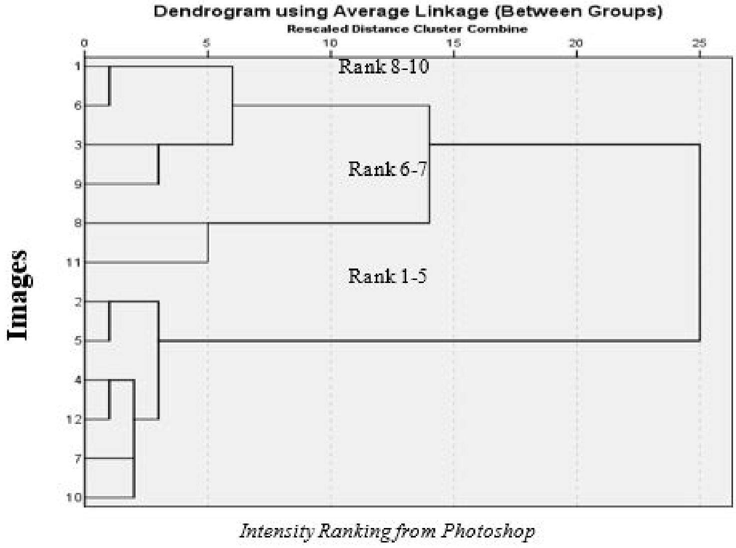


Figure 4.11: Dendrogram for clustering by participants in task 2 A

The dendrogram in figure 4.11 shows clearly a correspondence between the grouping by participants and the ranking by Photoshop. For example, observe that the colour chips with ranking 8, 9, and 10 are all within the same group in the dendrogram.

Table 4.6: The intensity of the colour chips used in experiments 2A and 2B

Photoshop ranking	Colour chip
Rank 1	Image 5
Rank 2	Image 2
Rank 3	Image 4
Rank 4	Image 10 and Image 12
Rank 5	Image 7
Rank 6	Image 8
Rank 7	Image 11
Rank 8	Image 1 and Image 6
Rank 9	Image 3
Rank 10	Image 9

Table 4.6 lists the intensity of colour on the colour chips used in experiments 2A and 2B, as defined by the file metadata in Adobe Photoshop. This utilises a sliding scale from 0 to 100 in order to establish a reference level scaled from 1 to 10. Figure 4.12 is a dendrogram representation of the results obtained from pilot experiment 2B. The figure shows an overlay of image ranking based on intensity data obtained from the Adobe Photoshop file metadata.

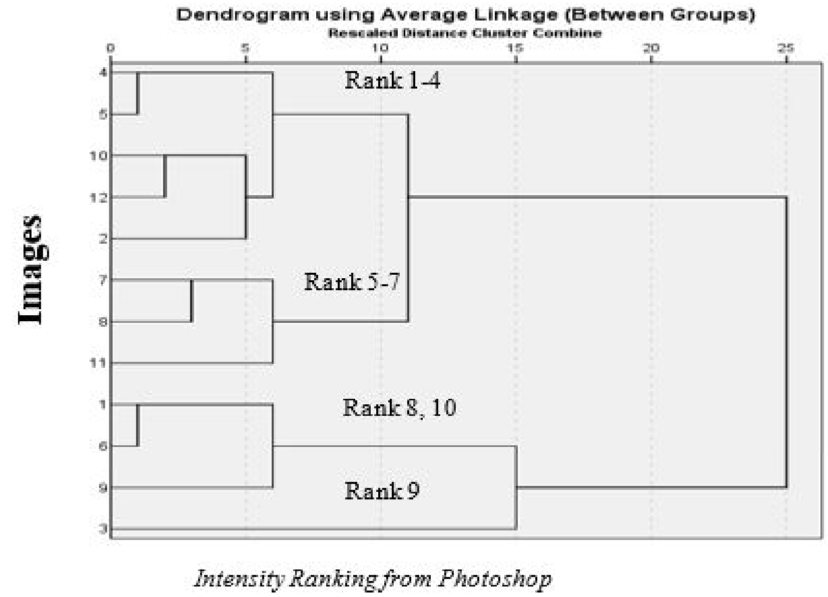


Figure 4.12: Dendrogram for clustering by participants in task 2 B

The figure 4.12 also shows an evident similarity, similar to that shown by figure 4.11, but demonstrates a higher degree of accuracy in the participants' grouping of the colours comparison.

Intensity clusters and actual intensity values in task 2A and 2B The following section focuses on how the clustering of the colour intensities as assigned by the participants compares with an intensity ranking based on actual intensity values derived from the image file metadata. In relation to the consistency of the answers of the participants, the results showed that the intensity clustering by the participants in 2A and 2B is similar, as shown the dendrogram in figure 4.11 and 4.12. In relation to consistency between the intensity clustering of the participants and the ranking of the images based on actual image file metadata, the results showed that the ranking that used actual values in both tasks 2A and 2B corresponds closely to the participant-assigned clustering. Detailed comparison of the clustering with the specific ranking values shown in table 4.6 furthermore indicates that the results are marginally more accurate after use of the guidance notes (task 2B).

Variations in colour clustering by individual participants

Because the sample participants were the same in all of the pilot experiments, it was possible to directly compare the effect of the extra guidance on each individual. Table 4.7 below shows the number of clusters created by each individual in tasks 2A and 2B and records the changes in clustering.

Table 4.7: The number clusters generated by individual participants

Participant ID	Task 2A	Task 2B	Difference
1	3	4	+1
2	3	4	+1
3	3	4	+1
4	3	4	+1
5	4	5	+1
6	4	4	0
7	6	3	-3
8	5	7	+2
9	4	12	+8
10	4	5	+1
12	6	4	-2
13	5	5	0
14	4	6	+2
15	3	3	0
16	4	3	- 1
17	3	3	0
18	3	3	0
19	3	3	0
20	3	3	0
21	2	3	+1
22	3	4	+1
25	5	7	+2
27	3	3	0
29	4	12	+8
30	4	5	+1
31	4	5	+1
32	5	4	-1
33	5	6	+1
34	9	5	+4
35	3	5	+2
37	5	7	+2

Table 4.7 shows that in the majority of cases there are clear changes in the number of clusters created in the two tasks. In most cases, there was a difference in number of one cluster between the two tasks. This reflects knowledge and experience that the participants developed in between performing the two tasks and that made them more specific and focused in their referencing of the images. Since the two tasks used the same questions it is possible that the repetition of the same experience may also have had an impact on the results. Figure 4.6 gives a pictorial representation of the effect of using a colour guide on the performance of participants. The chart also demonstrates the presence of outliers; most apparently, participants number 9 and number 29. This turned out to be a result of questions being misunderstood and ranking done incorrectly.

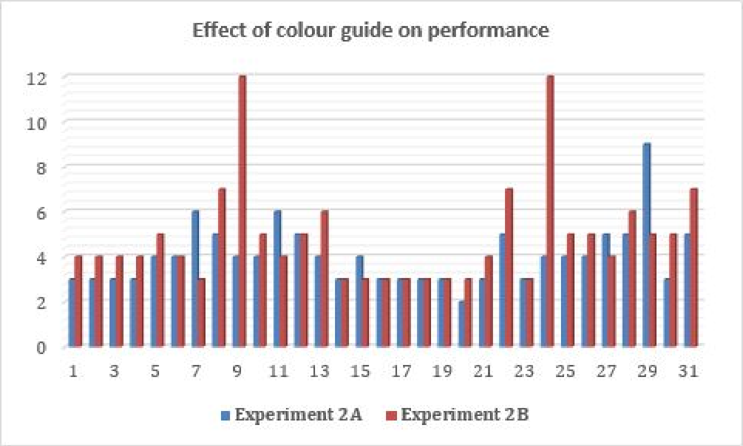


Figure 4.13: The effect on participants' performance of using a colour guide

4.7 Results of experimental tasks 3A and 3B

In task 3A and 3B the same participants as previously were asked to view the same colour chip images as those used in task 2A and 2B. On this occasion, instead of clustering the colours, the participants were asked to rank them based on their intensity. Examination of the results of task 3A highlighted a couple of issues. Firstly, it was clear that a number of the participants had either failed to understand the instructions, or had failed to apply them. This resulted in answers that bore no resemblance to the patterns observed in the answers of the other participants. The second issue arose from the lack of clarity in the question itself. Around 40% of participants ran the scale in the opposite direction to the rest, i.e. 1 to 12 going from less to more intense instead of 1 to 12 going from more to less intense. In order to analyse the results, it was therefore decided to firstly discard the answers of the two participants whose answers bore no resemblance to those of the other participants. This still left an acceptable sample size of 29. Secondly, in order to correct the anomalous answers, it was decided to apply a statistical transformation on the data using a factor of 13-n (where n represents the answer of the participant). Having made these adjustments, the data was analysed in the form of both a dendrogram and Spearman's correlation. Spearman's rho correlations were conducted to investigate the relationship between the answers given by each participant and the image ranking values sourced from the Adobe Photoshop file metadata.

The reason for using this method is that rankings are non-parametric. They indicate only rank order and their numerical values have no meaning and cannot be manipulated arithmetically. The value for Spearman's rho (ρ) is always between -1 and +1 and the closer a value is to either -1 or +1 then the more indicative this is of a stronger correlation (Hayes, 2000).

A positive relationship occurs when both images are given a high ranking score. A negative relationship means the opposite, i.e. when one image gets a high ranking and the other gets a low one.

Calculation of Spearman's correlation per participant showed accurate correlation in the case of all but four of the participants. Closer examination of their individual answers made it clear that they had failed to understand the question or had not followed the instructions. They had again clustered the images rather than ranked them as was required, thus producing anomalous data. These results were therefore discarded prior to further analysis of the data. Friedman and Kendall tests were then performed to determine the level of agreement between participants. These results were then listed in table form to directly provide a comparison with rankings derived from the Photoshop image file metadata (Table 4.8). Spearman correlations for the participants may be downloaded from the Dropbox link at http://bit.ly/1l5WyTy. Participants Nos. 3,5,9 and 22 were rejected on the basis of anomalous Spearman results. Participants 30 and 31 had previously been rejected as neither participant made any effort to rank the images.

Table 4.8: The ranking of participants and values in Photoshop (Task 3A)

Images	Photoshop Ranking	Mean of Friedman & Kendall's of participants ranking
Image 1	9.50	9.56
Image 2	2.00	2.20
Image 3	11.00	11.04
Image 4	3.00	3.30
Image 5	1.00	1.88
Image 6	9.50	9.52
Image 7	6.00	5.62
Image 8	7.00	7.62
Image 9	12.00	11.78
Image 10	4.50	4.34
Image 11	8.00	7.36
Image 12	4.50	3.78

Table 4.8 clearly illustrates the high level of agreement between the ranking of the images by participants and the ranking of the same images based on values in the Photoshop file metadata. Analysis of the level of agreement of 25 participants showed a Kendall's W of 0.923, which indicated a high level of concordance. It also showed that the p-value = 0.001 ¡ 0.05, thereby allowing the rejection of the null hypothesis that there is no agreement among the participants. Even with the inclusion of the anomalous results Kendall's value still shows a moderate degree of agreement at 0.676. Further confirmation of the closeness of the correlation can be seen in figure 4.7 with a dendrogram where the actual Photoshop-based rankings are tabulated in the right-hand column

The results of task 3B demonstrated similar problems to those of 3A and the same procedure as above was performed in mitigation except that, in this case, a total of seven participants' answers were rejected. It is a possibility that, once again, the participants chose to group rather than rank the images, but the fact is that eight participants, given a scale from 1 to 12, failed to evaluate any of the images at a higher level than 4. This did not even approximate to the answers given by the majority of the participants, and was totally in disagreement with their own evaluation of the same images in task 3A. For this reason participant numbers 1, 3, 5, 6, 12, 14, 22 and 31 were excluded from the analysis. Analysis of the data was conducted as in the previous task and the results are presented in tables 4.11 to 4.13. Calculations of Spearman's rho, per participant, may be accessed via the Dropbox link at http://bit.ly/1l5WyTy.

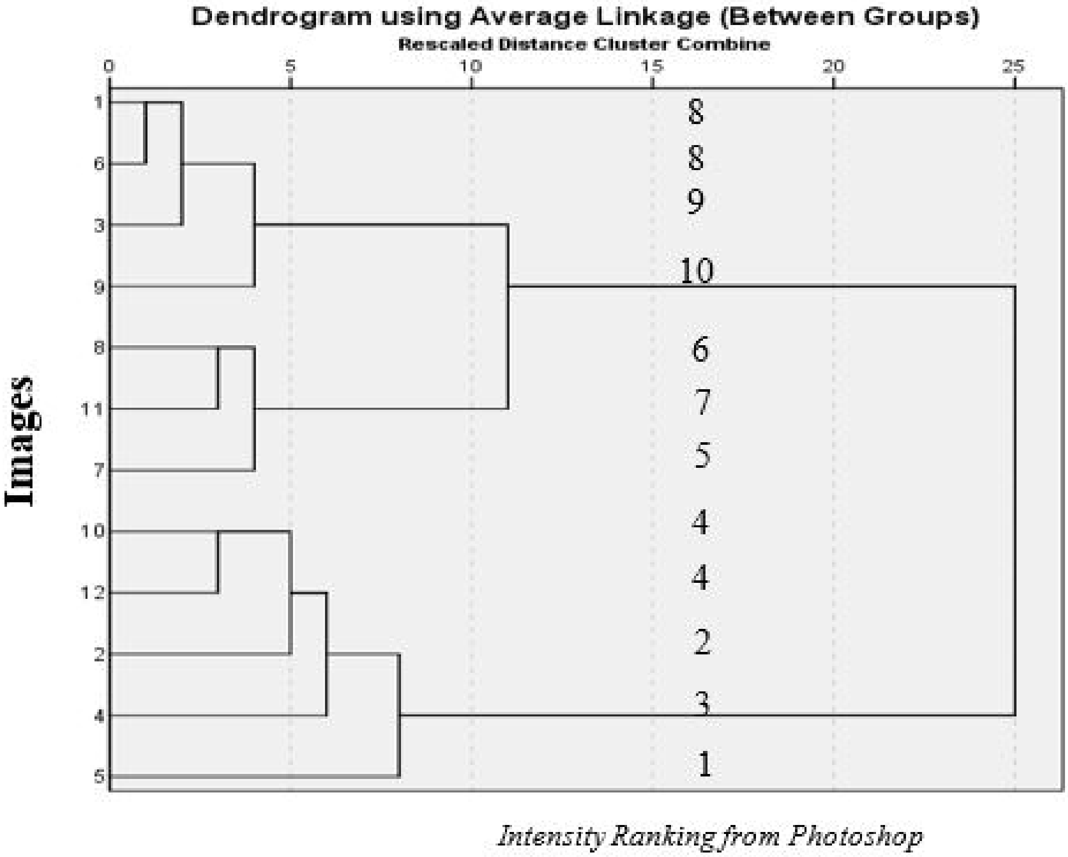


Figure 4.14: The effect on participants' performance of using a colour guide

Table 4.9: The ranking of participants and values in Photoshop (Task 3B)

Images	Photoshop Ranking	Mean of Friedman & Kendall's of participants ranking
Image 1	9.50	9.48
Image 2	2.00	2.54
Image 3	11.00	10.39
Image 4	3.00	2.76
Image 5	1.00	2.85
Image 6	9.50	9.57
Image 7	6.00	5.63
Image 8	7.00	7.65
Image 9	12.00	11.17
Image 10	4.50	3.52
Image 11	8.00	7.65
Image 12	4.50	2.54

For the 23 participants, both Friedman's test and Kendall's coefficient produced results of, chi-square = 200.977 and Df =11. A moderately high degree of concordance was indicated by Kendall's W = 0.794. The dendrogram representation of the results (figure 4.8) clearly shows that the ranking of the images by the participants still closely matches the ranking values based on image metadata in Photoshop. The Photoshop values are shown in the right-hand column.

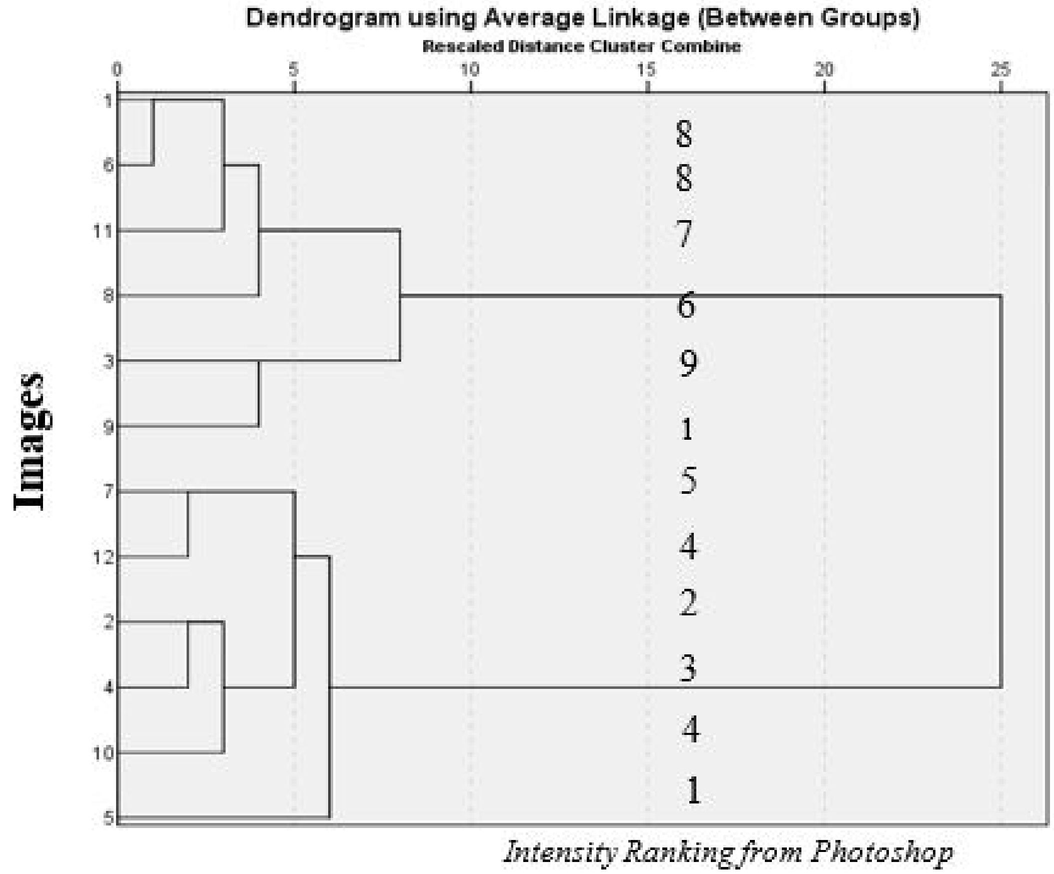


Figure 4.15: The effect on participants' performance of using a colour guide

A comparison of the results of tasks 3A and 3B reveals that there is no significant influence on the results whether guidance is or is not provided. Statistically, the differences between pairs of results are minor and no clear case can be made in favour of either procedure. There are several possible reasons behind this;

- The participants were looking at plain computer-generated colour chips with no context.

- The repetitive nature of the tasks may have resulted in inaccuracies due to overconfidence.

- The additional colour guidance provided may have been insufficiently clear.
- The anomalies, which were observed in the results, would tend to indicate that instructions were not sufficiently clear.

4.8 Critical Discussion

The following section is a critical discussion of the pilot study results and of the study's limitations. This discussion also includes material prefatory to the next series of experiments. Preliminary examination of the results made it clear that merely asking participants to describe a colour (as in task1) resulted in a wide variety of descriptors. This was probably due to individual differences in knowledge and in experience with colours. However, in task 2 and 3, where participants were asked to cluster and rank colours, the consistency of the results was much higher. The only significant difference between the descriptive experiments and the clustering and ranking experiments was the existence of a pre-defined scale of measurement. This shows that a common language, such as a scale, aided standardisation of answers and ensured greater accuracy.

Comparing experiment 2A and experiment 2B, it was notable that there was an increase in the number of clusters and subclusters created by the participants. This may have been due to the additional instructions or the participants' experiences from task 2A. It appears that being given a common focus in ranking and classifying colour produced different results from the use of personal judgment alone, as such is more subject to individual differences. As a result of these findings, subsequent experiments were designed to incorporate a colour scale to be utilised as a common language. This was intended to familiarise participants with the concepts of colour characteristics and to expose them to model examples with accurate colour descriptions.

There is much controversy regarding sex-related differences in visual processing. Relevant studies by Chaudhari and Shaw (2012); Abramov et al. (2012) and Kuehni (2001) have already been described in the literature review in chapter 2. Thus, it would have been better had the sample included females for comparative purposes. Unfortunately, it was not possible to recruit women for the study as none volunteered to take part.

4.9 Conclusion

In summary, the pilot study used non-medical images and colour chips the experiments undertaken using them showed a high level of consistency in the answers of the participants in the majority of tasks. Such tasks involved the identification, description and classification of colours. Following the pilot, a furher series of experiments were designed which used images showing actual clinical cases of cellulitis (chapters 5 and 6) and conjunctivitis (chapter 7). The images were collected from the A&E department of ARI. These experiments also investigated the confidence of participants in their answers as an additional indication of their accuracy level in relation to colour perception.

At this stage, it was considered that the general structure and methodology of the experimental design was suitable for application in subsequent experiments. A lesson was learned from the apparent misinterpretation of the questions posed and the consequent effects on the results. Applying the principles of agile methodology, the questions and instructions were worded more precisely for subsequent experiments.

Chapter 5

Skin Infection Experimental Series

5.1 Introduction

Chapter 5 presents a series of experiments which utilised clinical images of cellulitis, a skin infection. These images had the advantage of showing the red areas of infection in context with the surrounding tissue. Applying the agile methodology discussed earlier it was deemed appropriate to make a transition from plain colour to contextual colour in order to avoid the problems encountered in the pilots.

5.2 Cellulitis Experimental Objectives

1. To investigate the consistency and accuracy among participants when describing, clustering (grouping), and ranking colours.
2. To compare accuracy, confidence, and time between doctors and non-doctors when judging colours in digital images that show cellulitis.
3. To test the relationship between accuracy, confidence, and time when classifying colours.
4. To test to what extent the participants were able to discern different levels of image quality when comparing images that showed the colour red.
5. To investigate whether the participants were able to recognise subtle differences in colour characteristics such as brightness, saturation, and intensity (chapter

2.4, “Understanding colour and digital images” and section 2.5.1 “Human colour perception in key previous studies”).

6. To test the confidence of the participants in their answers and how the objective confidence of the participants in their answers correlates with their actual accuracy level (accuracy being defined in terms of the number of correct answers).
7. To make a comparison between the relative accuracy of doctor and non-doctor groups.

Five principal experiments were conducted and then repeated either with modified guidance or with no changes, in order to directly compare the accuracy of repeated procedures. For convenience of identification the experiments were numbered from 1 to 5 with suffixes A and B. The tasks associated with the experiments are outlined below.

5.3 Experimental Tasks

- 1A. The participants were presented with 5 images of skin infection and were asked to describe, in their own words, the degree of redness.
- 1B. The above task was repeated but this time the most common descriptors appearing in the results of experiment 1A were provided as categories into which the images were to be assigned.
- 2A. Nine images were provided and the participants were asked to group them by degrees of similarity.
- 2B. This was a repeat of 2A, except that categories of “high”, “medium” and “low” were provided for the participants’ guidance.
- 3A. The same images as those used in experiment 2 were provided and the participants were asked to rank the images by intensity of colour.
- 3B. This was a repeat of 3A except that a scale of intensity from 1 to 9 was provided, into which the images were to be categorised.
- 4A. A total of 144 images were supplied, being 12 images each with 12 progressive degradations in image quality. The participants were asked to rate the images in terms of their quality and definition.
- 4B. This was a repeat of 4A but with a different group of participants, and using only 36 images (based on 12 quality-degraded versions of 3 original images). These images were a subset of those which had been used in experiment 4A.

- 5A. 12 pairs of images were provided and the participants were asked to compare and grade each pair in relation to the other, based on a binary same or different).
- 5B. This was a repeat of experiment 5A, but using a different group of participants.

5.4 Participants

In experiments 1A to 5A, a total of 190 participants took part. They comprised 73 non-medically-qualified students of computing, and 64 medical doctors. On this occasion, 24 female participants were included, which made the sample more representative.

In experiments 1B to 5B, a total of 53 participants took part, 23 of whom were qualified medical doctors and 30 non-medically-qualified computing students. In this instance, 8 female participants took part.

5.5 Design and Variables

Although these experiments took place in two different buildings, great care was taken to ensure that the two environments were as close to identical as was possible, in terms of ambient lighting and supplementary electric lighting.

Computers and display equipment were standardised and independent checks were carried out by a professional medical illustrator to ensure that the two sets of circumstances were as near as possible to identical. All participants received the same instructions and were given the same time to perform the experimental tasks.

The experiments were designed with handouts that contained the images, instructions for performing the experimental tasks and blank answer sheets. All tasks were explained to the participants in both verbal and written form.

The independent variables were the tasks of the experiments, colour characteristics, the use of a colour scale, image quality, confidence level, and the environmental conditions. The dependent variables were the participants' responses to the experiments (in describing, grouping, ranking and rating the images).

5.6 Materials

The following were the materials and tools used in this experiment:

5.6.1 Images used for the experiments

The 5 images used in experiments 1A and 1B; the 9 images used in experiments 2A, 2B, 3A, 3B; the 144 images used in experiments 4A and 4B and the 12 pairs of images used in experiments 5A and 5B may all be accessed via the Dropbox link at http://bit.ly/1l5WyTy. The original unmodified images may also be accessed in the same location. Examples of images used in the experiments are reproduced in figure 5.1.

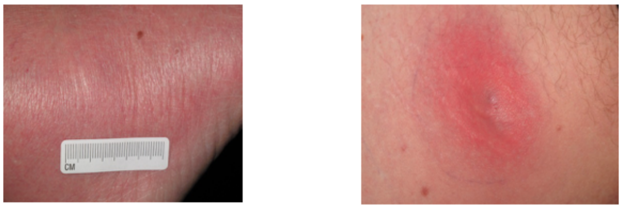


Image 3

Image 5

Figure 5.1: Examples of images from 1A and 1B

Q: Experiments 1A and 1B. Describe the colour based on redness and state the level of confidence.

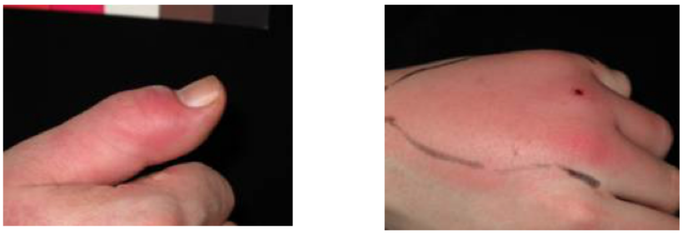


Image 4

Image 6

Figure 5.2: Examples of images from 2A and 2B

Q: Experiments 2A and 2B. Group the images based on intensity of redness and state the level of confidence.

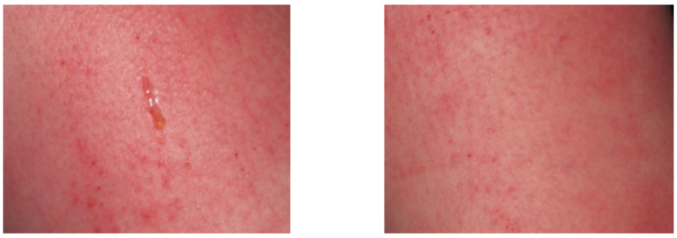


Image 3 Image 7

Figure 5.3: Examples of images from 3A and 3B

Q: Experiments 3A and 3B. Rank the images based on degree of redness and state the level of confidence.

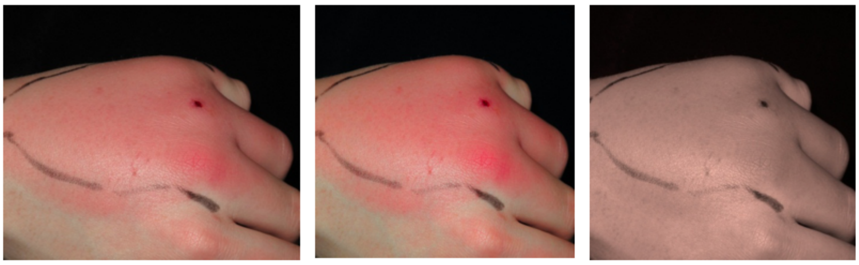


Image1 Image 2 Image3

Figure 5.4: Examples of images from 4A and 4B

Q: Experiments 4A and 4B. Rate the images based on image quality and state the level of confidence.

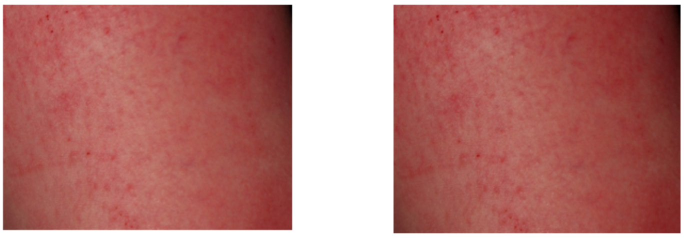


Image 5, Copy1 Image 5, Copy2

Figure 5.5: Examples of images from 5A and 5B

Q: Experiments 5A and 5B. State whether the images are the same or different, evaluate the degree of difference and state the level of confidence.

The questions related to each experiment were distributed in a random order, which ensured that the performance of the users was not influenced by the order of the tasks. All experimental data was collected in Microsoft Excel spreadsheet tables.

These contained participants' information and performance in describing, grouping, ranking, and rating the images provided. The data tables may be accessed via the Dropbox link at http://bit.ly/1l5WyTy.

Colour blindness test sheet
As the pilot experiment the same Ishihara plates were used to test participants for colour blindness (appendix number 8). All participants were assessed for colour blindness before participating in the experiment and the data analysed was only that applicable to users who were not colour blind.

Question and Answer sheets (experimental tasks)
The participants were given question and answer sheets along with the images. These sheets consisted of all the five tasks and provided space for the users to fill their answers. These sheets may be accessed via the Dropbox link at http://bit.ly/1l5WyTy.

Instruction sheet
Instructions were provided directly with the question sheets to guide the participants step by step (See examples in appendix 9).

Consent form
Formal consent was elicited from all participants for all experiments of this series (an example of the consent form used may be viewed in appendix 7.2). The participants were informed at the beginning of the experiment about all pertinent ethical issues, such as their right to withdraw at any time from the experiment without providing any reason.

5.7 The Results

Detailed below are the results of all experiments, together with analyses of the performance of all participants and the comparison between doctors and non-doctors.

5.7.1 Results for experiments 1A and 1B (image description)

In all five images, the majority of opinion amongst all participants was that the red area of infection could be described as “medium red”. What was interesting, however, was that after the introduction of additional guidance, in experiment 2B, there was a clear shift of opinion across the range in several cases. In image 1, for example, there was a 14-point movement from ‘dark’ to ‘medium’ and a corresponding shift of 17 points from ‘medium’ to ‘light’. A similar shift of opinion could also be observed in the case of image 5. This indicates that the presence of additional information had a definite effect on the participants’ judgement.

Table 5.1: Frequency of answers for all participants in experiments 1A and 1B

Images	Description	Percent Experiment 1A	Percent Experiment 1B
	1-Dark red	25.2	11.3
	2-Medium red	59.7	56.6
Image 1	3-Light red	15.1	32.1
	1-Dark red	14.3	3.8
	2-Medium red	78.2	71.7
Image 2	3-Light red	7.6	24.5
	1-Dark red	23.5	18.9
	2-Medium red	55.5	69.8
Image3	3-Light red	21	11.3
	1-Dark red	14.3	17
	2-Medium red	66.4	66
Image 4	3-Light red	19.3	17
	1-Dark red	21.8	18.9
	2-Medium red	66.4	52.8
Image 5	3-Light red	11.8	28.3

Table 5.2: Frequency of answers for doctor participants in experiments 1A and 1B

Images	Description	Experiment 1A (%)	Experiment 1B (%)
	1-Dark red	5.8	7.2
	2-Medium red	81.2	79.7
Image 1	3-Light red	13	13
	1-Dark red	4.3	4.3
	2-Medium red	82.6	88.4
Image 2	3-Light red	13	7.2
	1-Dark red	20.3	24.6
	2-Medium red	58	40.6
Image3	3-Light red	21.7	34.8
	1-Dark red	7.2	40.6
	2-Medium red	84.1	53.6
Image 4	3-Light red	8.7	5.8
	1-Dark red	10.1	40.6
	2-Medium red	78.3	53.6
Image 5	3-Light red	11.6	5.8

Table 5.2 shows the changes of opinion of participating doctors from experiment 1A (with no guidance) to experiment 1B (with guidance). The results show little shift in opinion in relation to images 1 and 2, but there is a clear shift in opinions in the cases of images 3, 4, and 5.

Table 5.3 shows the changes of opinion of participating non-doctors from experiment 1A (with no guidance) to experiment 1B (with guidance). The results show little shift in opinion in relation to images 1 and 2, but there is a clear shift in opinions in the cases of images 3, 4, and 5.

Table 5.3: Frequency of answers for non-doctor participants in experiments 1A and 1B

Images	Description	Experiment 1A (%)	Experiment 1B (%)
Image 1	1-Dark red	31.1	32
	2-Medium red	43.7	42.7
	3-Light red	25.2	25.2
Image 2	1-Dark red	15.5	20.4
	2-Medium red	71.8	58.3
	3-Light red	12.6	21.4
Image3	1-Dark red	23.3	21.4
	2-Medium red	61.2	27.2
	3-Light red	15.5	51.5
Image 4	1-Dark red	20.4	39.8
	2-Medium red	54.4	33
	3-Light red	25.2	27.2
Image 5	1-Dark red	28.2	53.4
	2-Medium red	51.5	26.2
	3-Light red	20.4	20.4

Table 5.4: Direct comparison of doctor and non-doctor participants in experiment 1A

Images	Description	Non-doctors (%)	Doctors (%)
Image 1	1-Dark red	31.1	5.8
	2-Medium red	43.7	81.2
	3-Light red	25.2	13
Image 2	1-Dark red	15.5	4.3
	2-Medium red	71.8	82.6
	3-Light red	12.6	13
Image3	1-Dark red	23.3	20.3
	2-Medium red	61.2	58
	3-Light red	15.5	21.7
Image 4	1-Dark red	20.4	7.2
	2-Medium red	54.4	84.1
	3-Light red	25.2	8.7
Image 5	1-Dark red	28.2	10.1
	2-Medium red	51.5	78.3
	3-Light red	20.4	11.6

Table 5.5: Direct comparison of doctor and non-doctor participants in experiment 1B

Images	Description	Non-doctors (%)	Doctors (%)
Image 1	1-Dark red	32	7.2
	2-Medium red	42.7	79.7
	3-Light red	25.2	13
Image 2	1-Dark red	20.4	4.3
	2-Medium red	58.3	88.4
	3-Light red	21.4	7.2
Image3	1-Dark red	21.4	24.6
	2-Medium red	27.2	40.6
	3-Light red	51.5	34.8
Image 4	1-Dark red	39.8	40.6
	2-Medium red	33	53.6
	3-Light red	27.2	5.8
Image 5	1-Dark red	53.4	40.6
	2-Medium red	26.2	53.6
	3-Light red	20.4	5.8

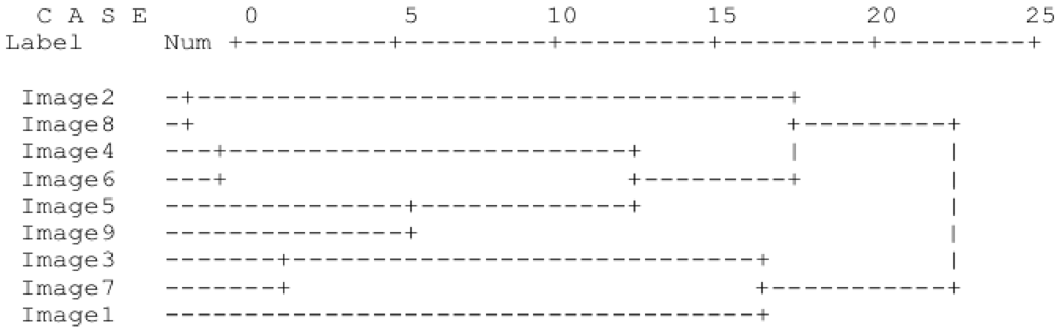
Table 5.4 represents a comparison of the evaluation of redness by non-doctors with that of doctors in experiment 1A (without guidance). Meanwhile Table 5.5 represents a comparison of the evaluation of redness by non-doctors with that of doctors in experiment 1B (with guidance).

With a few exceptions, there is a reasonable degree of concordance, in the evaluation of red intensity, between both doctor and non-doctor participants. It seems that a likely reason for this may be the absence of natural, uninfected, skin within the images that could have provided context for comparison with the inflamed areas. This may be a case where the training and experience of qualified doctors permitted them to judge the degree of red intensity differently from the non-medical participants. Of more significance is the fact that both groups were influenced to similar degrees by the additional guidance provided in experiment 1B. In the results for both groups, there is little change of opinion until image 3 where the assessment of both groups of participants begins to change. This continues to be the case through images 4 and 5. Again, this may be because of the presence in the images areas of unaffected tissue, which provided context for the inflamed areas and made visualisation of the degree of infection simpler.

In relation to research hypotheses 1 and 2 (chapter 3, section 3.7), the above experiments 1A and 1B confirm that there is a definite shift in opinion when a descriptive colour scale is introduced. However, although the doctors who took part in the experiment showed slightly more acumen in applying the scale, the results were not sufficiently conclusive in relation to hypothesis 1, i.e. that the ability of the doctors would be enhanced by their prior training.

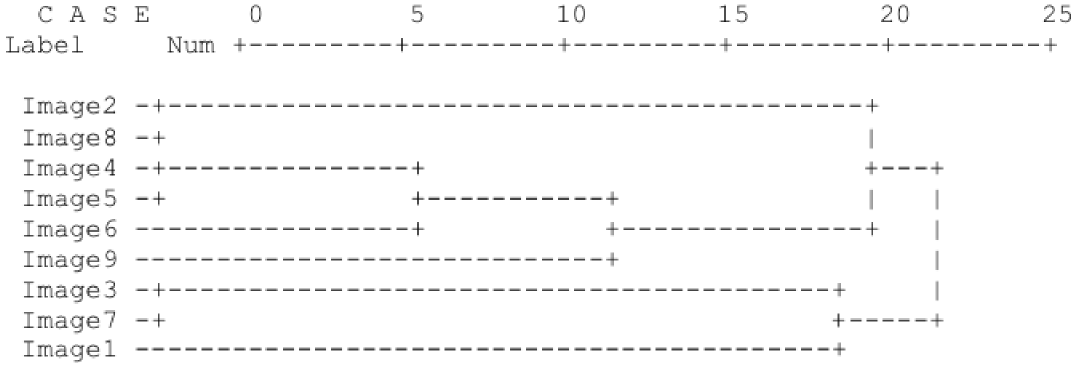
5.7.2 The results of experiments 2A and 2B (image clustering)

Figures 5.6 to 5.9 show hierarchical cluster analyses for image clustering experiments 2A and 2B. Dendrogram using average linkage (between clusters), rescaled distance cluster combine.



```
  C A S E      0         5        10        15        20        25
Label     Num +---------+---------+---------+---------+---------+

 Image2   -+-------------------------------------+
 Image8   -+                                     +---------+
 Image4   ---+-------------------------+         |         |
 Image6   ---+                         +---------+         |
 Image5   ---------------+-------------+                   |
 Image9   ---------------+                                 |
 Image3   -------+-----------------------------+           |
 Image7   -------+                             +-----------+
 Image1   -------------------------------------+
```

Figure 5.6: Dendrogram of image clustering in experiment 2A for non-doctors



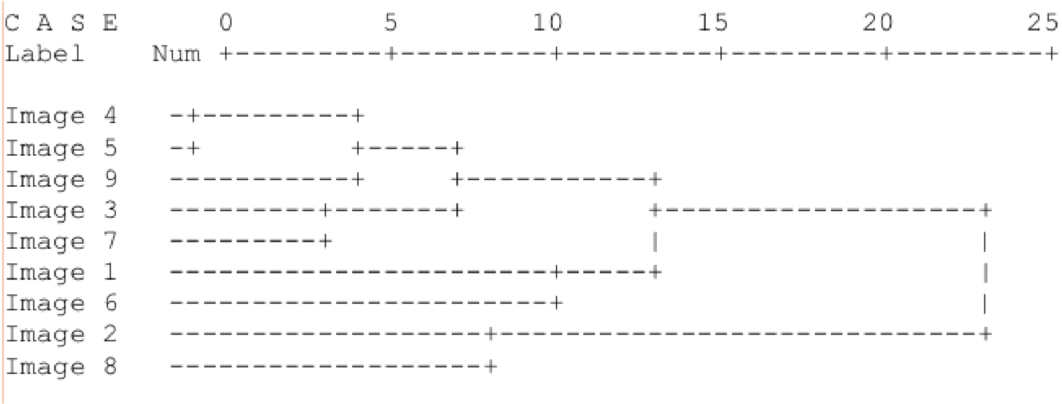
```
  C A S E      0         5        10        15        20        25
Label     Num +---------+---------+---------+---------+---------+

 Image2 -+-------------------------------------------+
 Image8 -+                                           |
 Image4 -+---------------+                           +---+
 Image5 -+               +-----------+               |   |
 Image6 -----------------+           +---------------+   |
 Image9 -----------------------------+                   |
 Image3 -+-----------------------------------------+     |
 Image7 -+                                         +-----+
 Image1 -------------------------------------------+
```

Figure 5.7: Dendrogram of image clustering in experiment 2A for doctors

The four dendrograms shown above were analysed and for clarity of presentation the results are summarised in tables 5.6 and 5.7 below.

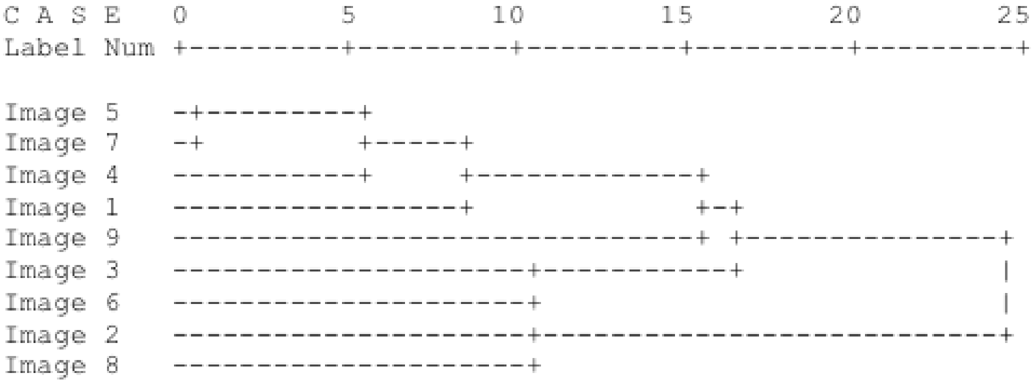
Table 5.6: Two-cluster solution in experiment 2A (image clustering)

Group	No. of participants	Cluster 1	Cluster2
Doctors	46	Images 2, 4, 5, 6, 8, 9	Images1,3,7
Non-doctors	73	Images 2, 4, 5, 6, 8, 9	Images1,3,7



```
C A S E      0         5        10        15        20        25
Label     Num +---------+---------+---------+---------+---------+

Image 4   -+---------+
Image 5   -+         +-----+
Image 9   -----------+     +-----------+
Image 3   ---------+-------+           +-------------------+
Image 7   ---------+                   |                   |
Image 1   -----------------------+-----+                   |
Image 6   -----------------------+                         |
Image 2   -------------------+-----------------------------+
Image 8   -------------------+
```

Figure 5.8: Dendrogram of image clustering in experiment 2B for non-doctors



```
C A S E   0         5        10        15        20        25
Label Num +---------+---------+---------+---------+---------+

Image 5   -+---------+
Image 7   -+         +-----+
Image 4   -----------+     +-------------+
Image 1   -----------------+             +-+
Image 9   -------------------------------+ +---------------+
Image 3   ---------------------+-----------+               |
Image 6   ---------------------+                           |
Image 2   ---------------------+---------------------------+
Image 8   ---------------------+
```

Figure 5.9: Dendrogram of image clustering in experiment 2B for doctors

Table 5.7: Two-cluster solution in experiment 2B (image clustering)

Group	No. of participants	Cluster 1	Cluster2
Doctors	23	Images 1, 3, 4, 5, 7, 9	Images 8,2,6
Non-doctors	30	Images 1,3,4,5,6,7,9	Images 8,2

As is clearly shown in tables 5.6 and 5.7, there is virtually no difference between doctor and non-doctor participants in clustering digital images into groups based on their colour intensity. As with experiment 1A and 1B, these experiments again fail to affirm hypothesis1 (chapter3, section 3.7), regarding the respective ability of doctors and non-doctors.

5.7.3 The results of experiments 3A and 3B (image ranking)

The following section presents the results of experiments 3A and 3B. It also provides a comparison of the performance of doctors and non-doctors and an analysis of the degree of agreement among participants' responses. Using the experience gained from previous experiments, and applying the principles of agile methodology, in experiment 3B the participants were given clear instructions that the numeric grading scale would begin at 1 for the lowest level of colour intensity and end at 9 for the most intense colours. This precaution avoided the occurrence of reversed scales as had been experienced in previous experiments.

Because there is variation of data in this case (9 images were to be ranked by each participant using a 9-point scale), Friedman's and Kendall's tests were used to show the degree of agreement among the participants in how they ranked the images. Kendall's coefficient of concordance is the most relevant in this case. (Actually, Friedman's and Kendall's are identical in this case because the data is already ranked). Kendall's coefficient ranges from 0, indicating no agreement among the sample individuals, to 1, indicating perfect agreement.

Table 5.8: Friedman and Kendall's values in experiment 3A

Key figures	Friedman Test Statistics	Kendall's Coefficient of Concordance Statistics
N	119	119
Chi-Square	236.766	236.766
Df	8	8
Asymp. Sig.	0.000	0.000
Kendal's WA		0.249

The Kendall's coefficient (W) value of 0.249 in table 5-8 indicates low agreement among the participants. In other words, the significance of Friedman's test and Kendall's coefficient indicate strong differences in the rankings.

Table 5.9: Average on 9-point scale by participants in experiment 3A

Images	Mean of Friedman & Kendall's for doctors	Mean of Friedman & Kendall's for non-doctors
Ranking _Image1	7.43	5.63
Ranking _Image2	4.83	4.86
Ranking _Image3	3.83	4.31
Ranking _Image4	4.07	5.08
Ranking _Image5	3.47	4.01
Ranking _Image6	6.22	6.58
Ranking _Image7	4.05	4.58
Ranking _Image8	7.15	7.21
Ranking _Image9	3.96	2.75

Table 5.10: Kendall's Coefficient of Concordance for doctors for experiment 3A

N	46
Kendall's Wa	0.320
Chi-Square	117.862
Df	8
Asymp. Sig.	0.000

Table 5.11: Kendall's Coefficient of Concordance for non-doctors for experiment 3A

N	73
Kendall's Wa	.248
Chi-Square	144.763
Df	8
Asymp. Sig.	0.000

Poor levels of concordance were demonstrated by both groups (tables 5.11 and 5.12), with a marginally better performance by doctor participants. Kendall's W for doctors was 0.320 and for non-doctors was 0.248. Low levels of agreement were demonstrated in table 5.9, despite the inclusion of a numeric scale, and precise notes on its application, within the question paper of experiment 3B.

The level of agreement in experiment 3B was actually lower than in 3A, although the results may have been skewed by the smaller sample size (53 participants in 3B as against 119 in 3A).

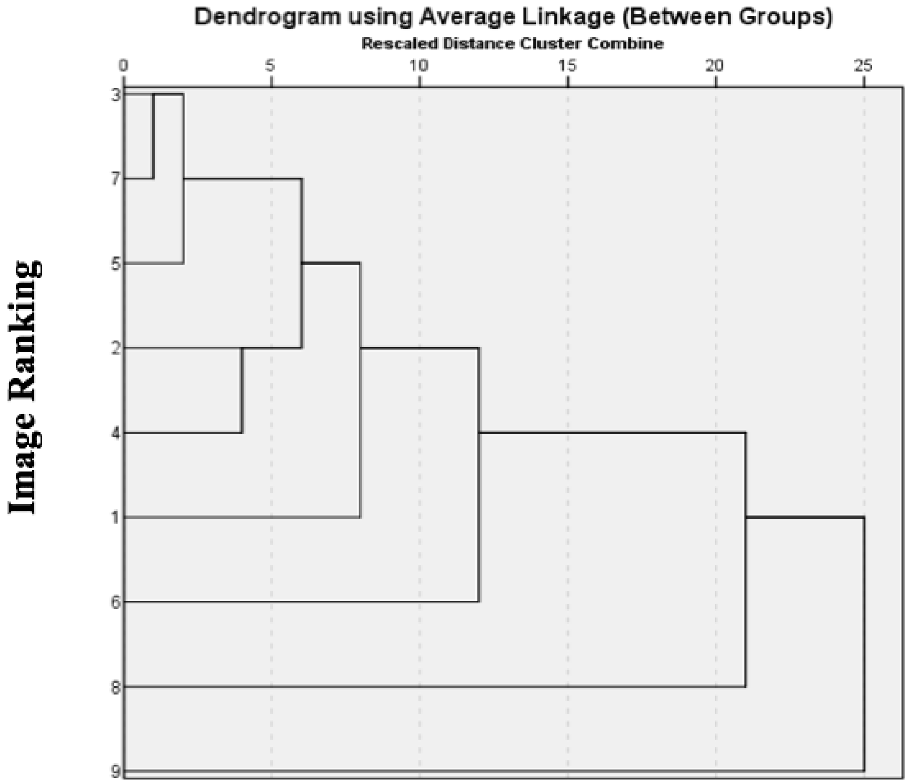


Figure 5.10: Dendrogram of image ranking by non-doctor participants in task 3A

Table 5.12: Friedman and Kendall's values in experiment 3B

Key figures	Friedman Test Statistics	Kendall's Coefficient of Concordance Statistics
N	53	53
Chi-Square	49.047	49.047
Df	8	8
Asymp. Sig.	0.000	0.000
Kendal's WA		0.116

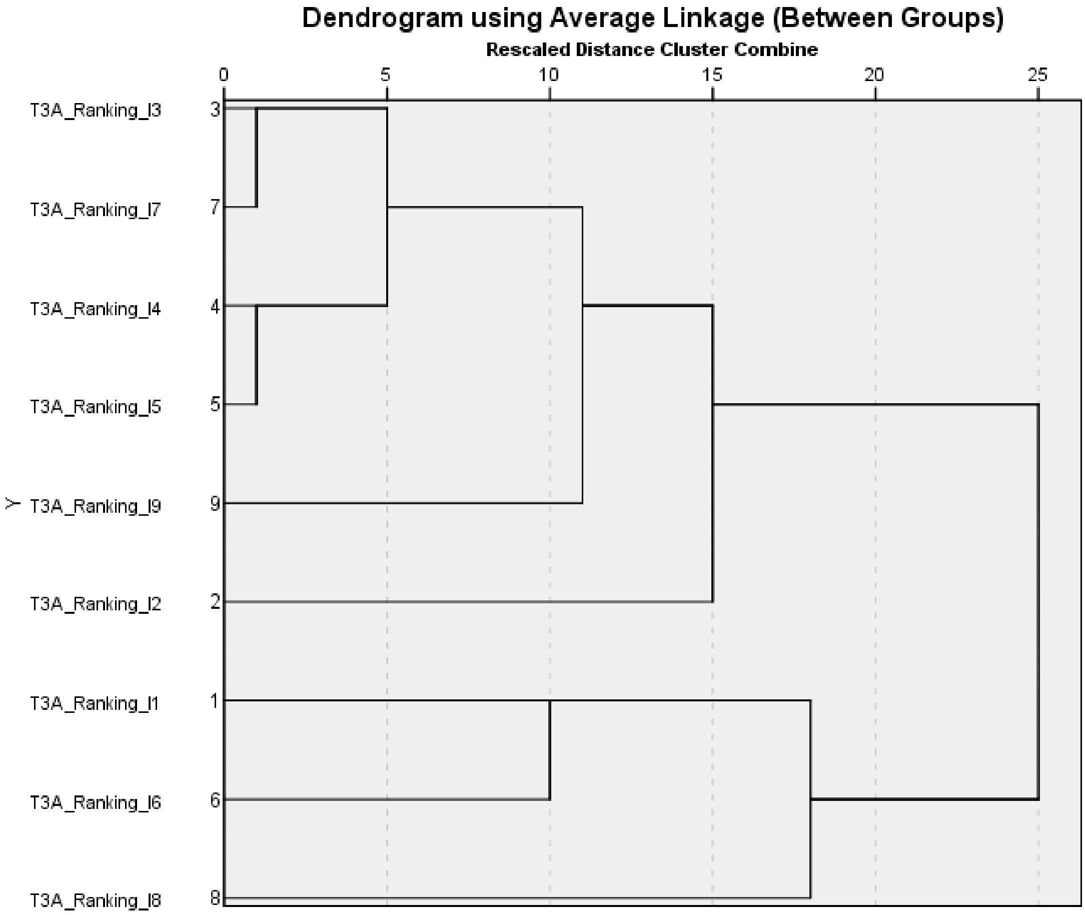


Figure 5.11: Dendrogram of image ranking by doctor participants in task 3A

Table 5.13: Average position of ranking by doctor and non-doctor participants in 3B

Images	Mean of Friedman & Kendall's for doctors	Mean of Friedman & Kendall's for non-doctors
Cellu_T3B_I1_Ranking	4.41	5.72
Cellu_T3B_I2_Ranking	3.54	5.43
Cellu_T3B_I3_Ranking	5.26	4.13
Cellu_T3B_I4_Ranking	5.43	4.98
Cellu_T3B_I5_Ranking	5.46	4.35
Cellu_T3B_I6_Ranking	5.72	4.63
Cellu_T3B_I7_Ranking	5.5	4.85
Cellu_T3B_I8_Ranking	3.17	5.03
Cellu_T3B_I9_Ranking	6.5	5.87

Table 5.14: Kendall's Coefficient of Concordance for doctors in experiment 3B

N	23
Kendall's Wa	0.165
Chi-Square	30.376
Df	8
Asymp. Sig.	0.000

Table 5.15: Kendall's Coefficient of Concordance for non-doctors in experiment 3B

N	30
Kendall's Wa	0.050
Chi-Square	12.103
Df	8
Asymp. Sig.	0.147

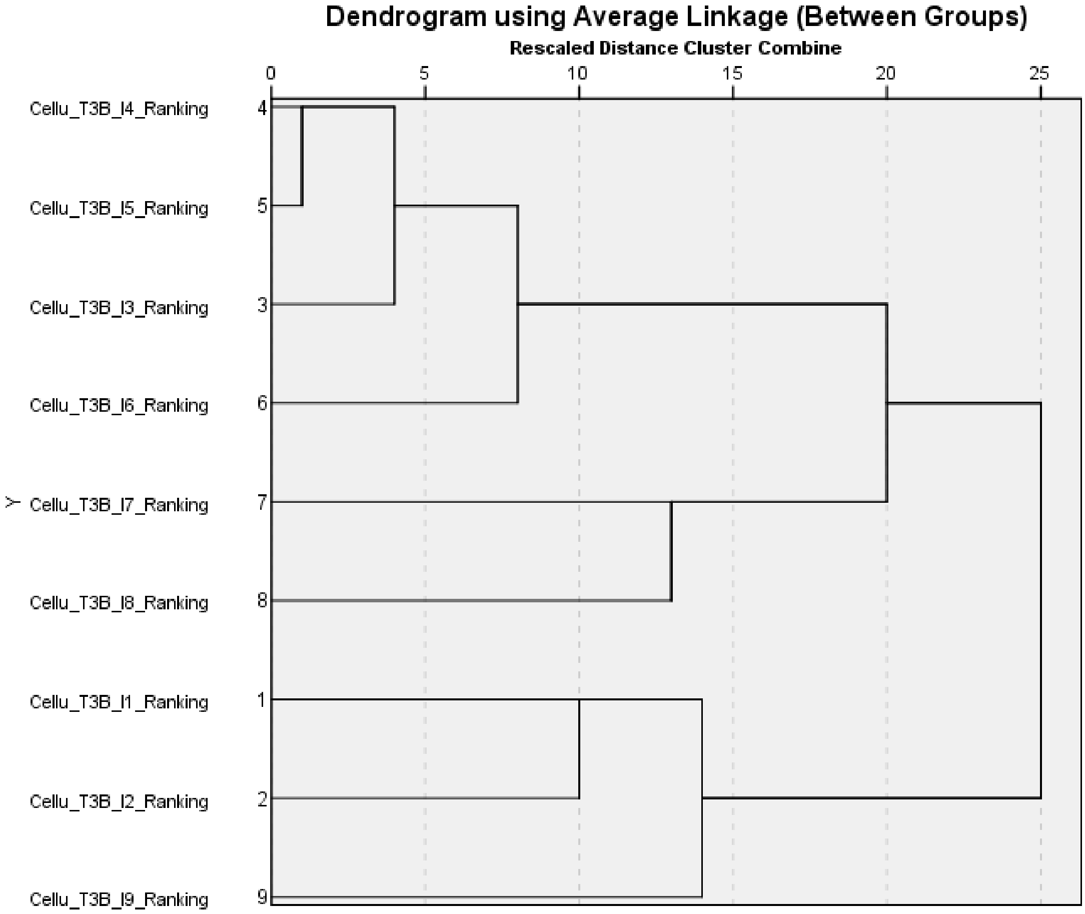


Figure 5.12: Dendrogram of image ranking by non-doctors in experiment 3B

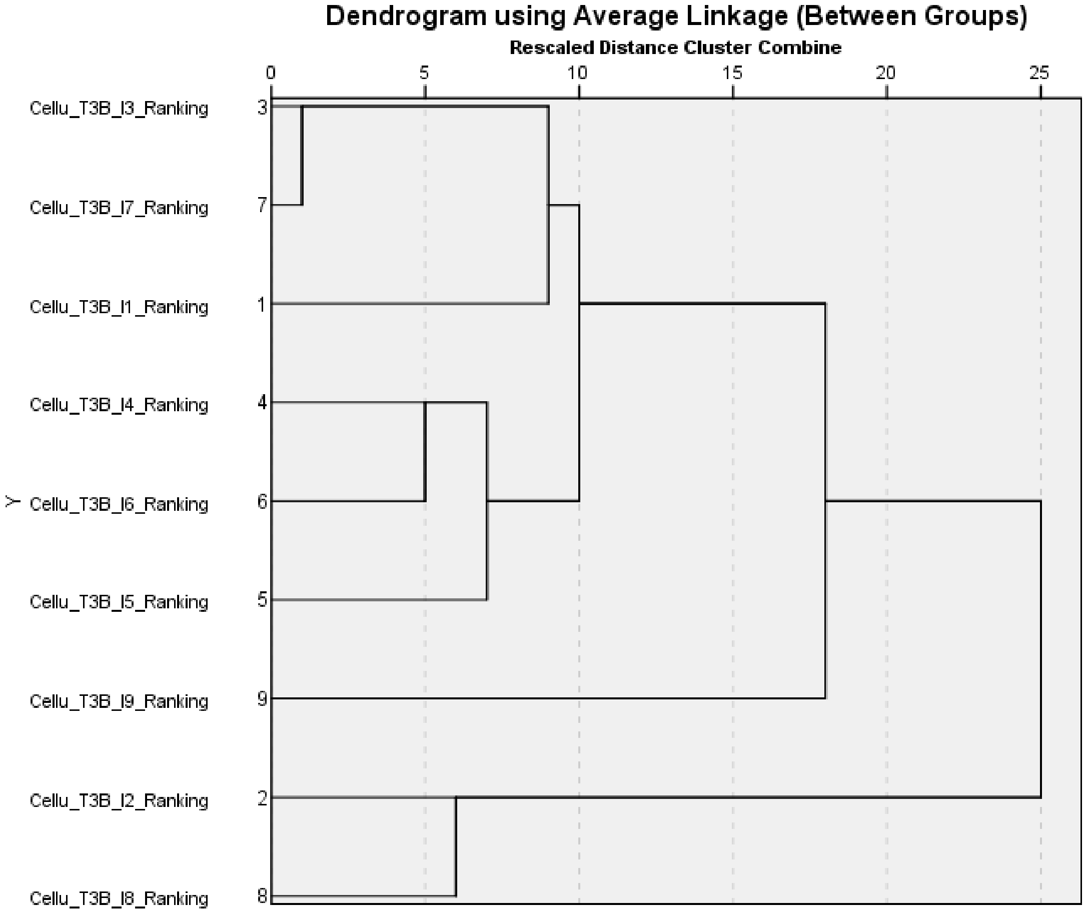


Figure 5.13: Dendrogram of image ranking by doctors in experiment 3B

Comparison of the performance of doctors and non-doctors in experiment 3B showed that there was little difference between the two groups. Both groups showed low levels of agreement, with doctors being only slightly more in accord. Kendall's W for doctors was 0.165 and for non-doctors was 0.050. The dendrograms in figures 5.10 to 5.13 graphically illustrate the generally random nature of the distribution and the lack of concordance.

It should be noted that the majority of the images displayed only subtle variations in redness, only a couple of the images displaying extremes (as, for example, in images 8 and 9).

There were several possible reasons for the lack of concordance within these experiments. Firstly, the nature of cellulitis makes judgement difficult as its appearance can vary, not only according to the degree of severity, but also with differing skin types. Secondly, the photographic images provided may not have contained sufficient context. It was notable that when the clusters within the dendrograms were compared to the experimental images, those images containing context in the form of areas of healthy skin were almost invariably clustered together (images 4, 5 and 6). Finally, it is possible that a scale as large as 1 to 9 may have proved to be excessive and that perhaps a more limited range of scale would be more appropriate. Consideration was also given to the use of a photographic frame of reference as a possible improvement over a numeric system. This was further examined within the conjunctivitis group of experiments.

5.7.4 Results of experiments 4A and 4B (image quality rating)

Kendall's coefficient of concordance was calculated for both doctor and non-doctor participants in experiment 4A. For 72 non-doctors, Kendall's W = 0.270, p = 0.001. For 46 doctors, Kendall's W = 0.353, p = 0.001. The statistics showed that there was a poor level of agreement between the participants in both groups.

Comparison of the two groups was carried out using Mann-Whitney U tests. The results ranked non-doctors slightly higher than doctors at 63 vs 55. Mann-Whitney U = 1457, Z = -1.211 and p = 0.226 showed that the differences between the two groups was not statistically significant.

As a further confirmation, and in order to assist with analysis of the results a boxplot diagram was generated that displayed the correlation between the answers of individual participants and the true values of the images (obtained from Photoshop modifications made to the images) for both doctor and non-doctor participants. As experiment 4B was a cut-down repeat of 4A, it was decided to carry out analysis of the reasons for the inconsistencies after consideration of the basic results of 4B.

Kendall's coefficient of concordance was calculated for both doctor and non-doctor participants in experiment 4B. For 29 non-doctors, Kendall's W = 0.218, p = 0.001. For 23 doctors, Kendall's W = 0.470, p = 0.001. The statistics showed that there was a poor level of agreement between the participants in both groups. Comparison of the two groups was carried out using Mann-Whitney U tests. The results ranked non-doctors slightly lower than doctors at 26 vs 27. Mann-Whitney U = 332, Z = -0.233 and p = 0.816 showed that the differences between the two groups was not statistically significant. The difference between the performances of the two groups of participants becomes very clear when presented in a boxplot format (figure 5.15).

Examination of the responses of all participants to the questions brought a number of issues to light. It is fair to say that, in all cases, the results were poor in terms of both accuracy and consistency. It was clear from the data that, had the results of the participants been averaged, they would, in the majority of questions, have been

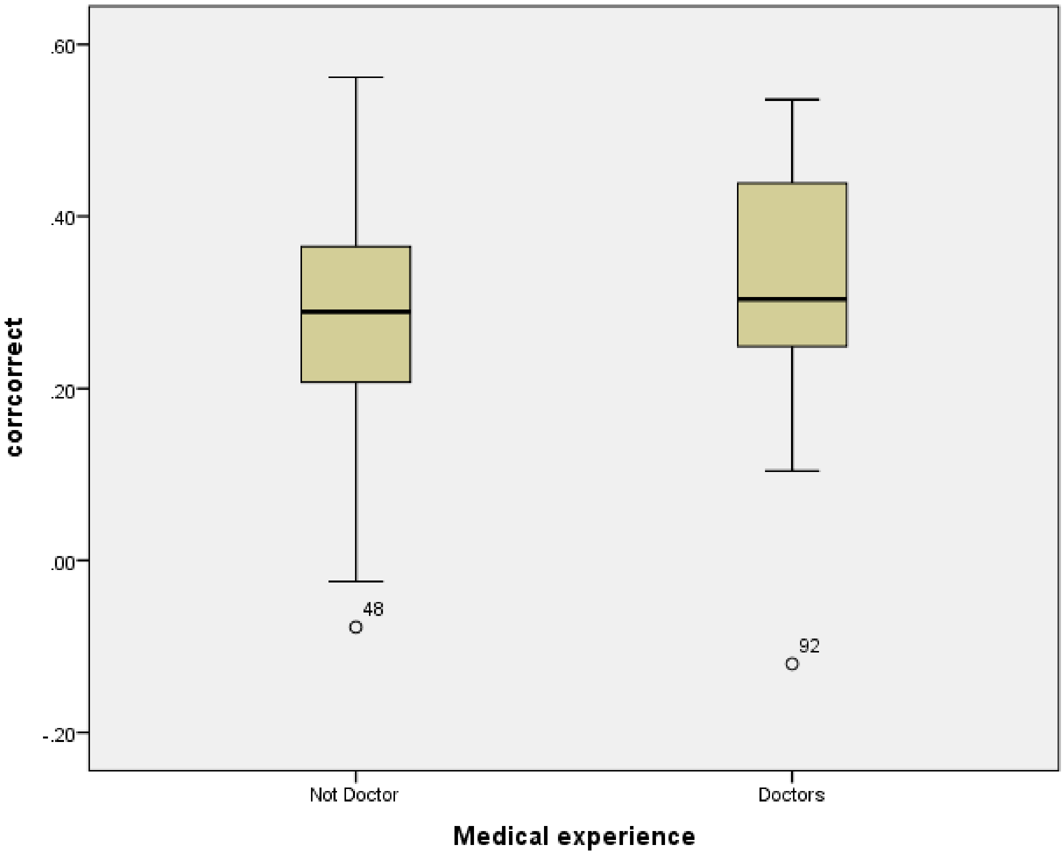


Figure 5.14: Boxplot of correlation in experiment 4A

close to the correct result (albeit with large degrees of variation), except in the case of the highest quality images. There appears to have been an aversion to the use of the higher values in the scale. Even the highest quality images (those rated at 9) were evaluated at only 5 or 6 by the participants, who applied similar values to those images which were a true 5 or 6. This raised a question as to whether the 9-point scale of reference was too large and if a narrower scale, of say 5 points, might have been more appropriate. Another view of this may be to suggest that human ability to differentiate between levels of image quality peaks at a point below that which we are able to attain with modern imaging equipment. A further consideration is that perhaps the degrees of degradation used in the experimental images were too close together to be accurately discerned.

It was apparent that, once again, the generally poor levels of concordance were exacerbated by a number of errors which would be hard to explain other than to assume that they were the result of lapses in concentration. In several instances participants who had evaluated the images reasonably closely to the true values, inexplicably rated

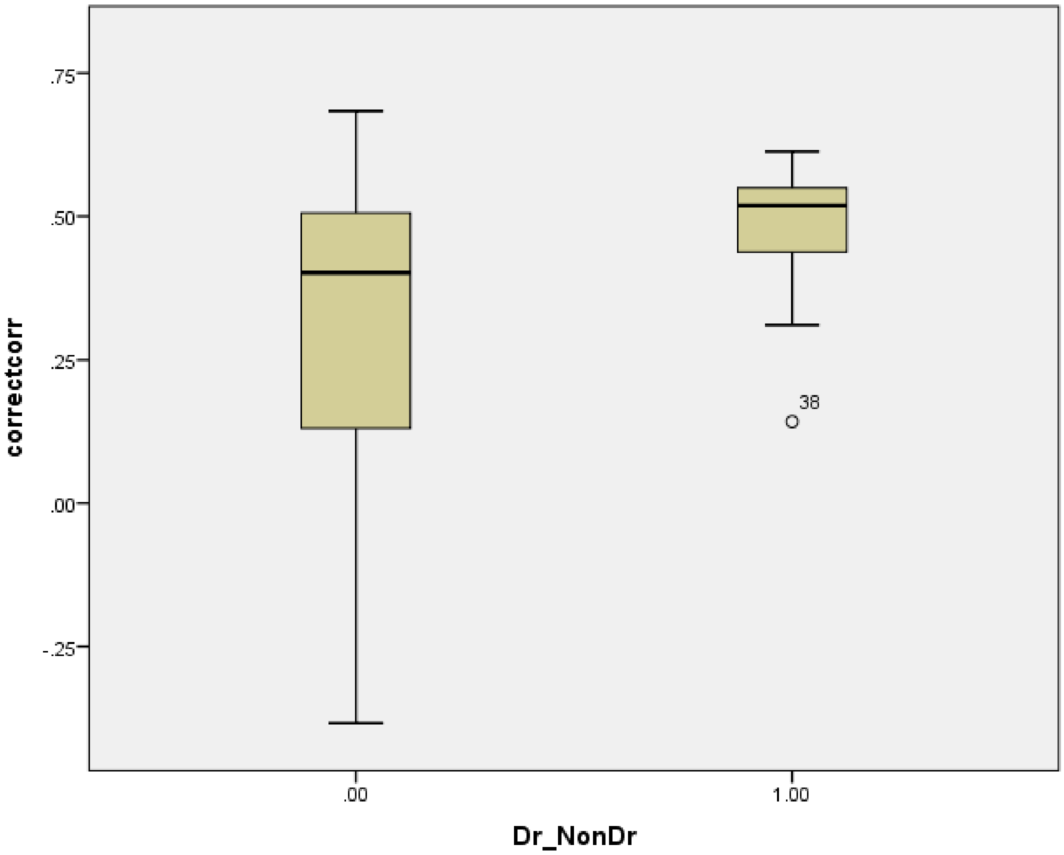


Figure 5.15: Boxplot of correlation in experiment 4B

images with true values of 1 or 2 as being a 9. This latter represented a total reversal of the scale and could only be interpreted as inattention in application of it. Such lapses may be unsurprising in view of the fact that the participants were dealing with large numbers of images and, as was explained in the literature review, it has been shown that such undemanding and repetitive work may lead to such slips (Reason 2006).

Table 5.16: The correct answers when images were the same

Images	Intensity difference	Experiment 5A		Experiment 5B	
		Doctors	Non-doctors	Doctors	Non-doctors
Image4	None/same	70	64	30.4	56.7
Image8	None/same	78	62	39.1	50
Image12	None/same	54	55	60.9	56.7

5.7.5 Results of experiments 5A and 5B (image matching)

Table 5.17: The correct answers when images had slight differences

Images	Intensity difference	Experiment 5A		Experiment 5B	
		Doctors	Non-doctors	Doctors	Non-doctors
Image1	Slight	39	44	56.5	76.7
Image3	Slight	76	75	87	76.7
Image5	Slight	59	42	26.1	43.3
Image6	Slight	37	40	34.8	36.7
Image7	Slight	67	41	43.5	60
Image9	Slight	30	33	26.1	20
Image11	Slight	61	60	73.9	70

Table 5.18: The correct answers when images had moderate differences

Images	Intensity difference	Experiment 5A		Experiment 5B	
		Doctors	Non-doctors	Doctors	Non-doctors
Image2	Moderate	74.0	66.0	100.0	66.7
Image8	None/same	74.0	78.0	91.3	96.7

Mann-Whitney U test comparison for experiment 5A, showed 73 non-doctors ranked at 56, slightly lower than the 46 doctors who were ranked at 67. Mann-Whitney U = 1374, Z = -1.681, p = 0.093, demonstrating no significant difference between the groups. In the case of experiment 5B, 30 non-doctors ranked slightly higher than doctors, 29 vs 23, although the tests showed that, once again, the statistical difference between the two was not significant (U = 264, Z = -1.477, p = 0.140).

In this series of experiments the only difference between conditions was that, a different group of participants were used in 5B. Data for both experiments (5a and 5b) was joint and analysed, Man Whitney test was conducted and the results matched the result of both experiments individually, with no statistically significant difference (U = 3312, Z = -1.761, p = 0.447) between doctors (69) and non-doctors (103).

Statistical comparison of the performances of doctors and non-doctors, using Mann-Whitney U tests, indicated that there was no significant difference between the two groups in either experiment 5A (p = 0.093) or 5B (p = 0.140). It was considered appropriate, therefore, to combine the results of the two experiments and to analyse the outcomes as though they had been a single experiment. Graphical representations of the percentage of accurate answers for doctors, non-doctors and for the two groups combined were generated for comparison purposes. The participants were asked only to state whether the image pairs were the same or different, but, in the graphs, the images which were different from each other were further sub-divided into those with slight and those with moderate differences (based on the file metadata in Adobe Photoshop). This was done in order to visualise the effect on the accuracy level of the participant.

Friedman test was performed in order to confirm the effects of image difference levels on the agreement levels of all 172 participants.

Table 5.19: Friedman Test (ranking) for all participants

Image categories	Mean Rank
No difference	1.87
Slight difference	1.65
Moderate difference	2.48

Table 5.20: Agreement level between the image categories for all participants

N	172
Chi-Square	68.417
Df	2
Asymp. Sig.	0.000

The data presented in tables 5-19 and 5-20 showed clear and statistically significant differences among the three categories for all participant groups when combined together.

Table 5.21: Friedman Test (ranking) for non-doctors

Image categories	Mean Rank
No difference	1.94
Slight difference	1.63
Moderate difference	2.43

Table 5.22: Agreement level between the image categories for non-doctors

N	103
Chi-Square	36.348
Df	2
Asymp. Sig.	0.000

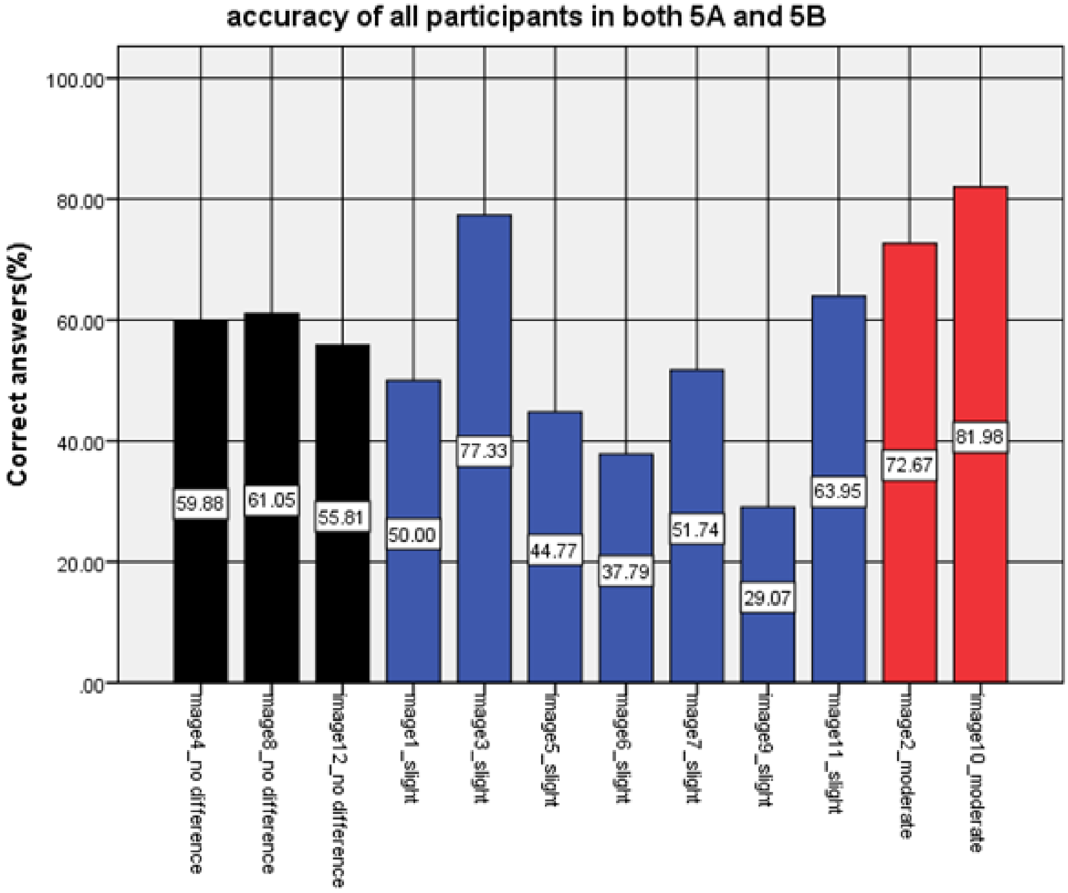


Figure 5.16: Comparison of accuracy with image differences

The data presented in tables 5-21 and 5-22 also showed clear and statically significant differences among the three categories in the case of non-doctor participants.

Table 5.23: Friedman Test (ranking) for doctors

Image categories	Mean Rank
No difference	1.76
Slight difference	1.69
Moderate difference	2.55

Table 5.24: Agreement level between the image categories for doctors

N	69
Chi-Square	34.443
Df	2
Asymp. Sig.	0.000

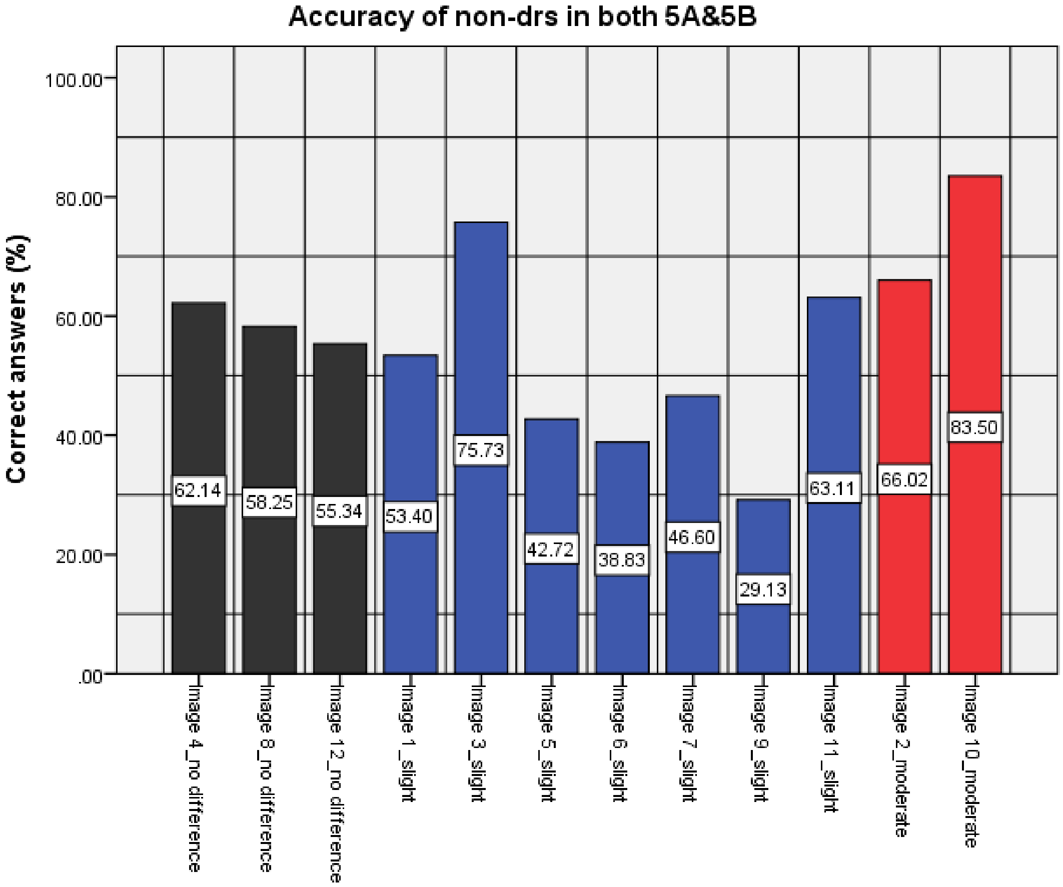


Figure 5.17: Comparison of accuracy with image differences

The data shown in tables 5-23 and 5-24 revealed clear and statically significant differences among the three categories in the case of doctor participants.

The graphs clearly show that the accuracy of the participants was generally higher in the case of the identical images, and where the difference between the two images was moderate. This would tend to confirm the findings of earlier experiments, where the results suggested that the degree of difference in some of the images was too subtle for humans to distinguish between them. The Friedman tests confirmed the graphical representations and showed that the differences were statistically significant. The notable exceptions to this generalisation occurred in images 3 and 11. In these images a large amount of contextual background was present which, as was concluded in previous experiments, probably contributed to the overall accuracy of the evaluation. As in the majority of previous experiments there was little difference shown in the respective accuracy of doctors and non-doctors in the experiment.

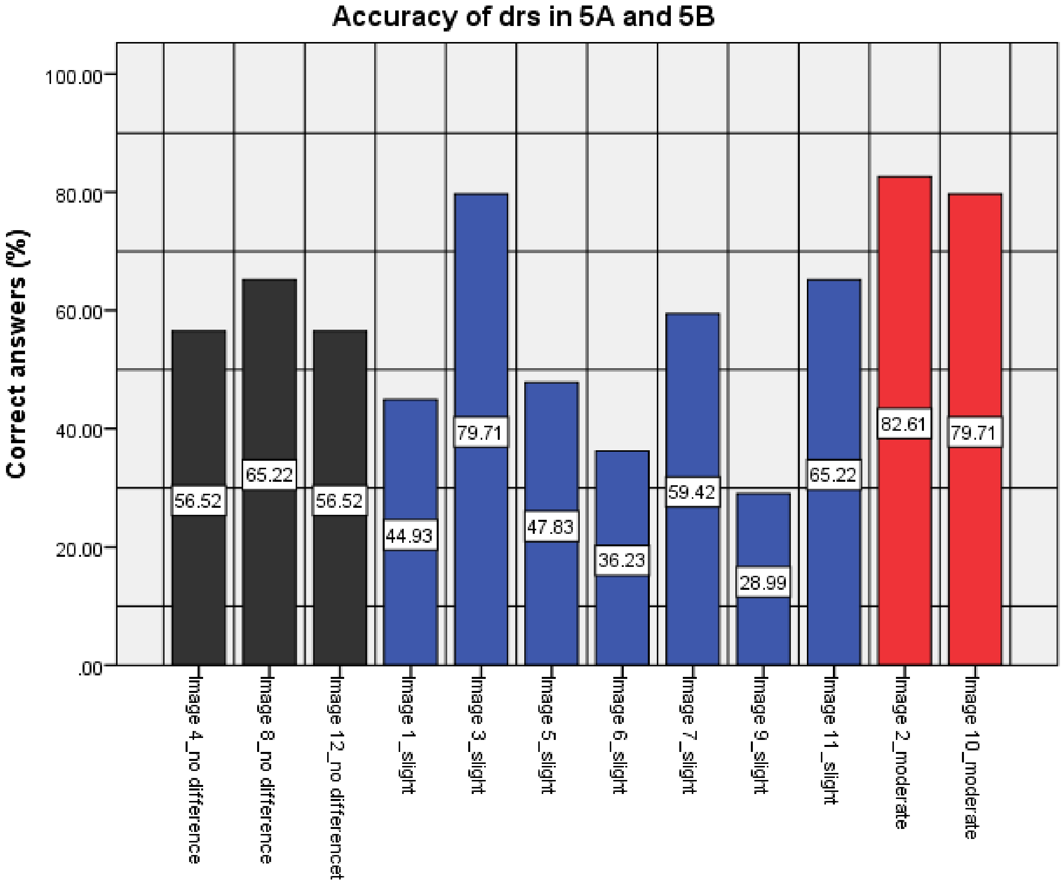


Figure 5.18: Comparison of accuracy with image differences

Examination of the design of experiment 5 led to the conclusion that the numbers of images used in each category was unbalanced and insufficient for truly accurate analysis. It was also clear that experiment 5 was a key part of the work, as it examined several key areas of the study within one experimental design. It was decided, therefore, that a full repeat of this experiment would be conducted, prior to moving on to the study of conjunctivitis. Chapter six contains the details of the repeated experiment 5 with a statistically more representative sample of images and with a larger population of different participants.

The confidence of the participants and its relationship with their accuracy

In all of the experiments described in chapter 5, the participants were asked to report the level of confidence that they had in their answers. Given the number of experiments involved, it would have been reasonable to expect that a range of confidence levels would have been recorded, reflecting the degree of difficulty experienced within the individual

Table 5.25: Mean confidence for all participants in experiment 4

Overall confidence	Mean	Std. Deviation	N
Experiment 4A	6.4	1.4	119
Experiment 4A for doctors	6.4	1.5	46
Experiment 4A for non-doctors	6.4	1.2	73

Table 5.26: Mean accuracy for all participants for experiment 4

Overall confidence	Mean	Std. Deviation	N
Experiment 4A	0.55	0.09	119
Experiment 4A for doctors	0.57	0.11	46
Experiment 4A for non-doctors	0.54	0.09	73

experiments. Surprisingly, however, both doctors and non-doctors, in all experiments, recorded confidence levels of between 6 and 7 (standard deviation of 1) on a scale of 0 (no confidence) to 9 (full confidence).

With a degree of confidence as consistent as this, it would be expected that the same levels of consistent accuracy would have been demonstrated in the experiments. This was not the case however.

Experiments 1, 2 and 3 involved subjective judgement in assessing, grouping and ranking images and therefore no measurable accuracy was involved. It should still have been the case, however, that high confidence should have been reflected in the consistency of the answers. Whilst this was true of experiment 1, experiments 2 and 3 showed considerably less agreement but with no change in confidence levels.

In the case of experiments 4 and 5, specific measurement of accuracy was possible. In both experiments, the accuracy for both groups of participants was of the order 55% to 56%, but still the level of confidence remained the same among all participants. Examples of the confidence and accuracy results for experiment 4 are shown in tables 5.26 and 5.27.

Spearman's ' correlations were conducted in experiments 4 and 5 and an analysis of the results is presented in this chapter. Spearman's rho correlation was chosen as the data was non-parametric. The values for (') are always between -1 and +1 and the closer the values are to either -1 or +1 the stronger the correlation (Hayes, 2000). Positive correlation occurs when the two related values increase or decrease together. Movement of the two values in opposite directions results in a negative correlation. Spearman correlation of the figures for experiment 4A are shown in table 5.27 (Correlation Coefficient = -.283, negative poor correlation). This is an example only and all other analyses may be accessed via the Dropbox link at http://bit.ly/1l5WyTy.

It may be that what was encountered here was an example of overconfidence due to

Table 5.27: Spearman correlation between confidence and accuracy for all participants

Spearman's rho correlations		Overall confidence (across 12 images)	Overall accuracy (across 12 images)
Overall confidence (across 12 images)	Correlation coefficient Sig. (2-tailed)	1 .	-.283** 0.002
Overall accuracy (across 12 images)	Correlation coefficient Sig. (2-tailed) N	-.283** 0.002 119	1 . 119

what were perceived as undemanding repetitive tasks (Reason 2006) or it is equally possible that the participants gave no consideration to their true confidence levels and simply inserted what they considered to be an acceptable number. In either case, it is apparent that in this group of experiments there was no correlation whatsoever between confidence and accuracy. This would confirm the null hypothesis.

5.8 Critical Discussion and Conclusion

It was apparent from the results of these experiments that diagnosis of cellulitis using current telemedicine techniques presents a challenge. Throughout the experiments, low levels of concordance, consistency and accuracy were achieved. The question which had to be addressed was whether these results were caused by human inaccuracy when performing the experiments or by the design of the experiments themselves. Several points became clear during analysis of the results. Firstly, there was little or no difference between the respective performances of doctor and non-doctor participants in any of the experiments. This confirmed the null hypothesis number 1; that no difference between them was to be expected.

Secondly, it was clearly the case that evaluation of the images by the participants was affected measurably by the introduction of numeric scales as a term of reference. This finding disagreed with the null hypothesis number 2; that no difference was expected. What became apparent, however, was that the introduction of a scale brought problems of a different type into the equation. It was clear that, on a number of occasions, the instructions for the use of the numeric scale were misunderstood, misapplied or forgotten. There were also reasons to believe that a scale utilising pictorial comparisons would prove to be of more value than a simple numeric scale. A further indication was that the numeric scales employed were often too large in range to be useful for the purpose of accurate evaluation, and that a shorter scale with wider parameters may have been more effective.

With regard to hypothesis number 3 (that no relationship would be established between confidence and accuracy), the findings of these experiments confirmed that no such relationship existed. Additionally, although no time constraints were placed on the participants for the completion of the experiments, the time taken appeared to have no influence on the level of accuracy. Null hypothesis number 4 proposed that there would be no significant difference between doctors and non-doctors in relation to their confidence levels or in the time spent in the performance of the experiments. Once again, the results tended to confirm the null hypothesis.

As was discussed in the conclusion to the results of experiments 5A and 5B, there were indications that both overconfidence (Kahneman 2011) and lack of attention (Reason 2006) may have contributed to the inconsistency of the results.

It was also considered possible that the design of the experiment was flawed in that some of the images provided for the experiments may have been insufficiently different for the human eye to distinguish between them. This applied particularly to experiments 5A and 5B, which otherwise provided many useful results. It may equally have been the case that the limitations of human perception were inadvertently established by the experiment. Furthermore, in experiment 5, insufficient images were provided to represent a fair distribution of the image degradation levels. For this reason, experiment 5 was expanded and is re-examined in chapter 6. A detailed comparison between time taken and accuracy is also there included.

Chapter 6

Image Matching Experiment

6.1 Introduction

As was explained in the results section of chapter 5, the unexpectedly large number of errors encountered during experiments 5A and 5B gave cause for concern and required these experiments to be revisited.

Applying agile methodology principles, it was therefore decided to perform amended x of the image matching experiment with different participants, a larger range of images and with a numeric scale of reference for image evaluation.

As a further refinement, it was decided to use this experiment as a pilot for the use of online testing methods, rather than paper methods. Online testing had initially been intended for use in the conjunctivitis experiments described in chapter 7.

6.2 Objectives

The following are the main objectives of this experiment:

- To determine the ability of the human eye to differentiate between two similar images.
- To compare the error rate with that experienced in experiments 5A and 5B (chapter five).
- To compare the relative accuracy of doctor and non-doctor participants.

- To establish a threshold to the capability of the human eye to distinguish between the image quality of two images. A 25% difference threshold was indicated during experiments 5A and 5B.
- To examine the effectiveness of an online testing system.

6.3 Participants

A total of 148 participants, with ages ranging from 20 to 60 years, were engaged in this experiment. The breakdown was as follows:

- 64 male doctors
- 11 female doctors
- 55 male non-doctors
- 18 female non-doctors

This equates to a reasonably representative sample in terms of age, qualifications and gender.

6.4 Design and Variables

Due to the differing locations of the participants, the experiment took place in two different environments, namely the IT facility at RGU and the A&E in ARI. As with the experiments conducted in chapter 5, great care was taken to ensure that the rooms, the IT equipment and the two environments were as near to identical as practicable. The requirements for identical items were not as great as in the preceding experiment as no specific rating or ranking was to be applied to the images, this experiment requiring only the comparison of an original image with an image-processed version of the same image. All questions, images and instruction sheets, together with the colour blindness test and explanations of the ethical issues involved, were supplied online to the participants. The experimental results were also gathered and collated online. A number of images of cellulitis pathologies were obtained from the E&A department of ARI. For control purposes, five of the images were duplicated without modification. The remaining images were electronically modified in Adobe Photoshop image processing software in order to provide an original and a modified image for comparison in the experiment. Thus were created a total of 25 pairs of images with varying degrees of colour intensity.

Table 6.1 lists the degree of modification applied to the images and example pairs are shown in figures 6.1 and 6.3. All images used in the experiment may be accessed via the Dropbox link at http://bit.ly/1l5WyTy.

Table 6.1: Image pair descriptions

Experimental questions	Degree of difference	Degree of degradation
Q1	Slight difference	25%
Q2	Moderate/great difference	35%
Q3	Slight difference	25%
Q4	Slight difference	25%
Q5	No difference	0%
Q6	No difference	0%
Q7	Moderate/great difference	35%
Q8	Slight difference	25%
Q9	Moderate/great difference	35%
Q10	Slight difference	25%
Q11	Moderate/great difference	35%
Q12	Moderate/great difference	35%
Q13	Moderate/great difference	35%
Q14	Moderate/great difference	35%
Q15	No difference	0%
Q16	No difference	0%
Q17	Slight difference	25%
Q18	Moderate/great difference	35%
Q19	Slight difference	25%
Q20	Moderate/great difference	35%
Q21	Moderate/great difference	35%
Q22	Slight difference	25%
Q23	Slight difference	25%
Q24	Moderate/great difference	35%
Q25	No difference	0%

Following are the examples of image pairs used in the experiment:

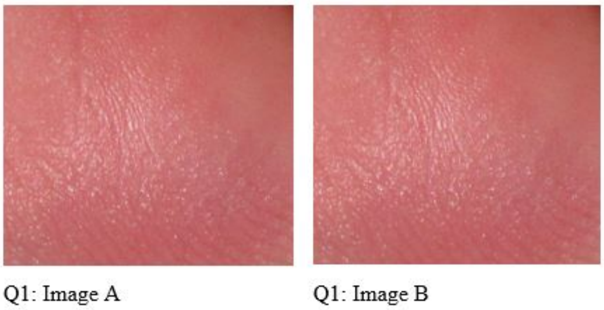


Figure 6.1: Q1 in image matching experiment

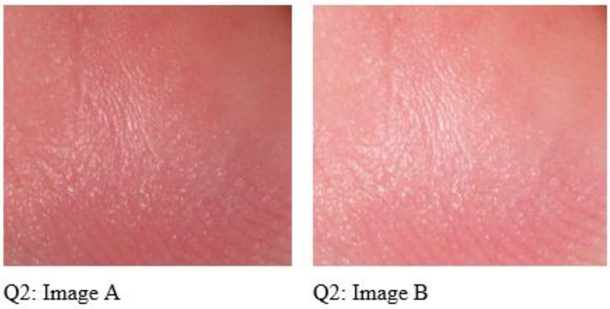


Figure 6.2: Q2 in image matching experiment

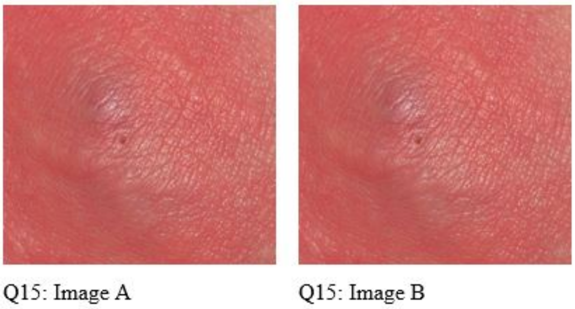


Figure 6.3: Q3 in image matching experiment

6.5 Procedure and Tasks

Before commencement of the experiments, the participants were asked to complete several administrative preliminaries including signing a consent form and undergoing a colour-blindness test. They were then required to complete an online form in order to obtain personal information about their age, sex, education level, occupation and eyesight (normal or corrected to normal with prescription lenses). Two additional questions were asked of the medically qualified participants. These concerned their areas of specialization and their years of experience with cellulitis and other skin infections.

Twenty-five pairs of images were presented to the participants who were asked to determine whether the pairs were the same or different, and to specify their confidence in each answer from 1 (minimum) to 9 (maximum). The participants were asked to complete the experiment as quickly as possible without compromising accuracy.

Independent variables in this experiment were the tasks of the experiment, the colour characteristics of the images and the image quality. The only dependent variables were the answers of the participants. Examples of the data collected from all users can be viewed in appendix 10.

6.6 Results

The following are the results of the experiment. IBM SPSS software version 17.0 was used for analyses. 147 participants took part, 72 of whom were non-doctors and 75 doctors.

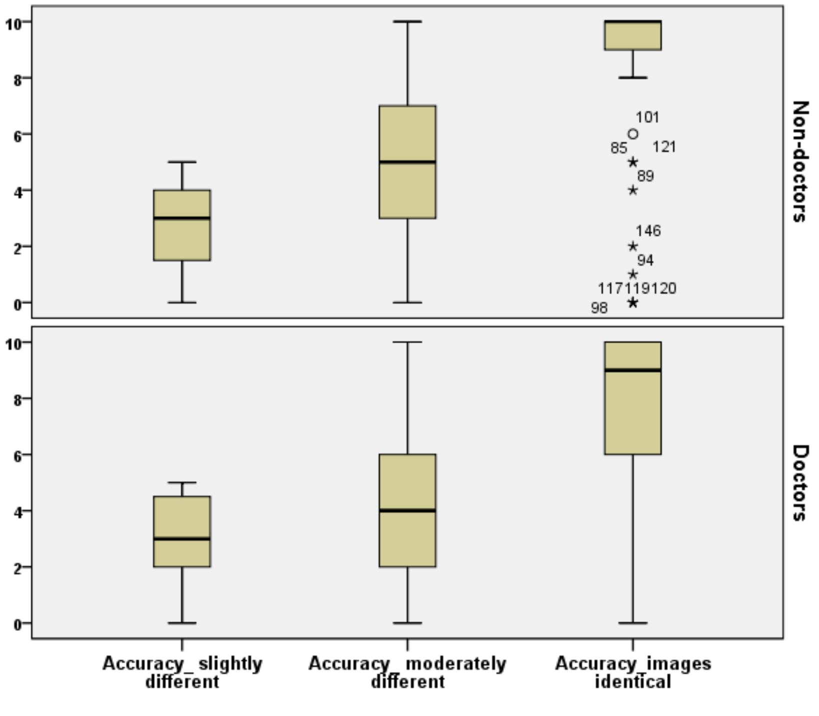


Figure 6.4: Relative accuracy of doctors and non-doctors at different levels of image modification

The boxplot in figure 6.4 shows the consistency and accuracy when doctors and non-doctors matched images with different levels of modification. The median of the total accuracy for non-doctors (mdn = 18) was higher than the median of the total accuracy for doctors (mdn =16). Additionally, the medians of the accuracy, when the images were slightly different, moderately different, and identical (5, 10, and 3) for non-doctors were higher than for doctors, with (4, 9, and 3) respectively. Mann-Whitney U tests were performed to establish the respective accuracies of doctors and non-doctors. These demonstrated that significant differences occurred between the two groups when the images differed from each other (P = 0.001 and U = 1877.000, mean rank of 85.43 for non-doctors and 63.03 mean rank for doctors). In the case of identical images, however, the performance of the two groups showed no difference (see results summarised in table 6-2).

Table 6.2: Mann-Whitney U test summary

	Participants	N	Mean Rank	P value	U value
Accuracy when images identical	Non-doctors	72	72.60	0.692	2599.500
	Doctors	75	75.34	0.692	2599.500
Accuracy when sight difference	Non-doctors	72	82.19	0.021	2110.000
	Doctors	75	66.13	0.021	2110.000
Accuracy when moderate/high difference	Non-doctors	72	82.44	0.008	2092.500
	Doctors	75	65.90	0.008	2092.500

Statistical analysis demonstrated that the only real concurrence occurred with the unmodified images. Both slight and moderate/high difference categories showed clear statistical significance in the difference between the answers of the two groups of participants, with non-doctors being more accurate than doctors. A Wilcoxon signed ranks test was also used to examine the difference in accuracy between the three categories and showed a major shift in ranking in the comparison of the three levels. Table 6-3 clearly shows that the participants had little difficulty in recognising identical images, but their ability to differentiate between the images was at its best when there was a clear distinction between the two, and progressively diminished the closer the two images became in degree of difference.

Table 6.3: Wilcoxon signed ranks test summary for all participants

For all participants		N	Mean Rank	P value	Z value
- Accuracy when slight difference	Negative Ranks	37a	61.59	0.001	-4.805b
- Accuracy when no difference	Positive Ranks	95b	68.41	0.001	-4.805b
	Ties	15c		0.001	-4.805b
	Total	147		0.001	-4.805b
- Accuracy when moderate	Negative Ranks	19d	25.74	0.001	-9.457b
/ high difference	Positive Ranks	125e	79.61	0.001	-9.457b
- Accuracy when no difference	Ties	3f		0.001	-9.457b
	Total	147		0.001	-9.457b
- Accuracy when moderate	Negative Ranks	6g	24.58	0.001	-9.573b
/ high difference	Positive Ranks	124h	67.48	0.001	-9.573b
- Accuracy when slight difference	Ties	17i		0.001	-9.573b
	Total	147		0.001	-9.573b

Key

(a) Accuracy when slight difference < Accuracy when no difference

(b) Accuracy when slight difference > Accuracy when no difference

(c) Accuracy when slight difference = Accuracy when no difference

(d) Accuracy when slight difference < Accuracy when no difference

(e) Accuracy when moderate / high difference > Accuracy when no difference

(f) Accuracy when moderate / high difference = Accuracy when no difference

(g) Accuracy when moderate / high difference < Accuracy when slight difference

(h) Accuracy when moderate / high difference > Accuracy when slight difference

(i) Accuracy when moderate / high difference = Accuracy when slight difference

A Wilcoxon test for doctors and non-doctors was also calculated and showed similar results. The data may be accessed via the Dropbox link at `http://bit.ly/1l5WyTy`.
A Friedman's test confirmed these findings, showing mean ranks of:
2.63 for moderate/high difference.
1.53 for slight difference.
1.84 for identical.
P = 0.001

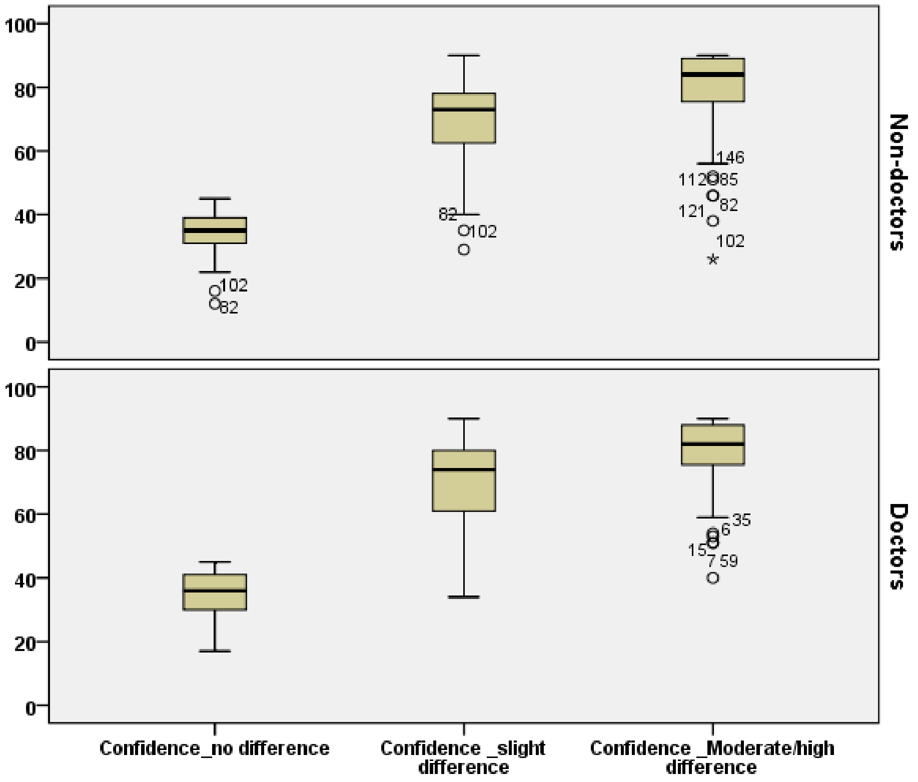


Figure 6.5: The confidence levels of doctors and non-doctors closely mirrored each other

6.7 Confidence and Accuracy

The boxplot shown in figure 6.5 provides a clear demonstration of how the confidence levels of doctors and non-doctors closely mirrored each other but bore no relationship to the accuracy of the results. As was shown earlier, the most accurate results in these experiments occurred when the images being compared were identical, yet this was also when confidence was at its lowest for both groups.

Hypothesis numbers 3 and 4 proposed that there would be no relationship between time taken, confidence and accuracy, and also that there would be no difference between doctors and non-doctors in terms of their confidence or the time taken to answer the questions.

As part of the experiment, participants were asked to state their levels of confidence in their answers. This data was used in Mann-Whitney U tests to compare the confidence levels of doctors and non-doctors. Mann-Whitney allocated mean ranks of 74.16 to the 73 non-doctors and 74.83 to the 75 doctors (U = 2713, Z = -0.094, p = 0.925). No statistically significant difference was exhibited between the confidence levels of the two groups.

A Wilcoxon signed ranks test was used to compare the effects of the three image difference classifications. The findings were as follows:

Comparing slight to identical, neg. ranks = 0, pos. ranks = 147, Z = -10.521, p = 0.001.

Comparing moderate/high to identical, neg. ranks = 0, pos. ranks = 147, Z = -10.522, p = 0.001.

Comparing moderate/high to slight, neg. ranks = 17, pos. ranks = 118, Z = -9.029, p = 0.001.

This analysis confirmed what was apparent in the boxplot; that there were clear distinctions in confidence, based on the objective difficulty of the task.

The full tables of statistics for all participants, doctors only and non-doctors only can be found via the Dropbox link at http://bit.ly/1l5WyTy.

6.8 Time Results and Analysis

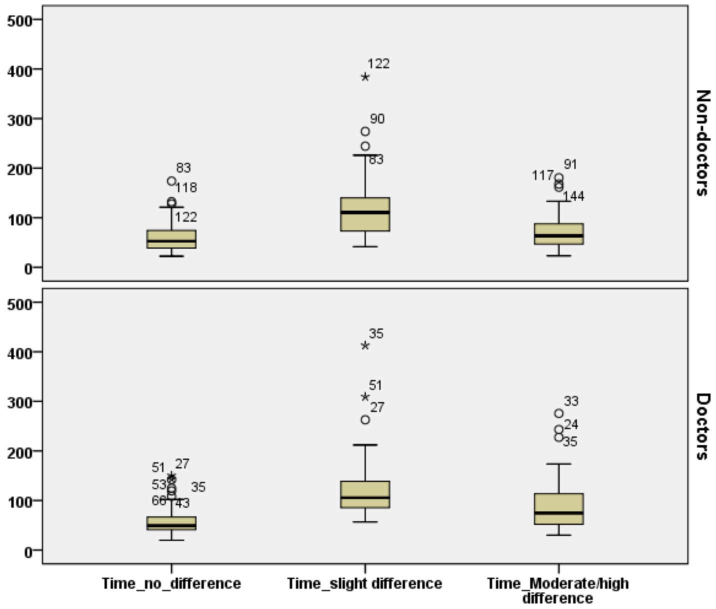


Figure 6.6: Time taken with different redness degrees

Times taken by all of the participants were recorded in order to determine if there was any significant relation between time and accuracy. It should be noted that there were 6 fewer participants in this analysis than in the confidence experiment. These individuals chose to take a protracted break part way through the experiment, causing their results to be eliminated from the analysis. They can, however, still be observed as outliers in the boxplot representations. Comparison between doctors and non-doctors for the time taken to complete the experiment was carried out using Mann-Whitney U tests. These revealed that there was no statistically significant difference between the two groups. The mean rank for non-doctors was 67.99 and for doctors was 75.01. U = 2271, Z = -1.018, p = 0.309.

These findings corresponded closely with those of the confidence level experiments where, again, the two groups of participants were closely matched.

A Wilcoxon signed ranks test was used to compare the effects of different degrees of image quality on the time taken by the participants. The findings were as follows:

Comparison of slight to identical, neg. ranks = 4, pos. ranks = 143, Z = -10.398, p = 0.001. Comparison of moderate/high to identical, neg. ranks = 43, pos. ranks = 104, Z = -5.778, p = 0.001. Comparison of moderate/high to slight, neg. ranks = 127, pos. ranks = 20, Z = -8.774, p = 0.001.

Statistical tables for all participants, non-doctors and doctors may be accessed via the Dropbox link at http://bit.ly/1l5WyTy.

6.9 Relationship among accuracy, confidence and time

Spearman rank-order correlation tests were conducted in order to determine if there were any relationships among the accuracy, confidence, and time taken when performing the image matching experiments. In all cases (doctors, non-doctors and the two combined) there was low correlation indicated:

In the case of all participants combined r = 0.140 and P = 0.091 for accuracy and confidence, and r = -.364 and P = 0.001 for accuracy and time. In the case of non-doctor participants r = .298 and P = 0.011for accuracy and confidence. and r = -0.382 and P = 0.001 for accuracy and time. In the case of doctors r = -0.012 and P = 0.919 for accuracy and confidence, and r = -0.339 and P = 0.003 for accuracy and time.

This confirmed the findings of experiments 5A and 5B (chapter 5), and the null hypotheses of the thesis. There is no indicated relationship between accuracy and either time taken or confidence (hypothesis 3). There is no difference between doctors and non-doctors in either their confidence levels or their time taken for the completion of the experimental tasks (hypothesis 4). While the correlations were low, some of the p-values ¡ 0.05 indicated significant association, e.g. it appeared that non-doctors with high scores tended to answer quickly.

6.10 Critical Discussion of Results

In all of the experiments relating to human assessment of digital images of cellulitis, it could be tentatively suggested that the human visual system was able to recognise the differences between two images when this difference was great or moderate, but found it more difficult when the colour characteristics were similar. Typically, a degree of difference of 25% appeared to create a level at which problems were experienced.

This emphasised the need for high quality capture and transmission of images to be used in telemedicine, if inaccuracies due to poor image quality are to be avoided.

In the results of the image matching experiments, it was notable that non-doctor participants generally were slightly more accurate than the doctors. These findings would tend to be counter intuitive, in that doctors, given their experience, would be expected to be more accurate in their assessments (Tversky and Kahneman 1974). There are a number of considerations, however, which could account for this discrepancy. In the majority of instances, the experiences of doctors would be in face-to-face consultations, where other factors such as the overall health of the patient and the context of the infection, would aid accurate diagnosis. There also exists the possibility that the doctors may have been overconfident in their own ability and therefore were not as careful in their judgements (Kahneman 2011) as the non-doctors, for whom this type of work would be novel and require more attention. One further consideration was that the majority of non-doctor participants worked within the IT field and may, therefore, have been more familiar than the doctors with digital images.

It was clear that the inclusion of a scale caused evaluations to be changed, but problems experienced in the application of the scales highlighted a number of issues. Without clear instructions, it was possible for the scales to be misinterpreted and applied in the opposite direction to that intended. It was also clear that the use of a wide scale with narrow parameters was ineffective as the small differences in each step were too subtle to be easily discerned.

Neither the time taken, nor the participants' level of confidence showed any statistically significant correlation with the degree of accuracy of the answers to the image matching questions. It was also noted that confidence tended to be applied in a haphazard manner and was not reflected in the degree of difficulty of individual questions. In other words, it was possible for participants in either category to be equally confident about a wrong answer as about a correct one.

Several of the hypotheses of the work were examined in the course of the preceding experiments, the findings were as listed below:

Hypothesis 1 proposed that there would be no difference in the accuracy of doctors and non-doctors. This was not substantiated, non-doctors proving slightly more accurate than doctors.

Hypothesis 2 suggested that the use of scales would not affect the accuracy of either group. Scales had a definite effect on judgement, and further work was carried out on this hypothesis in chapter 7.

Hypothesis 3 proposed that there would be no relationship between accuracy, confidence and time taken. This was confirmed.

Hypothesis 4 proposed that there would be no significant difference between the two groups of participants in relation to either their confidence level or the time they would take to complete the experiments.

Hypothesis 5 remains unconfirmed, as further work is carried out in chapter 7.

Chapter 7

Conjunctivitis Experiment

7.1 Introduction

This chapter presents a series of, initially, five sets of experiments. As a later refinement, a further two experiment sets were added. All of these experiments involved the investigation of human perception of colour, in this case using conjunctivitis as the basis of stimuli in the experiments. The design of the experiments took into account lessons learned during previous experiments.

7.2 Objectives of Conjunctivitis Experiments

The list below details the objectives of these experiments:

1. To evaluate the consistency and accuracy of the participants in assessing degrees of redness within the presented stimuli.
2. To investigate the effectiveness of using a red colour scale relating to infective severity level, then using this to classify redness for diagnostic purposes.
3. To compare the accuracy in classifying degrees of redness of two different scales, one having five divisions and one having three.
4. To test the relationship among accuracy, confidence and time when classifying degrees of severity of conjunctivitis.
5. To compare the differences in consistency, accuracy, confidence and time between doctors and non-doctors in their performance of these experiments.

7.3 Participants

In experiments 1 to 5, 70 doctors and 70 non-doctors took part, 20 of the doctors and 27 of the non-doctors being female. In experiments 6 and 7, a total of 30 doctors, 10 of whom were female, and 30 non-doctors, including 13 females, took part. 85% of all participants had ages ranging from 20 to 40 years. All participants had either normal or corrected eyesight.

7.4 Design and Variables

As was the case in previous experiments, it was again necessary to carry out the experiments in two different locations. As previously, considerable care was taken to ensure that the two environments were as near to identical as possible. Unlike the previous experiments, one laptop computer was used for all participants, with the experimenter present in order to facilitate use of the online documentation (items such as colour-blindness test plates, ethical issues and consent forms, question and answer sheets and such like).

The independent variables in these experiments comprised the tasks, the use of colour scales, image quality scales and confidence scales. Owing to the standardisation measures taken, the environment and the IT equipment were not, on this occasion, considered to be variables. The dependent variables were the participants' answers to the experiments' questions.

7.5 Materials Used in Conjunctivitis Experiments

All materials used were standardised for all participants. The following are the materials and tools that were used online in all experiments.

7.5.1 Online question and answers sheets

All the experiments were uploaded online to the research webpage of the RGU website `http://www.comp.rgu.ac.uk/staff/iba/index.htm`. The participants were given the following links to use when they performed the experiments:

- Experiment 1: http://www.comp.rgu.ac.uk/staff/iba/start1intro.htm
- Experiment 2: http://www.comp.rgu.ac.uk/staff/iba/start2intro.htm

- Experiment 3: http://www.comp.rgu.ac.uk/staff/iba/start3intro.htm
- Experiment 4: http://www.comp.rgu.ac.uk/staff/iba/start4intro.htm
- Experiment 5: http://www.comp.rgu.ac.uk/staff/iba/start5intro.htm

Participants were given a password and, in the presence of the researcher, they accessed the actual experiments from the above links and viewed the images and questions. Examples of the images and online experiment questions can be viewed in appendix 13 and the complete experiment questions and images may be viewed via the Dropbox link at http://bit.ly/1l5WyTy.

Other materials which were provided online were as follows:

- Online instruction sheets (appendix 12).
- Online colour-blindness test (appendix 8)
- Online user consent form (appendix 7.1)
- Conjunctivitis grading guide (figure 7.1)
- Image quality scale (figure 7.2)

7.5.2 Stimuli (conjunctivitis pathology images)

A large number of images of red eye pathology were presented to three practicing eye specialist doctors. They independently rated the quality of the images and the level of redness displayed. When the results were compared, only the images on which they were in complete agreement were selected as suitable stimuli for these experiments. In total 410 images were used, all of which can be may be viewed in the extra appendix folder via the Dropbox link at http://bit.ly/1l5WyTy. Examples of the images are presented in figure 7.3. Examples of all the image categories may be viewed in appendix 11

7.6 Tasks of the experiments

The seven experiments each consisted of a number of conditions. For ease of reference, each condition was allocated a letter as a suffix to the experiment number, thus the experiments appear as 1A, 1B, 1C, etc. All the experiments and their questions may be viewed in appendix 13.

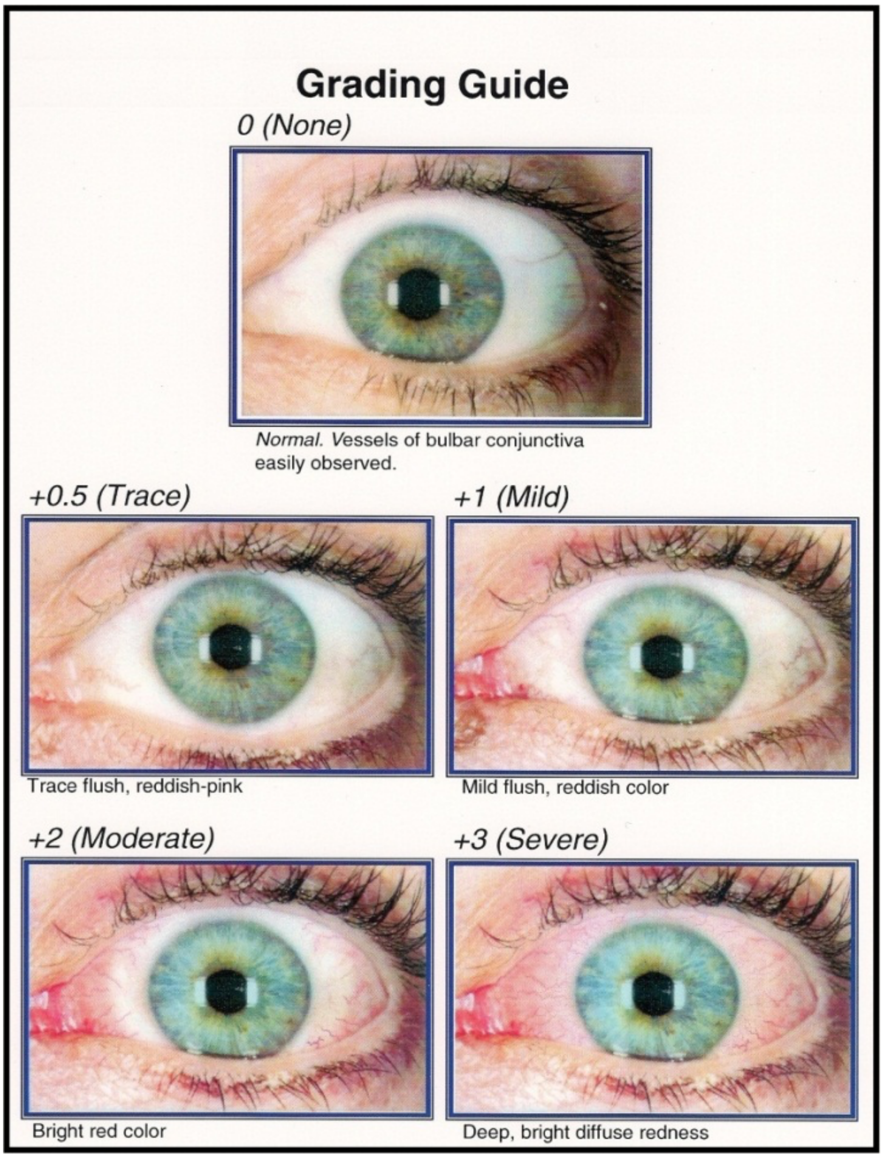

Grading Guide
0 (None)
Normal. Vessels of bulbar conjunctiva easily observed.
+0.5 (Trace)
Trace flush, reddish-pink
+1 (Mild)
Mild flush, reddish color
+2 (Moderate)
Bright red color
+3 (Severe)
Deep, bright diffuse redness

Figure 7.1: Conjunctivitis grading Grad designed by Allergan and used by the eye unit at ARI

7.6.1 Experiment 1A

Experiment 1A consisted of 10 questions, each of which presented an image of a conjunctivitis case. The participants were asked to grade the degree of redness using the grading guide shown in figure 7.1.

Experiment 1B presented the participants with 10 pairs of images of a conjunctivitis case and asked them to assess whether the first image displayed more, the same or less redness than the second image. Experiment 1C comprised 3 questions in which the

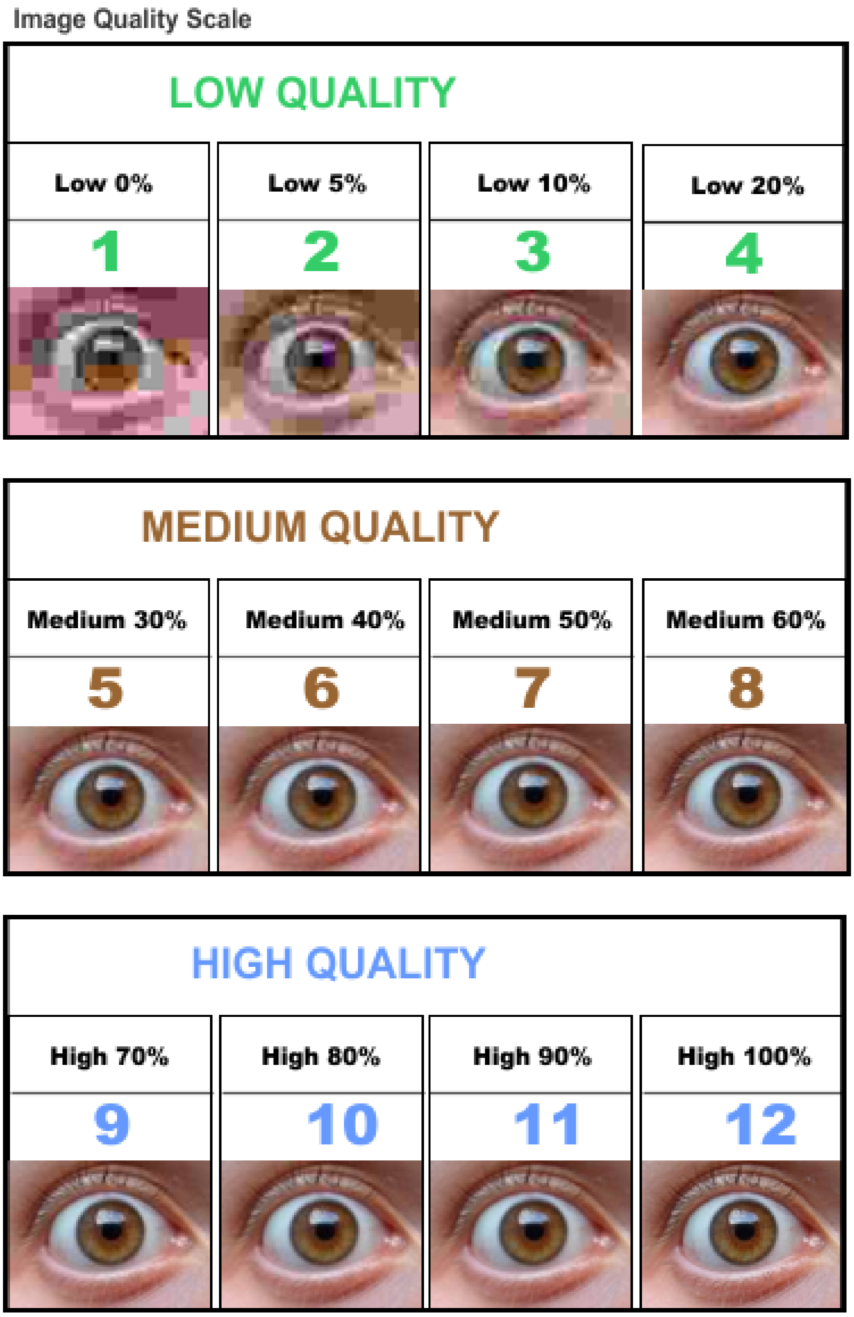


Figure 7.2: Image quality scale designed by the researcher

participants were presented with 3 images of a conjunctivitis case which they were then asked to grade with the aid of the grading guide shown in figure 7.1. In this experiment, all 3 images of pathology were shown on screen at the same time. Experiment 1D was a repeat of experiment 1A with the same images but without the assistance of the grading guide. Experiment 1E replicated experiment 1C, using the same images but without the assistance of the grading guide. Two sets of 5 images of conjunctivitis cases were

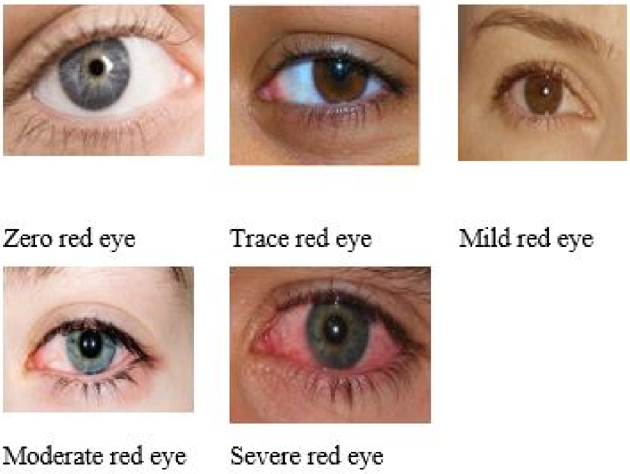


Figure 7.3: Degrees of redness in conjunctivitis

presented on screen simultaneously and the participants were asked, with the assistance of the grading guide, to classify them by degree of severity. Experiment 2B required the participants to grade 12 different images presented on screen simultaneously. The grading guide was included. Experiment 2C was a repeat of experiment 2A, using the same images but without the use of the grading guide. As experiment 2B, using the same images but without the grading guide.

In experiment 3A, 12 modified images of a single image of a conjunctivitis case were presented individually to the participants. With the first 6 images, participants were supplied with a quality scale (figure 7.2) and asked to grade the experimental images by quality of resolution, expressed as a percentage (as demonstrated in the scale). The participants were required to grade the succeeding 6 images in the same manner but without using the grading scale.

In experiment 3B, 10 pairs of conjunctivitis images were presented and the participants were asked to judge which image of each pair had higher quality, or if they were the same. In this experiment, an image quality scale was not provided.

In experiment 3C, Six groups of 3 images were presented on screen and the participants were asked to rate the quality of the images as ”high”, ”medium” or ”low”. The first three image groups were judged with the aid of a quality scale and the second three groups without.

Experiment 3D was a repeat of experiment 3C, except that a different set of images were presented in groups of 5. Experiment 3E consisted of a single question. Twelve images were simultaneously displayed on screen and the participants asked, without the aid of a quality scale, to rate their quality. A further twelve images were presented in experiment 3F to the participants, again without a quality scale. On this occasion, they were asked to determine which of the images were sufficiently well defined to permit classification of the degree of redness. Experiments 4 and 5 were repeats of experiments 1 and 2 utilising the same images and the same participants. In these experiments, however, a scale of only three divisions was employed, as opposed to the five divisions used in the original experiments. The intention was to compare the effect that the scale granularity had on the accuracy of the answers.

In experiments 6 and 7, A group of 60 additional participants offered to take part in experiments 1 and 2 only (their demographic is outlined in the 'participants' section). In order to retain the same sample of participants throughout all experiments 1 to 5, it was decided to rerun experiments 1 and 2 with the new participants and use the results as a direct comparison between them and the original sample. Both groups were large enough to constitute an acceptable experimental sample.

7.7 Confidence

Throughout all the experiments described in this chapter, the participants were asked to state their level of confidence in each answer on a scale of 0 (no confidence) to 9 (total confidence).

7.8 Results

7.8.1 Results of Experiment 1D

Experiments 1A and 1D were essentially the same experiment except that a reference colour scale was provided in 1A but not in 1D. As a safeguard against the participants using, in 1D, their memory of the images that they had already viewed in 1A, these two experiments were separated by two others (1B and 1C). This then permitted direct comparison, within a common statistical framework, of the two groups of participants in the two experiments.

In order to test the hypotheses 1 and 2 relating to the respective performances of the two participating groups, and the effects, if any, of including a scale of reference, Mann-Whitney U tests were performed to compare the accuracy of the two groups in experiment 1A (with scale) and in experiment 1D (without scale). The results of the Mann-Whitney U tests showed there to be no significant difference in the accuracy of the participating groups (P = 0.813 for 1A when using a colour scale and P =

0.518 for 1D when not using a colour scale). In both participating groups, there were 70 individuals. The mean rankings were 69.7 for non-doctors and 71.3 for doctors in experiment 1A (with scale). In experiment 1D (without scale) the mean rankings were 72.7 for non-doctors and 68.3 for doctors. The Mann-Whitney U values were 2394 (1A) and 2298 (1D).

These findings confirmed null hypothesis 1, that there would be no difference in the accuracy of doctors and non-doctors. In order to visualise the accuracy and consistency of both groups, boxplots were generated in IBM SPSS for both experiments.

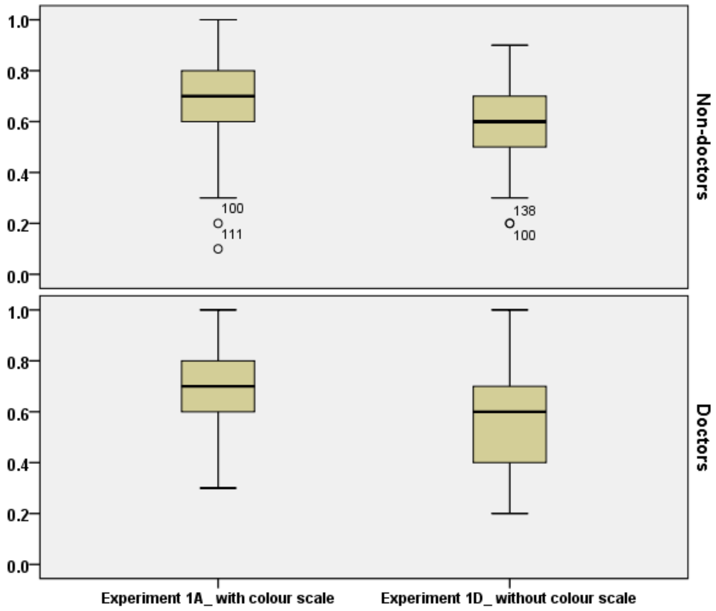


Figure 7.4: The effects on accuracy of using a colour scale

The boxplots (figure 7-4) clearly illustrate the effect of the inclusion of a scale, the levels of accuracy being noticeably higher, for both groups, where the scale was included. It is also evident that doctors, in particular, became less consistent in their answers after the scale was withdrawn. This despite the fact that the same images were used for both experiments.

It was clear that null hypothesis 2, that the inclusion of a scale for reference purposes would have no effect on the results, was refuted by these findings. In order to determine the statistical significance of the effects of using or not using a standard scale, a Wilcoxon signed ranks test was performed for both doctor and non-doctor groups. For both groups of participants, the use of the standard scale resulted in a higher degree of accuracy. 37 of 70 non-doctors and 43 of 70 doctors showed improved accuracy with the use of the scale. A significant degree of difference was demonstrated by the Wilcoxon test; Z=-2.693 and p=0.007, for non-doctors and Z=-3.453 and p=0.001 for doctors.

7.8.2 Results of Experiment 1B

Mean accuracy calculations demonstrated that the accuracy of both groups taking part in the experiment was in excess of 70%. Comparison by Mann-Whitney (tables 7.1 and 7.2) showed that the difference between the two groups' accuracy was significant.

Table 7.1: Mann-Whitney

	Participants	N	Mean Rank	Sum of Ranks
Experiment IB without colour scale	Non- doctors	70	63.76	4463.00
	Doctors	70	77.24	5407.00
	Total	140	3	4

Table 7.2: Mann-Whitney

	Experiment IB without colour scale
Mann-Whitney U	1978.000
Wilcoxon W	4463.000
Z	-2.024
Asymp. Sig. (2-tailed)	0.043
a. Grouping Variable: Dr. Non	

Analysis of the differences between the evaluations of the images by the two groups (table 7.3) indicated that the major area of disagreement was in images 7 and 8. Examination of these particular images revealed that the colour reproduction of the surrounding skin areas presented an overall greater impression of redness in one image compared to the other. It appears a reasonable assumption that the non-doctor participants may have been misled by this contextual colour and evaluated the images accordingly. Doctors, on the other hand, considered only the area of interest and therefore evaluated the degree of redness only, leading to a more accurate result.

Table 7.3: Comparison of the accuracy of doctors and non-doctors in each image

Questions	Participants	N	Mean Rank	Sum of Ranks
Q1_LESS THAN	Non-doctors	70	69.00	4830.00
	Doctors	70	72.00	5040.00
Q2_LESS THAN	Non-doctors	70	71.50	5005.00
	Doctors	70	69.50	4865.00
Q3_MORE THAN	Non-doctors	70	69.00	4830.00
	Doctors	70	72.00	5040.00
Q4_THE SAME	Non-doctors	70	71.50	5005.00
	Doctors	70	69.50	4865.00
Q5_LESS THAN	Non-doctors	70	69.50	4865.00
	Doctors	70	71.50	5005.00
Q6_THE SAME	Non-doctors	70	68.50	4795.00
	Doctors	70	72.50	5075.00
Q7_MORE THAN	Non-doctors	70	63.50	4445.00
	Doctors	70	77.50	5425.00
Q8_LESS THAN	Non-doctors	70	62.50	4375.00
	Doctors	70	78.50	5495.00
Q9_LESS THAN	Non-doctors	70	73.00	5110.00
	Doctors	70	68.00	4760.00
Q10_MORE THAN	Non-doctors	70	69.00	4830.00
	Doctors	70	72.00	5040.00
Total Participants		140		

Given these observations, it seems that contextual colour may be as important a consideration in assessing conjunctivitis images as it appears to be in evaluation of cases of cellulitis. This makes an even stronger case for the importance of using a standardised methodology in the capture of images to be used in telehealth diagnosis.

7.8.3 Results of Experiments 1C (with colour scale) and 1E (without using colour scale)

As can be seen in the SPSS boxplot (figure 7-5), the levels of accuracy of both doctors and non-doctors were again similar, although the overall level of accuracy was poor. Table 7-4 presents a Mann-Whitney U test demonstrating no significant difference between the two groups. In this case, however, the effects of the use of the colour scale were markedly different, as can be seen in the Wilcoxon signed ranks tests shown in tables 7-5 to 7-6.

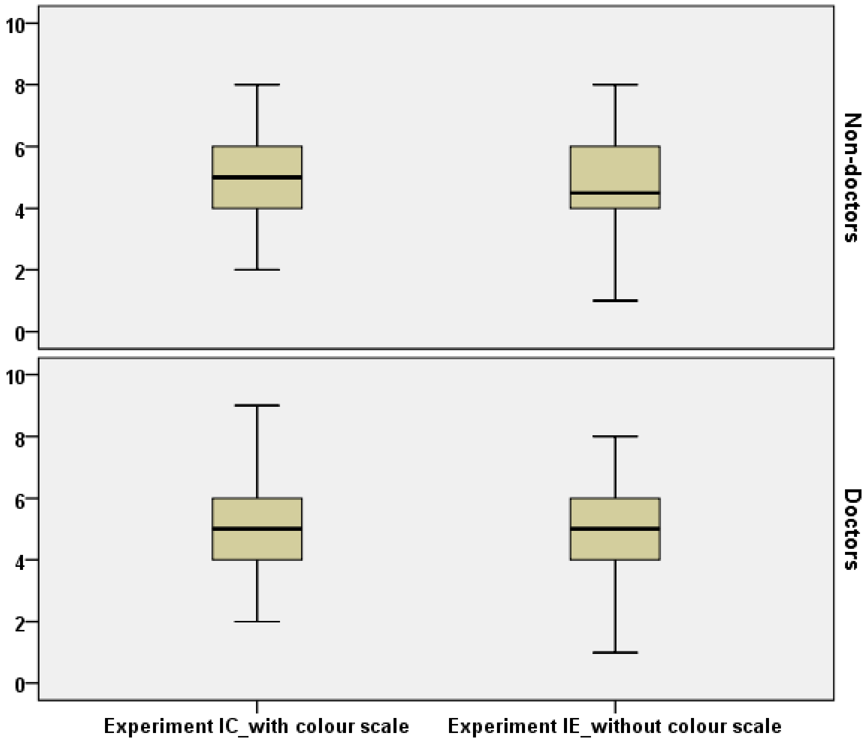


Figure 7.5: The effects on accuracy of using a colour scale

Table 7.4: Mann-Whitney test summary

Experiment	N	Mean Rank Mean Rank	Mann-Whitney U value	P value Asymp. Sig.
Experiment 1C with scale	70 Non-doctors 70 Doctors	71.17 69.83	2403.000 2403.000	0.841 0.841
Experiment 1E without scale	70 Non-doctors 70 Doctors	64.68 76.32	2042.500 2042.500	0.084 0.084

Table 7.5: Wilcoxon signed ranks tests summary for doctors

Experiment	N		Mean Rank	Z value	P value 1E-1C
Experiment 1C with scale & Experiment IE without scale	Negative Ranks	26	25.63	-.292 b a. Wilcoxon Signed Ranks Test	0.771 b. Based on positive ranks.
	Positive Ranks	24	25.35		
	Ties	20c			
	Total	70			
a. 1E <1C b. 1E >1C c. 1E = 1C					

As in previous experiments, the use of the colour scale resulted in significant improvements to the results of the non-doctor participants. In this case, however, there was no significant difference between the two sets of results for doctors. This may be a situation where the experience of doctors has come into its own. As this experiment

Table 7.6: Wilcoxon signed ranks tests summary for non-doctors

Experiment	N		Mean Rank	Z value	P value 1E-1C
Experiment 1C with scale & Experiment IE without scale	Negative Ranks	32a	27.53	-2.101b a. Wilcoxon Signed Ranks Test	0.036 b. Based on positive ranks.
	Positive Ranks	19b	23.42		
	Ties	19c			
	Total	70			
a. 1E <1C b. 1E >1C c. 1E = 1C					

consisted of the presentation and comparison of three simultaneous images, it is possible that, by application of their knowledge, the doctors were able to use the three images, as a scale in themselves, to work independently of the supplied scale. Later experiments should confirm whether this was the case.

7.8.4 Results of experiments 2A and 2C

All of the experiments of this group were designed to test the ability of the participants to work with the simultaneous presentation of multiple images. Additionally, the experiments were performed both with and without the assistance of a scale. This provided an opportunity to re-examine whether the doctors tended to use direct comparison of the conjunctivitis images, rather than use the scale, and also if there was any indication of non-doctors adopting the same technique. Results were compared initially in the form of SPSS boxplots, with statistical analyses added subsequently.

Non-doctor participants showed little difference in terms of accuracy and consistency either with or without the aid of the grading guide (figure 7-6). As previously, their results compared favourably with those of doctor participants. A major change was noted in the case of doctors, where both their accuracy and consistency declined with the use of the grading guide (the median changed from 4.5 to 6.1 where the scale was not present). If, as was suggested in experiment 1, the doctors were using the experimental images themselves as a means of assessing the degree of redness, then it would appear that the introduction of the grading guide caused sufficient confusion as to be detrimental to the accuracy.

In order to simplify interpretation of results, it was decided to consolidate the Mann-Whitney U tests for comparison of the participating groups in experiments 2A and 2C and experiments 2B and 2D into a single table (table 7-7).

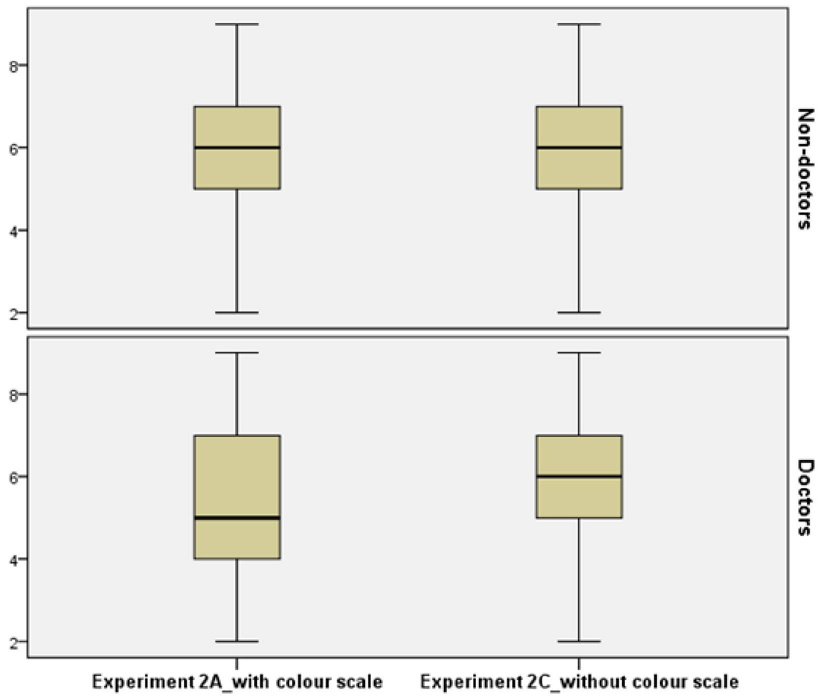


Figure 7.6: Comparison of accuracy and consistency for doctors and non-doctors

Table 7.7: Mann-Whitney test summary

	Participants	N	Mean Rank
Experiment 2A with colour scale	Non-doctors	70	75.24
	Doctors	70	65.76
Experiment 2C without colour scale	Non-doctors	70	67.40
	Doctors	70	73.60
Experiment 2B with colour scale	Non-doctors	70	75.56
	Doctors	70	65.44
Experiment 2D without colour scale	Non-doctors	70	69.79
	Doctors	70	71.21

The summary demonstrated that there was no statistically significant difference between the accuracy of the two groups. This supported previous findings that confirmed null hypothesis 1.

7.8.5 Results of experiments 2B and 2D

The SPSS boxplot results in figure 7.7 provided evidence that, as in experiments 2A and 2C, the doctor participants were slightly more accurate and consistent without the aid of the grading guide (the median changed from 5.0 to 5.8). This tended to confirm the earlier theory, that doctors were using the multiple images as an impromptu scale.

There was, however, an opposite effect with non-doctors. Their overall accuracy improved in the presence of a scale (the median changed from 6.0 to 5.0 when the scale was not present). Examination of the upper and lower quartiles indicated that the results were slightly less consistent when the scale was present but also that the extreme value outliers, that were seen when no scale was used, disappeared. Once again, these results tended to refute null hypothesis 2 in that the inclusion of a scale made a positive contribution to accuracy.

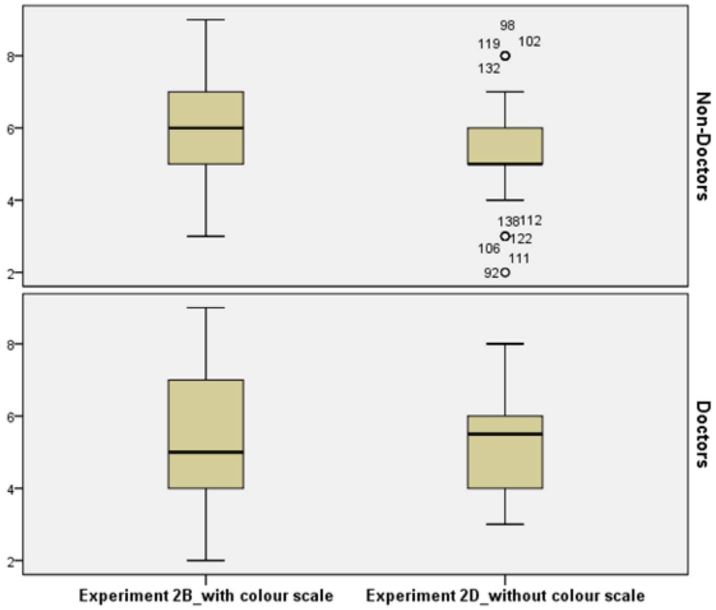


Figure 7.7: Comparison of accuracy and consistency for doctors and non-doctors

A Wilcoxon signed ranks test clearly displayed conflicting results for the use of a grading guide in the case of doctors. Comparing 2C and 2A, there was clear indication that the grading guide had a negative effect on accuracy, showing a negative rank shift of 35 and a positive rank shift of only 12. A high degree of significance was assigned to this by Wilcoxon (Z = -3.594, p = 0.001). In 2D and 2B, however, the shift was small and Wilcoxon attributed no significance to the result (Z = -0.170, p = 0.858). 70 doctors were involved in the experiment.

In the case of non-doctors, the findings were opposite to those seen with doctors in that there was a positive result from the introduction of the grading guide. Again, however, the degree of effect was inconsistent. In comparing 2C and 2A, a Wilcoxon signed ranks test on the 70 non-doctor participants indicated no significant difference between the two results (Z = -0.096, p = 0.924). In the case of 2D and 2B, a clear significance was calculated with a shift in negative ranks of 35 against a positive rank change of only 10 (Z = -3.169, p = 0.002).

Definite inconsistencies arose within these results both in the individual groups and in the comparison of the two groups. The only possible explanation for this was the small number of images used in the experiments and the consequently small sample of data available for analysis. In each analysis of 2A and 2C there were only 10 images in the data table and only 12 in the case of 2B and 2D. Due to these inconsistent results, it was decided that no reliable conclusions could be drawn in relation to hypothesis 2. In subsequent experiments, data would also, wherever possible, be combined to provide a more representative analysis.

7.8.6 Results of experiment 3A, 3C and 3D

Experiments 3A, 3C, and 3D were analysed together as they had the same questions and experimental conditions, the only variable being the number of images which were simultaneously displayed. All three experiments were divided so that half of each had a scale present and half had no scale. The results are therefore presented as 3̈ACD with scale; and 3̈ACD without scale.

At a glance, the boxplots in figure 7.8 indicate that there was little difference between doctors and non-doctors, regardless of whether a scale was used or not. These indications are more closely analysed below. As in the majority of preceding experiments, Mann-Whitney U tests comparing the accuracy of doctors and non-doctors showed no significant difference between the two groups. 70 participants took part in each group. Experiment 3ACD (with scale) resulted in Z = -1.356 and p = 0.175. Experiment 3ACD (without scale) produced results of Z = -0.391 and p = 0.695. As in previous experiments, the null hypothesis 1 was confirmed by these findings.

A Wilcoxon signed ranks test was performed for both groups in order to analyse the effects of the inclusion of a scale. In the case of the 70 doctors, a shift of 26 negative ranks was offset by a 35-point change in the positive rank. No significant difference was observed between the two conditions (Z = -1.152 and p = 0.249). For the 70 non-doctor participants, the shift in ranks was almost equal, 31 negative and 33 positive, resulting in Z = -0.964 and p = 0.335.

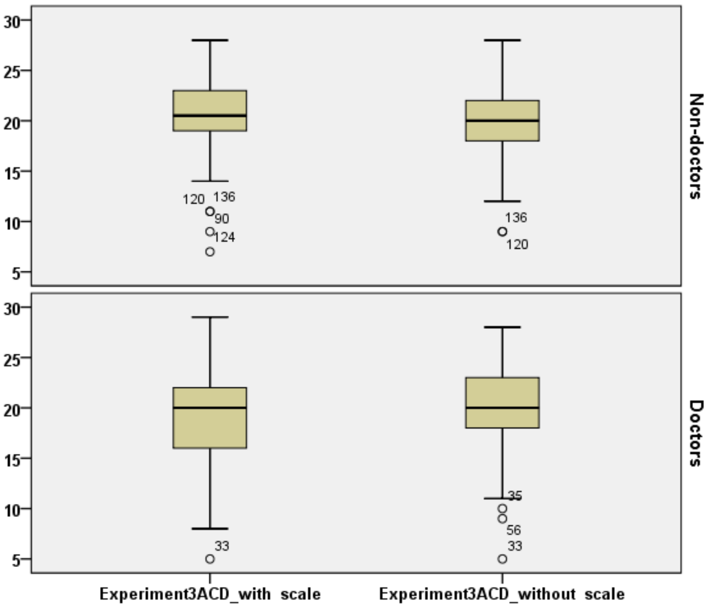


Figure 7.8: Comparison of accuracy and consistency for doctors and non-doctors

On this occasion, the results tended, for the first time, to support hypothesis 2. The reasons for this were investigated and it was concluded that the ultimate cause was due to features of the experimental design.

This group of experiments involved assessment of image quality, rather than degree of redness. The image quality scale that was used was structured as three general levels of quality (low, medium and high) but the scale then presented four slightly modified versions of the same example image within each of these three main levels, thus offering a total of twelve options by which to rate each image of the experiment.

With hindsight, this was far too extensive a choice and resulted only in confusion for the participants. Additionally, in experiments 2C and 2D, multiple versions of the same image were shown simultaneously, and these created, in themselves, a scale which was easier to apply than that provided. Thus, the presence of a scale had little effect in these experiments, and null hypothesis 2 was therefore supported by the results.

7.8.7 The results of experiment 3B

As experiment 3B involved the comparison of two images, no scale was required and thus this experiment was excluded from the preceding statistical analyses. In comparing the quality of pairs of images, doctor participants were generally more in accord than the non-doctors. Their results were, however, skewed by outliers. These affected the overall accuracy to the degree that non-doctors ranked higher, despite the fact that their spread of judgement was wider than that of the majority of doctors. These effects are clearly displayed in the SPSS boxplots shown in figure 7.9.

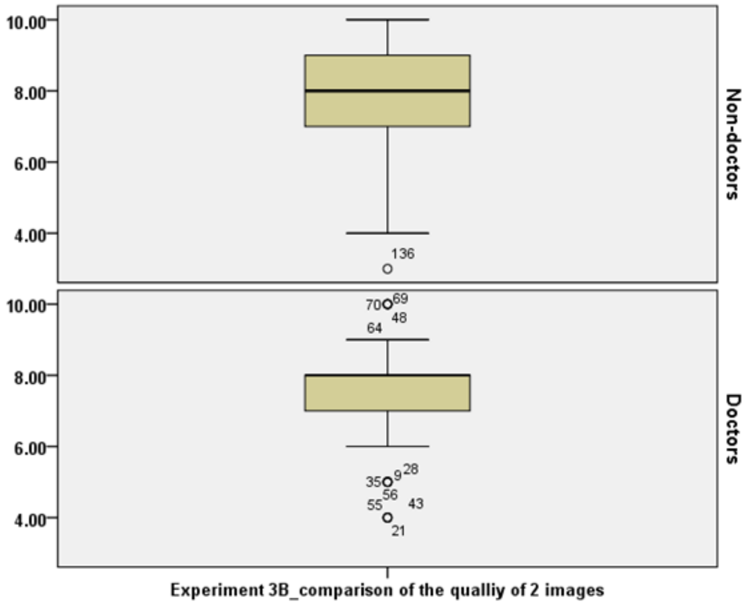


Figure 7.9: Comparison of accuracy for doctors and non-doctors

Comparison of the accuracy of the two groups of participants by Mann-Whitney U test indicated that non-doctors were significantly more accurate in their comparisons of image quality than were doctors (U = 1912, Z = -2.308, p = 0.021) although, as previously stated, the overall accuracy of the doctors was adversely affected by a number of extreme results. A further consideration was that this group of experiments involved assessing the quality of images, rather than the degree of redness shown in an image. As the majority of non-doctor participants were directly involved with the fields of computing and IT, the technical quality of digital images was a concept with which they might already have been familiar.

7.8.8 Results of experiment 3E

The SPSS boxplot in figure7-10 demonstrates close similarity in the accuracy of the two groups of participants. Non-doctors ranked slightly higher but were somewhat less consistent in their answers, as shown in the upper and lower quartiles of the boxplot. Statistical comparison by Mann-Whitney U test revealed that the difference between the two groups was not significant (U = 2154, Z = -1.254, p = 0.210). The results again affirmed hypothesis 1, but, as no scale was involved in this experiment, it was not possible to evaluate these results against hypothesis 2.

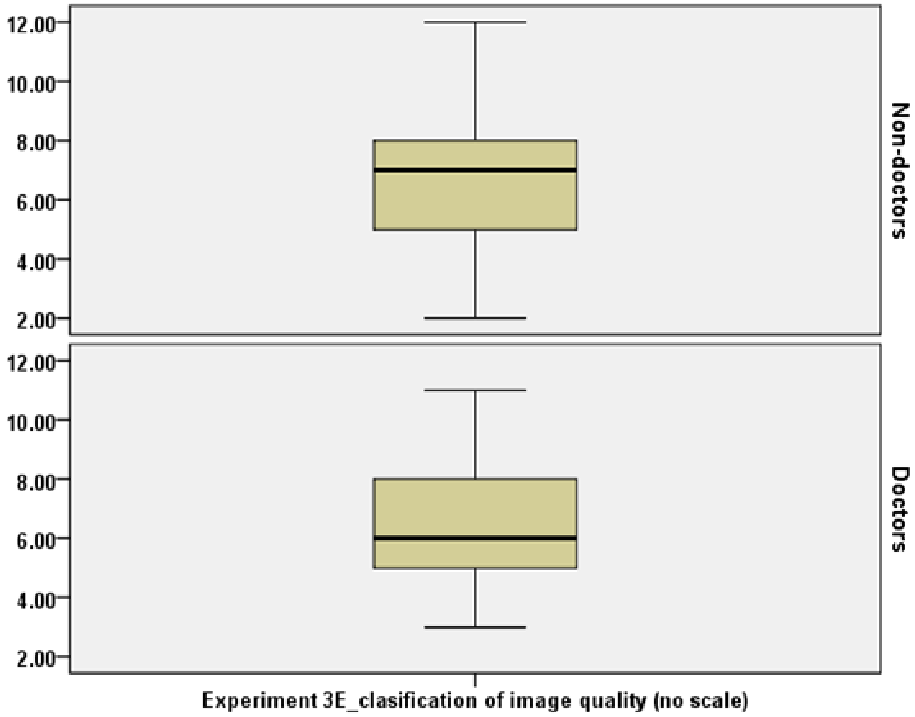


Figure 7.10: Comparison of accuracy of doctors and non-doctors

7.8.9 The results of experiment 3F

In terms of assessing whether the quality of the images presented was acceptable or not, both participating groups showed an overall good standard of accuracy and reasonable consistency, with only a few outliers indicating extreme values (figure 7.11). As with experiment 3E, comparison of the participant groups by Mann-Whitney U test showed no significant difference between them (U = 2273.5, Z = -0.808, p = 0.419) once more confirming the proposition of hypothesis 1. In terms of their construction, experiments 3E and 3F were similar. In both cases a series of twelve images was created by digitally modifying a single picture. Participants were then asked to categorise the images based on their quality. The essential difference between the two experiments was that experiment 3E required evaluation of the images into three categories of quality whilst in 3F it was only required that they classify the images as either acceptable or unacceptable in terms of quality. This would satisfactorily explain why the level of accuracy was higher in 3F. In this series of experiments, it was clear that non-doctors performed at least equally as well as doctors in terms of accuracy and consistency although, as mentioned earlier, this may well have been due to these particular non-doctor participants' backgrounds in IT.

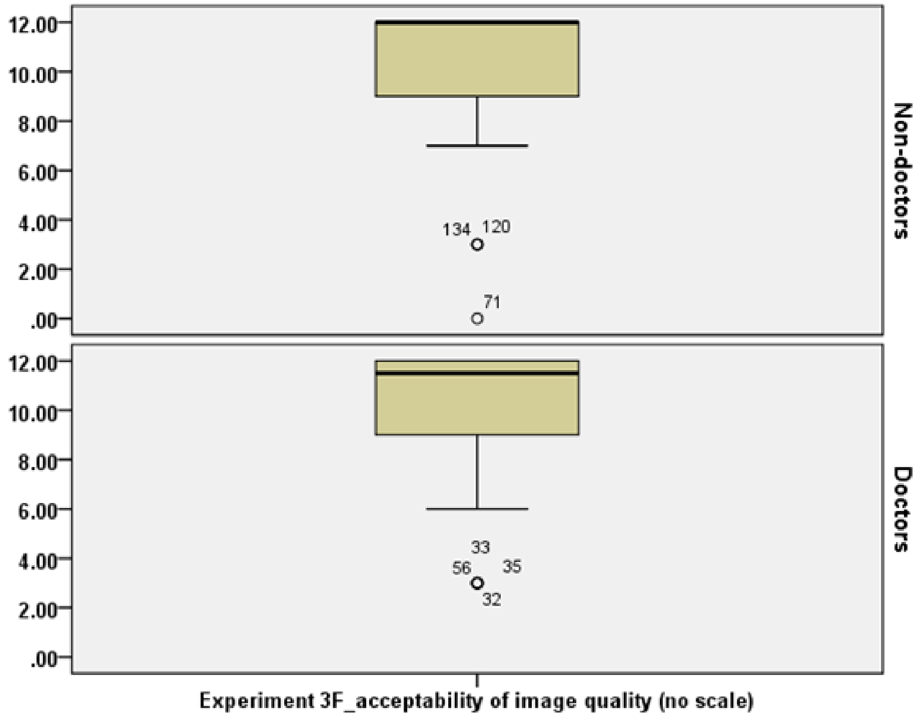


Figure 7.11: Comparison of accuracy of doctors and non-doctors

7.8.10 Results of experiment 4

The experiments of this section were identical to those of experiment group 1 with the exception of the application of a three-point scale system as opposed to the five-point scale utilised in group 1. This made it possible to carry out a direct comparison of the results in order to establish the effects of the different scales on the accuracy levels. This was not possible, however, with experiment 4B as no scale was involved. Its results had, therefore, to be considered in isolation.

As this series of experiments used the same images and involved the same participants as in experiment series 1, it was deemed necessary to separate them by interposing experiment groups 2 and 3, in order that the results should not be skewed by memory or learning.

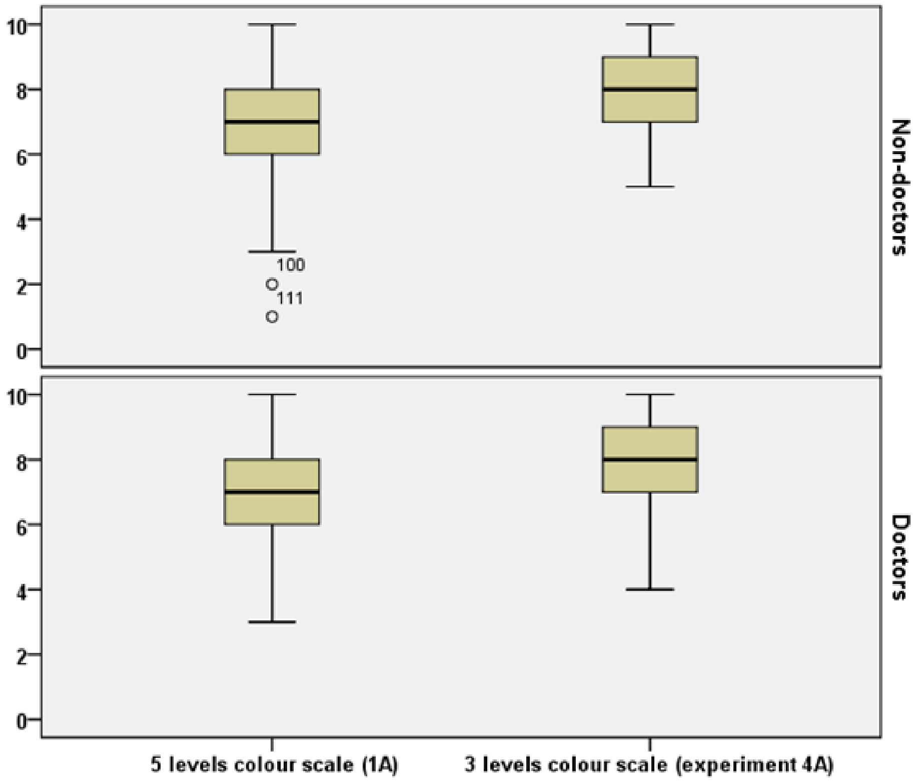


Figure 7.12: Comparison of accuracy between experiments 1A and 4A

7.8.11 Results of experiment 4A/1A

It is clear from the SPSS boxplot, shown in figure 7.12, that both doctors and non-doctors were more accurate when using the three-point scale. The results of Mann-Whitney U tests demonstrated that there was no difference between doctors and non-doctors in either experiment 1A (U = 2394, Z = -0.237, p = 0.813) or 1B (U = 2422.5, Z=-0.118, p=0.906) in experiment 4A. Wilcoxon signed ranks tests showed that comparison of the use of a 3-point scale with a 5-point scale resulted in a high degree of significant difference (p = 0.001). Comparing the use of a 5-point scale to a 3-point scale there was a negative rank movement of 94, whilst the positive rank movement was only 27.

7.8.12 Results of experiment 4B

As in previous experiments, doctors showed greater consistency overall than non-doctors, but again the results were skewed by outliers. Non-doctors showed a lesser degree of consistency, but all their results were within the representative quartiles (figure 7.13).

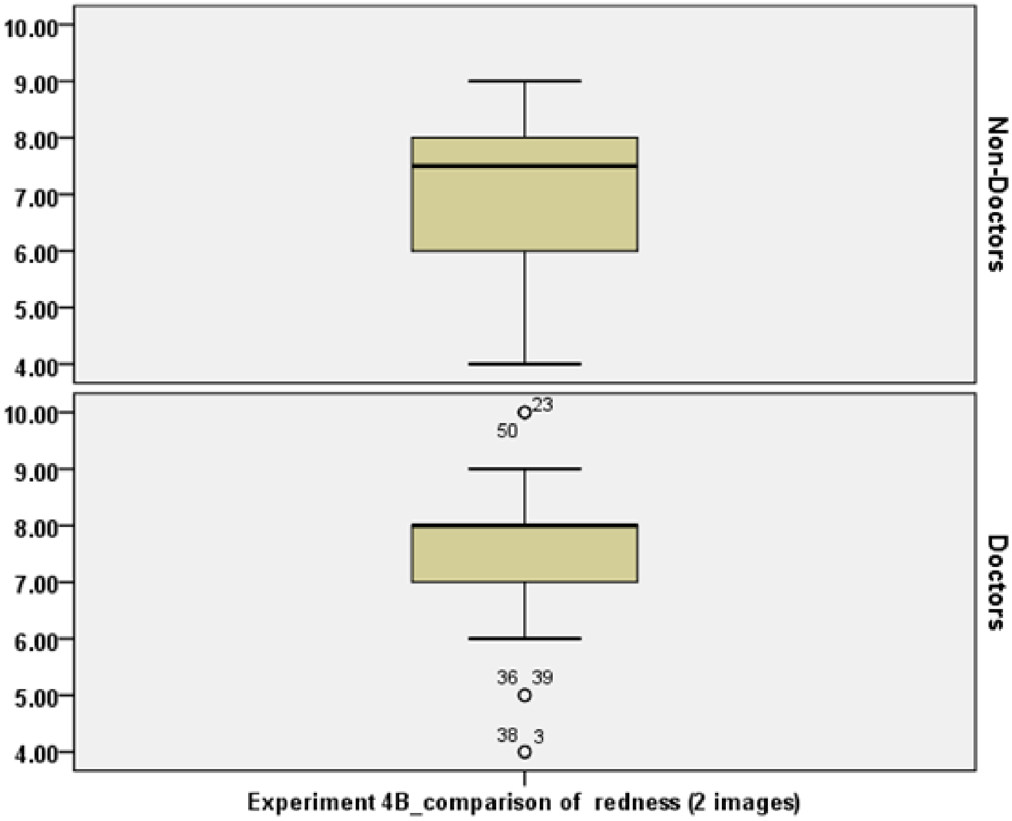


Figure 7.13: Comparison of accuracy of doctors and non-doctors

Using a Mann-Whitney U test to quantify the difference between the two groups showed that statistically there was no significant difference between them (U = 2255.5, Z = -0.837, p = 0.403). Hypothesis 1 was again confirmed.

7.8.13 Results of experiments 4C/1C

Figure 7.14 demonstrates a large increase in both accuracy and consistency for both doctor and non-doctor participants, although the relative accuracy of both groups remained similar. Mann-Whitney statistical comparison showed that there was no significant difference between the accuracy of the two groups when using either scale. Experiment 1C (U = 2403, Z=-0.201, p=0.841). Experiment 4C (U = 2417, Z = -0.144, p = 0.885), confirming null hypothesis 1. Comparison of the results when using the two scales, by means of Wilcoxon signed ranks tests, showed that there was a negative ranking shift of 127, whilst the corresponding positive ranking shift was only 4. This demonstrates a large increase in the accuracy of the results, (p=0.001).

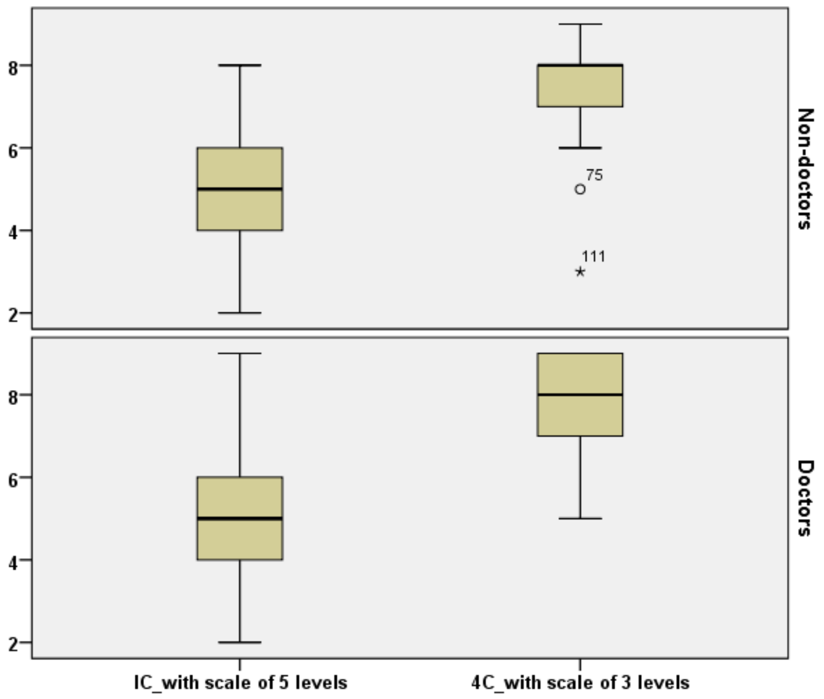


Figure 7.14: Comparison of doctors and non-doctors in 1C and 4C

7.8.14 Results of experiments 4D/1D

The boxplot analyses shown in figure 7.15 indicate that, using a 3-point scale, both doctors and non-doctors were more accurate in their evaluations. Comparison of relative accuracy by Mann-Whitney U test showed that, regardless of whether the scale used by a group featured 3 or 5 points, both groups ranked closely (between 68 and 72), and that there was no statistically significant difference between them. Experiment 1D ($U = 2297.5$, $Z = -0.647$, $p = 0.518$). Experiment 4D ($U = 2282.5$, $Z = -0.716$, $p = 0.474$). Null hypothesis 1 was therefore confirmed. A Wilcoxon signed ranks test was used to examine the difference between the effects of the 3-point and the 5-point scale. This showed a major shift in ranking with the 3-point scale compared to the 5-point scale. A positive rank change of 104 points versus a negative rank change of 18 points demonstrated a considerable increase in accuracy with the use of the 3-point scale, with a high degree of significance ($p = 0.001$).

7.8.15 Experiments 1E and 4E. Comparing scales of 5 and 3 levels

The boxplot in figure 7.16 illustrates the effects, on both groups of participants, when identical experiments were conducted with a scale of five points (1E) and with a scale of three points (4E). A considerably higher degree of accuracy may be observed in 4E.

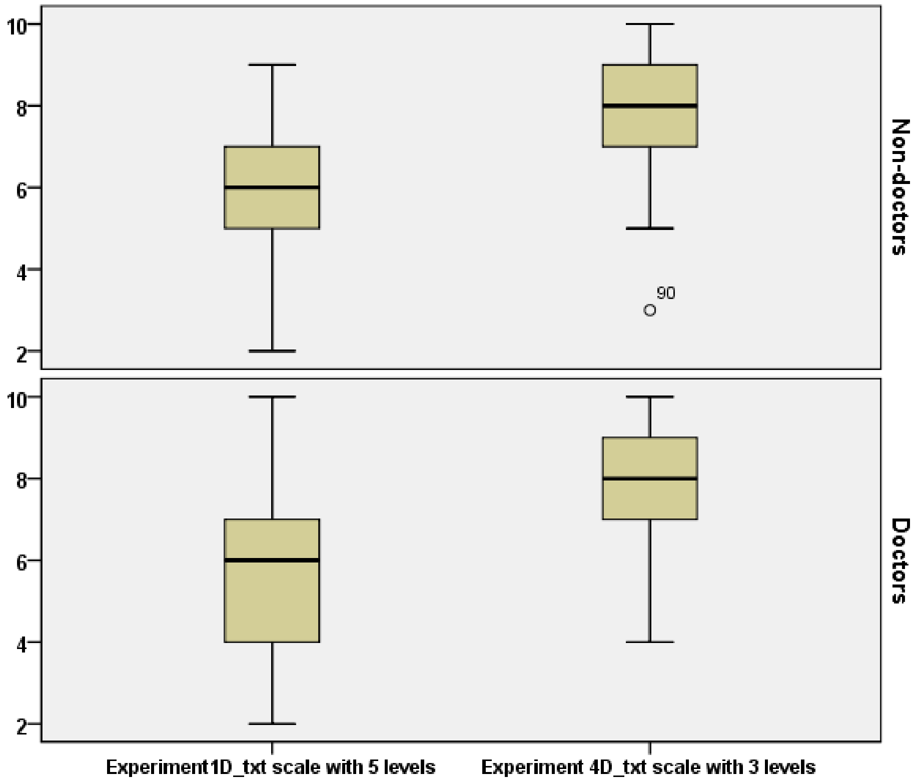


Figure 7.15: Comparison of doctors and non-doctors in 1D and 4D

Mann-Whitney U tests showed that there was no significant difference between the two participating groups using either scale. For experiment 1E U = 2042.5, Z = -1.728, p = 0.084. For experiment 4E U = 2308.5, Z = -0.619, p = 0.536. Hypothesis 1 was therefore confirmed. Wilcoxon signed ranks tests comparing the effects of the two scales showed that a major increase in accuracy was experienced in the case of the 3-point scale. Positive ranks changed by 123 whilst negative ranks changed by only 9 (p = 0.001).

Experiments 2A and 5A represented repeats with scales of five points and three points, as was the case with experiments 1 and 4. The results were therefore analysed in the same manner. The boxplot in figure 7.17 shows a significant increase in overall accuracy, for both groups, where the 3-point scale was employed. Comparison of the two groups of participants, by Mann-Whitney U tests, showed that there was no difference between them in experiment 2A (U = 2016, Z = -1.822, p = 0.068). In experiment 5A doctors were shown to be significantly more accurate than non-doctors, achieving a mean rank of 79 against 62 (U = 1859, Z = -2.475, p = 0.013). This was one of the few occasions on which doctors were significantly more accurate than non-doctors. It was not possible to find a specific reason for this although it seemed likely that the

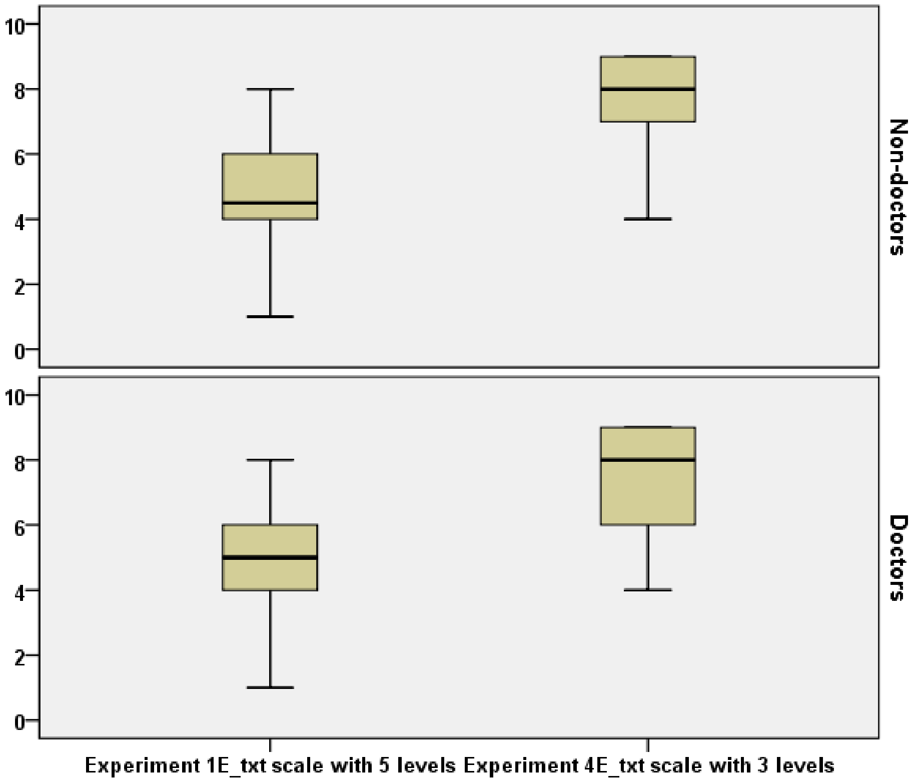


Figure 7.16: The effect on accuracy of scale granularity

results were within the bounds of normal experimental error. Wilcoxon signed ranks tests, comparing the effects of the two scales employed, indicated that an increase in accuracy was experienced with the 3-point scale used in 5A. Only 8 negative ranks were recorded whilst 127 positive ranks were observed (p = 0.001). The boxplot in figure 7.18 clearly shows the increase in accuracy when the 3-point scale was employed.

A Mann-Whitney U test comparison of doctors and non-doctors showed no significant difference between the groups in any of the experiments. Experiment 2B had results U = 2317, Z = -0.558, p = 0.577. Experiment 5B produced U = 2098, Z = -1.477, p = 0.140. Null hypothesis 1 was therefore confirmed.

Comparison by Wilcoxon signed ranks tests, demonstrated a significant increase in accuracy when using the 3-point scale (in experiment 5B). Negative ranks were 10 whilst positive ranks were 140 (p = 0.001).

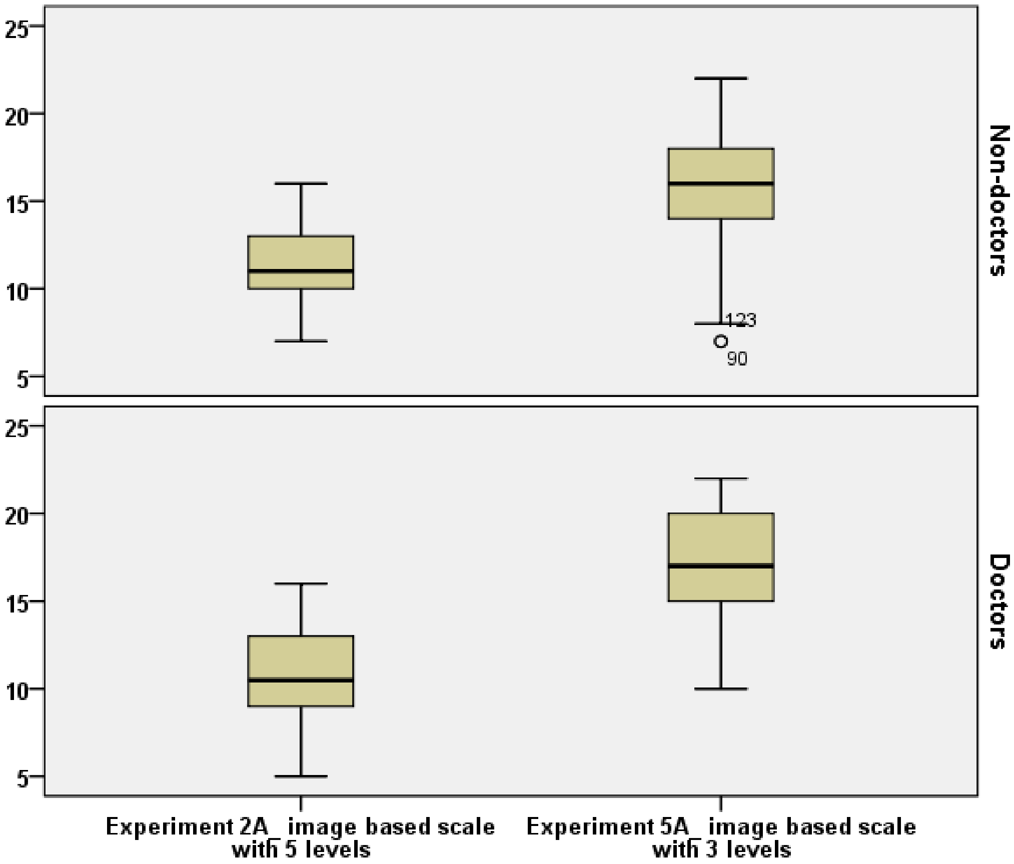


Figure 7.17: The effect on accuracy of scale granularity

7.8.16 Accuracy based on the degree of redness in images

It was observed in this group of conjunctivitis experiments, as in the earlier cellulitis experiments, that a small degree of difference in an image was much harder for the participants to distinguish than it was in either largely different or identical images. This was seen to apply equally in the case of the degree of redness presented in an image. In order to illustrate this effect, a boxplot representation was produced from the eligible data (figure 7-19). It can be clearly seen from this that both groups of participants showed similar results, demonstrating that experience or knowledge do not assist in the performance of this differentiating task. In addition to the boxplot representation, statistical confirmation was carried out. This data may be accessed via the Dropbox link at http://bit.ly/1l5WyTy.

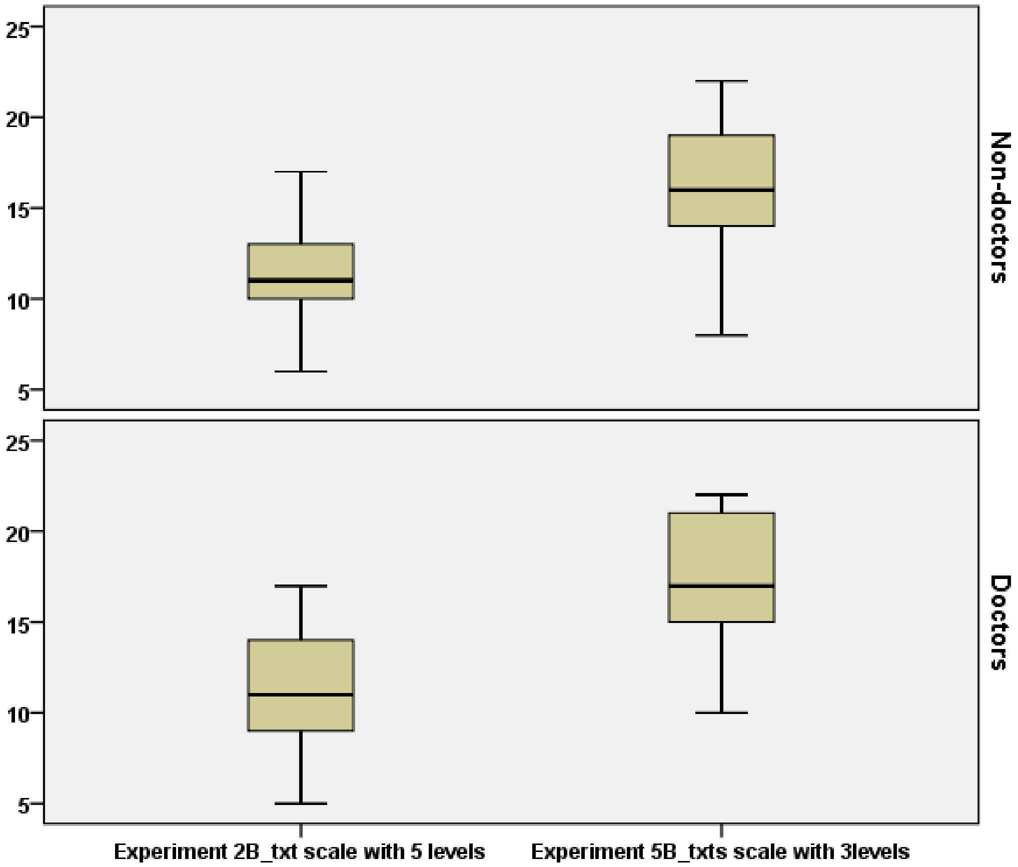


Figure 7.18: Scale of three levels was employed

Results of experiments 6 and 7

These experiments were repeats of experiments 1 and 2, but with different participants. These were conducted simply as a check on the previous findings because the new participants made themselves available for only two experiments. Analysis of the results revealed that the outcomes were similar to those of experiments 1 and 2 and thus confirmed those original findings. It was unnecessary, therefore, to list the results individually within the body of the thesis. All of the test data and tables of analysis may be accessed via the Dropbox link at http://bit.ly/1l5WyTy.

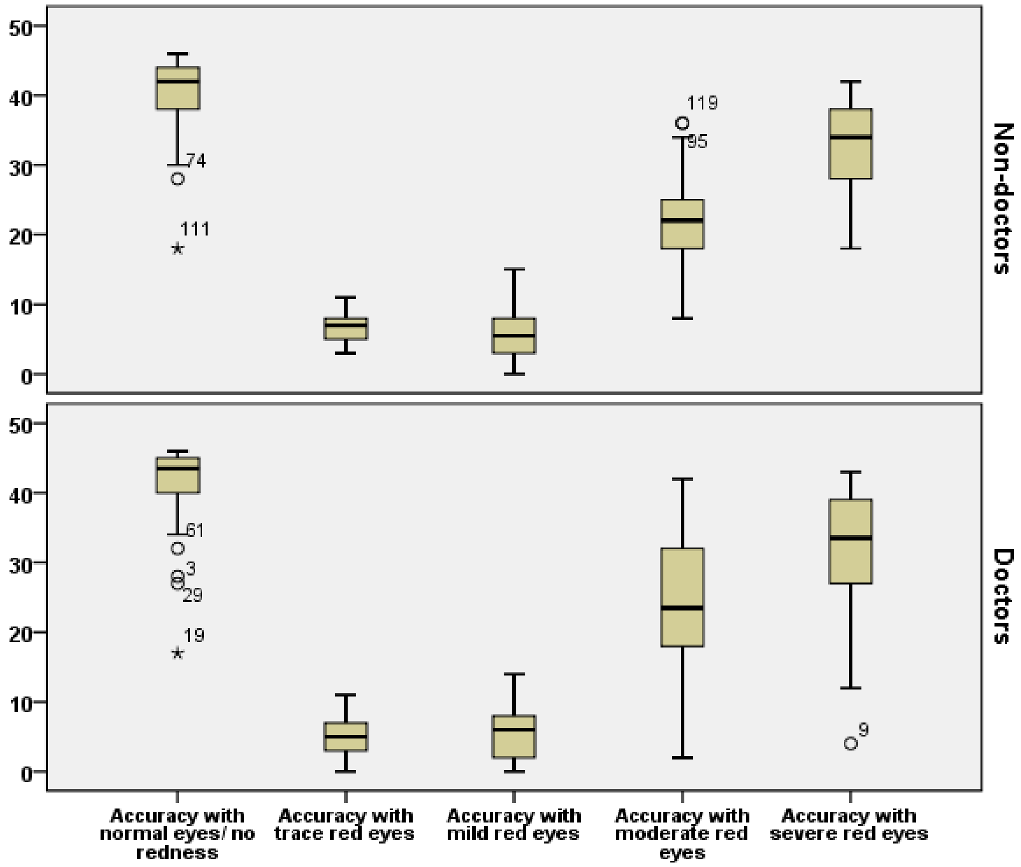


Figure 7.19: The effect of degree of redness on accuracy

7.9 Relationships between Accuracy, Confidence and Time

Spearman rank-order correlation tests were conducted for each experiment of this chapter in order to determine if there were any statistical relationships among the accuracy, confidence, and time taken when performing the experiments. The correlation tests were carried out for doctors and non-doctors separately, as well as for the two groups combined. As this resulted in a total of 21 tables, one example is shown as an example in table 7-8 and the others may all be accessed via the Dropbox link at http://bit.ly/1l5WyTy.

Table 7.8: Spearman correlation in accuracy, confidence and time-experiment 3

Experiment 3 for all participants		Accuracy	Confidence	Time
Accuracy	Correlation coefficient	1.000	0.129	0.090
	Sig. (2-tailed)	.	0.130	.288
	N	140	140	140
Confidence	Correlation coefficient	0.129	1.000	0.145
	Sig. (2-tailed)	0.130	.	0.087
	N	140	140	140
Time	Correlation coefficient	0.090	0.145	1.000
	Sig. (2-tailed)	.288	0.087	.
	N	140	140	140

Table 7-8 is a typical example of the results of the correlation tests, in that low levels of correlation were found in all cases. These findings supported all three of the null hypotheses relating to accuracy, confidence and time. This is further explained in the summary and discussion of the results.

7.10 Machine learning experiments

A supervised machine learning program was developed for the purpose of automatically classifying images showing conjunctivitis and cellulitis into two, three and five classes of severity. The results obtained from the program were then compared with the results of a human conducted analysis.

As mentioned in chapter 2 (Literature review) section 2.8.1, this classification task was most amenable to automation because of the systematic classification labelling that had previously been devised and applied to the images. The labels were generated from the agreed answers of three doctors who were expert in the field, as described in section 7.5.2.

7.10.1 Colour feature presentation

The images were represented in the form of feature vectors. Colour histograms were used for feature representation, since the principal feature distinguishing healthy from affected eyes or skin is redness. Prior consideration in the construction of an image classification is the colour space. It is common in image capture and transmission to use an RGB colour space. This however can have disadvantages in classification.

An alternative is to use an HSV colour space as this is more intuitive and is unaffected by changes in illumination and camera direction (Sergyan 2007). There exists a standard formula for the translocation of an RGB colour space into HSV. It was decided, therefore, to use an HSV colour space in the construction of the algorithm for the auto-evaluation system.

7.10.2 Nearest neighbour

The k-nearest neighbour classification algorithm used by the system was drawn from the OpenCV open source machine learning software library. K-nearest neighbour is a pattern recognition algorithm which matches an object with those others which have the most similar features. These are termed the k-nearest neighbours, k being an integer value. If k = 1, the object is matched with the single nearest neighbour. In this instance, 1-nearest neighbour was used. Images for analysis were matched by the system with the single nearest neighbour from a set of labelled training images which had been classified into 5, 3 or 2 categories ("Normal", "Trace", "Mild", "Moderate" and "Severe"; "Normal", "Moderate" and "Severe"; "Normal" and "Red"). The code used for image classification and the results of the auto-evaluation may be viewed in appendix 3 or downloaded from GitHub https://goo.gl/Ab7hsY.

7.10.3 Training and testing

The supervised machine learning was divided into two stages, training and testing. In the training stage, the algorithm was exposed to classified and labelled training images. The programme then analysed the image features, colour histograms and colour space, and associated their values with their respective severity defined classes. 76 cellulitis images and 25 conjunctivitis images were used for this process. Further classification was performed to accommodate three sets of analysis. The first set of analyses (class 2) defined two categories of image, "Normal" and "Red". The "Normal" category in this analysis contained all of the images defined as "Normal", which refers to healthy eye or skin, and the "Red" category contained all other images (i.e with Trace, Mild, Moderate and Severe degrees of redness). This analysis is a binary classification that looks at only images of patients that have indications of cellulitis/conjunctivitis and patients with no indications (absence of redness). The practical application of this analysis refers to the early stage of the hospital's diagnosis, when a patient is initially assessed. The second analysis had three categories, "Normal", "Moderate" and "Severe" with "Normal" and "Severe" images assigned to their eponymous categories, and with all other images assigned to the "Moderate" category. The practical application of this analysis is at the second stage of the hospital's diagnosis, when a patient is found to be affected, and therefore this analysis will classify the patient based on the severity of their condition to determine the urgency with which referral is required. The final analysis (class 5) assigned all types of image to their respective severity categories, "Normal", "Trace", "Mild", "Moderate" and "Severe". Doctors can use this level of analysis to understand more precisely the severity of the ailment, and therefore determine the right treatment and consultation approach. Experiments were carried out for both the cellulitis and conjunctivitis conditions.

In the testing process the algorithm analysed the features of unlabelled testing images and matched them to their nearest neighbour from the training images. The category label of the selected nearest neighbour training image was then assigned to the most similar unlabelled testing image, categorising it. The accuracy of the auto-evaluation system was analysed by comparing the labels predicted by the algorithm to the actual class labels of the images. The accuracy of the auto-evaluation system was then compared to that of human participants.

Machine Learning Classification Results

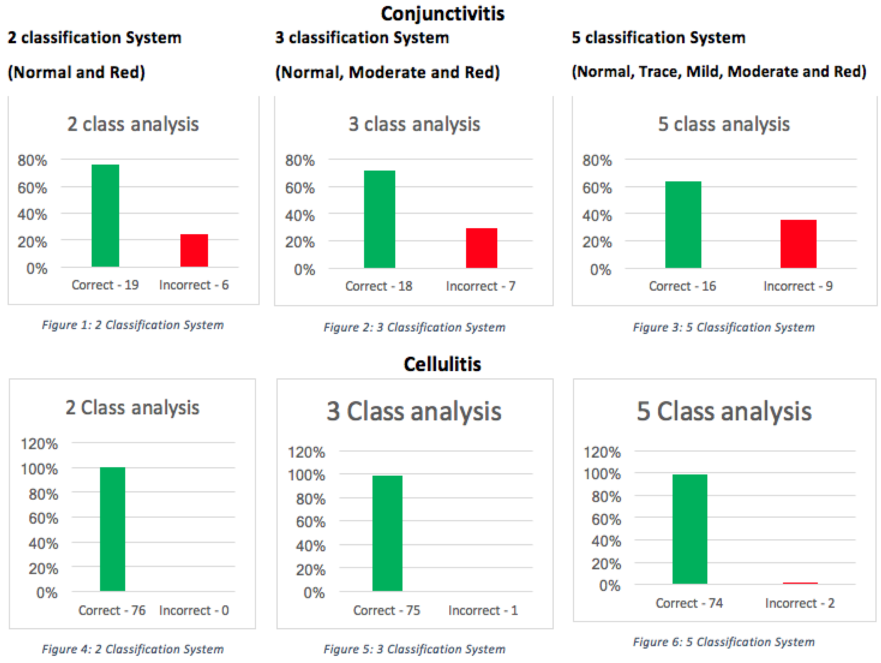


Figure 1: 2 Classification System

Figure 2: 3 Classification System

Figure 3: 5 Classification System

Figure 4: 2 Classification System

Figure 5: 3 Classification System

Figure 6: 5 Classification System

Figure 7.20: Machine Learning Classification Results

As seen from the results plots in Figures 7.20 above, the machine learning method achieved its highest accuracies in the 2 classification system, in both the cellulitis and conjunctivitis analyses, with 76% accuracy for conjunctivitis and 100% accuracy for cellulitis. This demonstrates that the method can easily distinguish the visual appearance of healthy and unhealthy tissues. This can be useful in the initial diagnostic stage in primary care centres to decide whether a pathology is present or not, and if the patient needs to be seen by a doctor for further examination. This can save time and resources for the health service provider.

The 3 classification system achieved an accuracy of 72% for conjunctivitis (figure 2 in 7.20) and 99% for cellulitis (figure 5 in 7.20). This level of analysis is particularly important in showing the severity level of the pathology. This would be of assistance when determining the most appropriate treatment.

As presented in the results of the 5-classification system, the machine learning method achieved its lowest accuracy and made more errors in the categorisation of the two pathologies. This analysis requires the levels of severity of the symptoms of a condition to be precisely categorised based on image features. Such an analysis will generally require to be fine grained with only small differences in features being sufficient to distinguish the categories. A classification system working at this level of granularity has application in monitoring the development of a condition and tracking the effect of treatment. This is an interesting area for future investigation.

It is apparent that the machine learning method achieved a higher accuracy in the cellulitis analysis. This is caused by inaccuracies in the learning system's model as a result of the higher volume of noise in the conjunctivitis images. An example of such noise is the skin colour around the eye, which is sometimes detected by the system's image processing apparatus and interpreted as part of the area of interest in the image. The cellulitis images that were used captured mainly the presumed affected area, restricting the analyses to the area of interest each time, whereas the conjunctivitis images used included areas of unaffected skin around the areas of interest. Further studies might be profitably directed at investigating ways of isolating specific areas of interest for a more accurate analysis.

The tables 1-3 in Appendix B show detailed results for conjunctivitis and tables 4-6 for cellulitis from the classification model.

Table 7.9: Classification Results for RGB and HSV - Cellulitis

Classification System	Cellulitis Results HSV	Cellulitis Results RGB
Class 2	1.0 (100%)	0.59 (59%)
Class 3	0.99 (99%)	0.64 (64%)
Class 5	0.98 (98%)	0.89 (89%)

Table 7.10: Classification Results for RGB and HSV - Conjunctivitis

Classification System	Conjunctivitis Results HSV	Cellulitis Results RGB
Class 2	0.76 (76%)	0.68 (68%)
Class 3	0.72 (72%)	0.52 (52%)
Class 5	0.64 (64%)	0.56 (56%)

Machine Learning Classification Results for RGB and HSV.

It is assumed that images of eyes with conjunctivitis will have a higher concentration of red colour. Thus, an attempt was made to analyse only the RGB red channel of the images (R = 1, G = 0, B = 0), in order to compare the results obtained by the system when using the HSV and the RGB colour spaces. The experiment proved that the HSV transformation system produced a more accurate result for all experiments, as presented in table 7.9 and 7.10. Full results can be seen in appendix E.

Machine Learning Errors

The tables 9 to 12 in appendix D show the level of error with description of misclassified images showing conjunctivitis by the auto-evaluation system. For level 1 errors, patients' number 10, 23 and 24 (see also figure 7.28 below) were identified by a doctor as Trace but were classified by the system as Normal. This is considered a level 1 misdiagnosis due to the closeness in the red colour features between the two groups of images. See table 10 in appendix D.

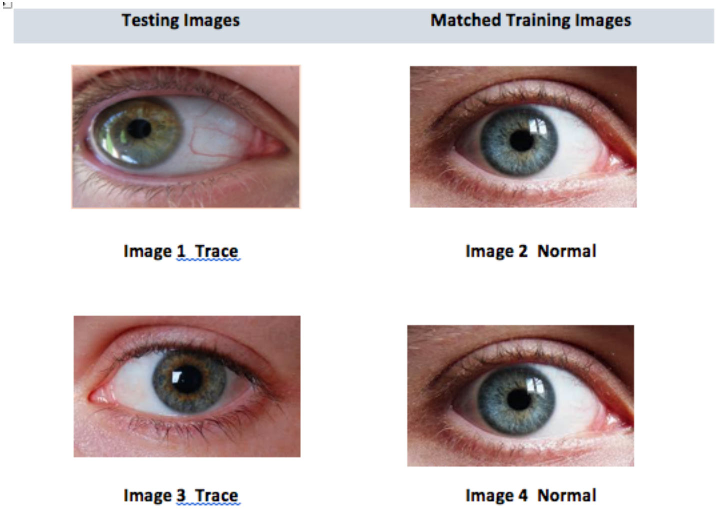


Figure 7.21: HSV classification error (1 level)

For 2 level errors, patients 7 and 14 (figure 7.22) that were identified as Mild were classified by the system as Normal, in this case the degree of misdiagnosis is higher and poses more risk because it could lead to worsening of a patient's condition if not treated.

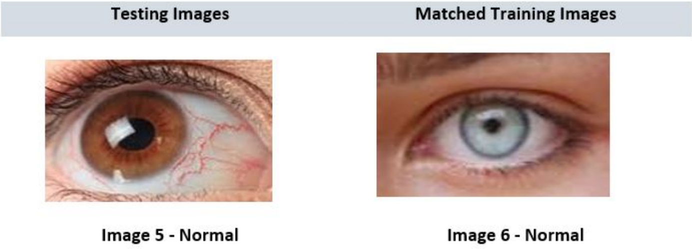


Figure 7.22: HSV classification error (2 levels)

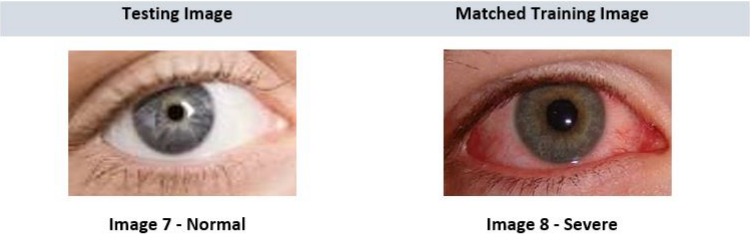


Figure 7.23: HSV classification error (5 levels)

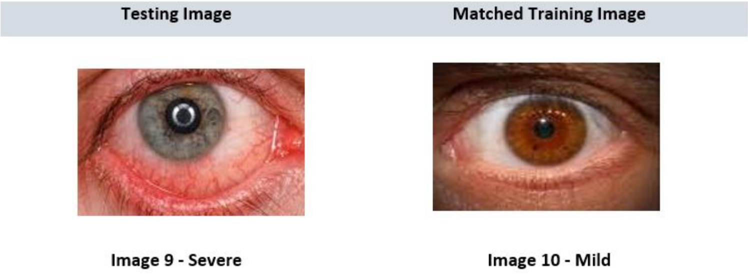


Figure 7.24: HSV classification error (1 level)

Of the 5 level errors that were observed, the diagnoses recorded for patient 20 displayed the widest discrepancy (Normal classified as Severe. As shown in figure 7.21), Such an error could potentially threaten a patient's eyesight if treatment were to be misprescribed as a result.

The results in table 7, appendix D are the same for the two, three and five classification systems, with addition of a 1 level error for patient 25 in the three and the five classification systems (table 8, and 9). In that case a Severe condition was classified as Mild, though this is not considered an error in the two-classification system.

The 5 classification system also made errors. Patients 11 and 19 had conditions that were rated by a human doctor as Trace but were classified by the system as Mild. This is a 1 level error and not considered an error in the 2 and 3 classification systems.

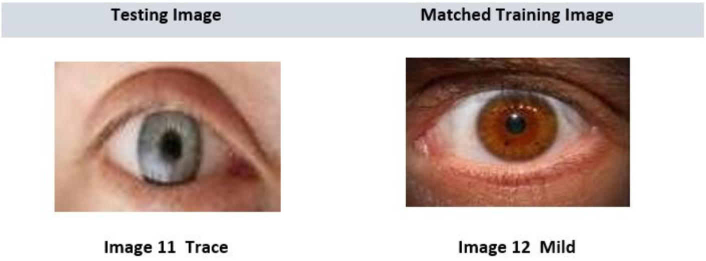


Figure 7.25: HSV classification error (1 level)

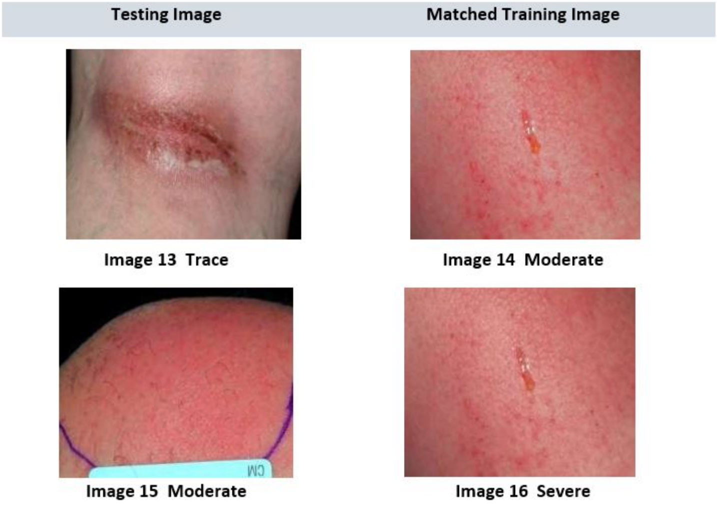


Figure 7.26: HSV classification error (1 level)

The table 10 in appendix D show the level of error with description of misclassified images, showing cellulitis by the auto evaluation system. This result was found in the two, three and five classification systems. The result from the cellulitis result shows that the two errors from images 8 and 34 are minor, with one and two levels. See table 10 in appendix D.

Conclusion

Although a considerable amount of development and refinement is still required, the basic principle of the proposed auto-evaluation system is sound. Even this elementary model has proved to be more accurate than the human participants, especially in the 2 and 3 classification systems. An auto-evaluation system, if it were to be realised, would facilitate the rapid analysis of large numbers of images, which could potentially range up to thousands. Such high-volume analysis would be time consuming and labour intensive for humans to undertake.

7.11 Summary and discussion of the results

Throughout this whole series of experiments, comparing the accuracy of doctors and non-doctors, whether in assessing redness levels or in comparing image pairs or in evaluating image quality, consistently found that there was no difference between the two.

This corroborated the findings of the cellulitis group of experiments, and also confirmed hypothesis 1, that there would be no difference between the respective performances of doctors and non-doctors. The boxplots in figure 7.27 present a graphical representation of these findings.

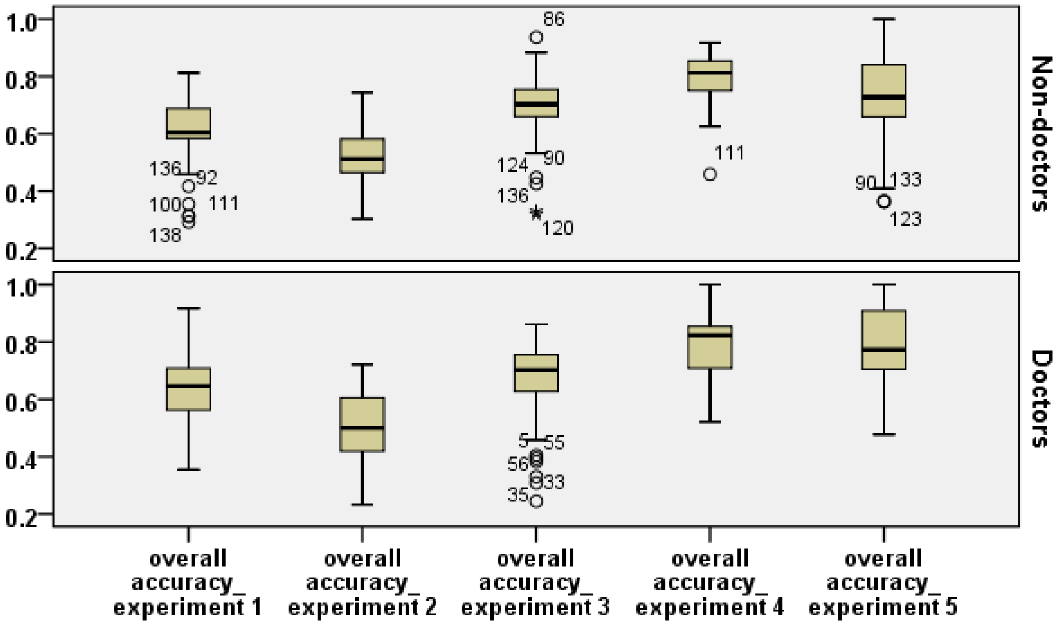


Figure 7.27: Accuracy and consistency levels in experiments 1 to 5

The introduction of guidance scales produced higher accuracy levels, but not to the same degree as observed in the cellulitis experiments. Many of the experiments in chapter 7 involved the presentation and rating of multiple images and it is possible that direct comparison of these images formed its own ad-hoc scale and therefore rendered the grading scales less necessary. Hypothesis number 2 was contraindicated in these results, as it proposed that the introduction of a scale would have no effect on accuracy.

Chapter 7 repeated experiments with scales of three and scales of five points in order to observe their respective effects on accuracy. Intuitively it seemed evident that reducing the number of intervals must reduce the potential for errors, and should therefore have resulted in higher accuracy levels. It was interesting to observe that the effects were indeed very much as expected, and that no sudden variations occurred.

The degree of confidence that the participants had in their decisions was recorded throughout the series of experiments, and were compared with their measured accuracy. Confidence was high throughout, regardless of the difficulty of the task, as evidenced by the boxplots in figure 7.28. Statistical analysis, however, showed that there was no correlation between confidence and accuracy.

It was notable that confidence levels also increased as the scale was reduced from five to three levels. This was contrary to expectation. As an example, comparing experiment 4 (using a 3-point scale) and experiment 1 (with a 5-point scale) by Wilcoxon signed ranks test resulted in a shift in ranking of 99 positive ranks against 32 negative ranks (Z = -5.029, p = 0.001.). Full tables of statistics may be accessed via the Dropbox link at http://bit.ly/1l5WyTy.

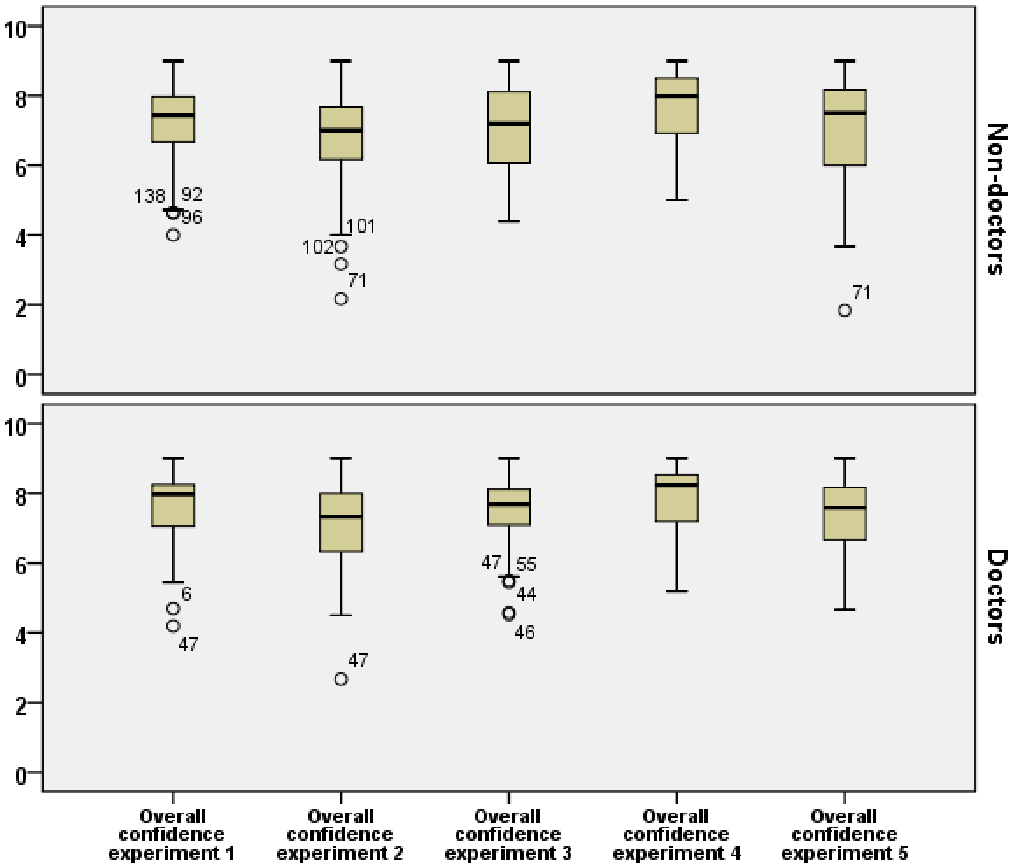


Figure 7.28: Confidence levels in experiments 1 to 5

The study also showed that the accuracy attained by the participants bore no relation to the time they took to complete the experiment tasks. As with confidence, the time taken was seen to decrease when the 3-point scale was used as opposed to the 5-point scale. As an example, comparing experiment 4 (3-point scale) with experiment 1 (5-point scale) by Wilcoxon signed ranks tests showed a shift in ranking of 119 negative ranks against 21positive ranks (Z = -7.359, p = 0.001.). Statistical tables for this example and other examples can be accessed via the Dropbox link at http://bit.ly/1l5WyTy. These findings confirmed hypothesis number 3, which proposed that there would be no relationship among accuracy, time taken and confidence. Hypothesis number 4 stated that there would be no difference between the two groups of participants in terms of time taken or levels of confidence. This was also confirmed.

The most interesting aspect of this series of experiments was observed in the contextual background of the images. In the cellulitis experiments, it was apparent that the presence of unaffected skin in the background of images resulted in more accurate judgement of redness levels. In the conjunctivitis experiments, however, the opposite appeared to be the case. It was noticeable, more in the case of non-doctors than doctors, that where the surrounding skin had a red tinge to it, the judgement of the participants was affected and wrong assessments of the degree of redness were more common.

This study also suggested that the implementation of a standard protocol for the minimum standard of image quality, and for the area to be captured for the two different pathologies, could well lead to the more effective application of telehealth to this field of medicine.

Further consideration of these points is developed, in conjunction with other findings, in the discussion and conclusions presented in chapter 8.

Chapter 8

Conclusion

This last chapter of the thesis starts with a critical discussion of the main results and a summary of the experimental work in relation to the hypotheses and objectives of the study. The second section of this chapter focuses on the originality of the research and its contributions to the field. The chapter then records the key limitations and reflections on the study. Finally, the chapter ends with proposals for further research.

8.1 Critical Discussion of the Main Results

The pilot experiment:

The pilot experiment (chapter 4) was intended to be a model by which to test the methodology of the main experiments of the thesis. It was immediately clear that the pilot achieved more than this. It was apparent, in the early stages, that attempting unguided verbal descriptions of colour could never work.

The range of language used to describe the presented images was too diverse to offer any practical value in terms of colour description. Hence it was obvious that some humongous reference point, e.g. a common language, had to be applied in order that the sender and receiver of descriptive terms could immediately understand the nature of the red colour being discussed.

As a logical extension of this idea, it was further considered that the use of a numeric scale of reporting might be a better means of communication. This concept was then investigated within the main experiments of the thesis (Chapters 5, 6, and 7). The existence of a pre-defined scale of measurement aided standardisation in reporting the answers of participants and the results of the experiments. It was also notable that there was an increase in the number of clusters and subclusters created by the participants after providing additional instructions to the participants in task 2B (image grouping). It was very clear that classifying colour produced different results from the use of personal judgment alone in describing colours which is more subject to individual differences.

An additional step followed the realisation that plain colour, without context, might be an inappropriate form of stimulus and a move was made to conduct the main experiments using images of real medical pathologies. Another controversy in the literature review (chapter 2) is sex-related differences in visual processing as shown by Chaudhari and Shaw (2012), Abramov et al. (2012) and Kuehni (2001).

It was not possible to recruit women for the pilot as none volunteered to take part. Later experiments included females, 61 females took part in the cellulitis experiments and 70 took part in the conjunctivitis experiments. No significant statistical difference was found between the results of males and females.

Hypothesis 1: The impact of clinical experience on the accuracy of colour perception

There is difference between doctors within their group, non-doctors within their group, and between the two groups in the level of accuracy in relation to colour perception due to medical background and clinical experience. (The null hypothesis of which is: There is no significant difference between the participants)

The results throughout the cellulitis and conjunctivitis experiments 1-5 were in generally in agreement with the null hypothesis as the accuracy of the answers showed clearly that, with two exceptions, there was no clear difference between the performance of doctors and non-doctors. The level of experience (Reason, 2000) and awareness (Schultz et al, 2011) were noted to be important factors in the minimisation of error; however the results showed a general homogeneity in the accuracy of the two groups highlighting that these did not play an important role for the subject of this thesis. The exceptions to this arose firstly in the cellulitis experiments 3B of chapter 5, where images of pathology, showing no contextual background, were evaluated more accurately by the doctors. This result was relatively unsurprising, as these images (8 and 9) played directly to the strengths of the doctors who would be more familiar with the pathology.

The second exception arose in the conjunctivitis experiments 1B of chapter 7 where, once again, contextual background had a marked influence. Images 7 and 8 displaying a redness of the skin of the eyelids and surrounding areas affected the accuracy of the non-doctors participants, who allowed their judgment to be affected by the general redness. Doctors, on the other hand, were more focused on the actual redness of the sclera, and were therefore more accurate. In general, however, the null hypothesis number 1 was confirmed as, thus far, there was little variation between the accuracy levels of the two participating groups.

It might reasonably have been assumed that the doctors might have been more accurate than non-doctors in their recognition of the differences in quality of two images of the same pathology. It must be borne in mind, however, that the non-doctor participants were all I.T. students who could have been equally familiar with varying image qualities, which might have compensated for their lack of familiarity with the pathology. Since the results of both groups was similar, it is fair to deduce that the inability to differentiate between the two images is common to humans, rather than being a trait which depends on either experience or area of expertise.

Two important considerations arose from these findings. Firstly, there appears to be potential for people who are not qualified doctors to be involved in the primary evaluation of telehealth images. Secondly, there is real reason to believe that the capture of images for telehealth should be standardised in such a way that context is provided in the case of cellulitis, but minimised in the case of conjunctivitis, in order to focus only on the area if interest.

Hypothesis 2: The impact of a numeric descriptive the accuracy of colour perception

There is difference between the accuracy of colour perception when using numeric descriptive colour scale with or without standard pictures with three or five divisions. (The null hypothesis of which is: There is no significant difference between the two groups)

Null hypothesis number two was refuted by the findings, which showed clearly that the use of a numeric scale could have a positive effect on accuracy levels. Cellulitis experiments 2 and 3 and conjunctivitis experiment 1 analysed the effect of a numeric scale in the conjunctivitis experiments. The statistical comparisons showed a clear increase in the accuracy owing to the use of these scales and therefore the null hypothesis was rejected. It was quickly apparent that using a scale of nine divisions was far too ambitious, contributing little to accuracy and causing confusion as to in which direction the scale should operate, even when instructions were clearly given, cellulitis experiment 4 is a classic example the results of which showed that participants showed higher degrees of inaccuracy rating images of the highest severity (9) similarly to those of a much lower severity (5). The main focus, therefore, fell on the comparison of scales with either three or five divisions.

For obvious reasons, it was found that a scale of three divisions resulted in higher accuracy levels. Statistically, a scale of three divisions offers a wider window into which answers can be located and fewer intervals equate to fewer opportunities for judgement errors. Comparison of the two scales convincingly confirmed this proposition. In real terms, a scale of three divisions would probably be perfectly satisfactory for the initial assessment of telehealth images, as it would enable the most serious cases to be identified and prioritised. Later stages of classification may well benefit from a scale of five levels, but this would need to be used by individuals who are more trained or more experienced than those performing first stage.

During the analysis of the results, it became apparent that human ability to distinguish between differing levels of image quality has a finite level, beyond which images possessing different degrees of definition become indistinguishable from each other. Identification of this critical level then becomes an important area for future research, as accurate knowledge of acceptable qualities of images is essential in defining the parameters for successful image capture and transmission, if telehealth is to be able to provide a means for accurate diagnosis.

In the analysis of the experiments conducted for this thesis, it became apparent that both doctor and non-doctor participants found it difficult to distinguish between two or more copies of the same image, which had been electronically modified, with respect to each other, by 25% or less. Accuracy improved as the degrees of modification moved further apart.

Hypothesis 3: The relationship between accuracy, confidence and time

There is a relationship between the accuracy and confidence as well as the time spent in teleconsultation when using digital images that are showing cellulitis or red eyes. (The null hypothesis of which is: There is no relationship between the three variables)

This hypothesis was explored by cellulitis and conjunctivitis experiments . No significant relationship was found between confidence and accuracy, an example of which is displayed in the cellulitis image matching experiments (p=0.091) in the case of all participants. Time taken was also found to have either no correlation or a weak correlation with accuracy levels as displayed, for example, by the conjunctivitis experiment 3 showing no correlation between the two variables (p=0.288). Therefore the null hypothesis is accepted. The literature exploring the relationship between confidence levels and accuracy was inconsistent. Some studies such as those by McNiel, Sandberg and Binder (1998) suggested that participants with higher levels of confidence were more accurate whereas other studies describe increased a negative correlation between confidence and accuracy for which "overconfidence" is an important factor Kahneman

(2011). A third subsection of the literature suggests that confidence is not a predictor of accuracy (Ordinot, Walters and Van Koppen, 2009) and it is this subsection with which the current study is concordant. The literature establishes a weak but negative relationship between time and accuracy wherein participants who were given more time showed slightly better performance (Lee et al. 2000 and Brewer et al. 2002). The results of the current study were mixed, cellulitis image matching experiments were consistent with those of the literature confirming the null hypothesis (r = -.364 and P = 0.001) whereas the results of the conjunctivitis experiments showed no relationship (r = 0.090 and P = 0.288) refuting the literature and null hypothesis. While the reason behind the latter result is undetermined, it may be reasonable to assume that the use of the conjunctivitis scale, given its well defined, standardised nature, may have influenced this relationship. Confidence and time were not found to important factors in the current study and therefore it can be concluded that for future work these may not need to be taken into consideration when trying to optimise diagnostic accuracy.

Hypothesis 4: The impact of a digital image scale on accuracy, confidence and time *There are changes to the accuracy, confidence or time in teleconsultation when using digital images showing cellulitis or red eyes, owing to the use of the proposed scales. (The null hypothesis of which is: There are no significant changes between the variables)*

In order to test the validity of hypothesis number 5, it was necessary to statistically compare all cellulitis experiments (1-5) of chapter seven which had been performed under two different conditions, i.e. with the use of a scale and without the use of a scale. In each case, the two conditions were analysed for their effects on the accuracy of, confidence level of, and the time taken by, the participants.

The effects of scale on time taken were clear. Cellulitis experiments 1-5 showed that the use of an image based scale increased the accuracy of classification supporting the previous literature by Kahn et al (1975) and Terry et al (1995), as well as increasing the time taken to classify the images. There was no statistical difference found between the confidence levels of the two groups. We can therefore reject the sections of the null hypothesis pertaining to accuracy and time taken.

Given the increase in accuracy and time taken it was reasonable to draw a parallel between these two factors. There were two possible reasons why this should have been the case. The presence of a scale would involve more time being spent, as it created an additional step in the evaluation process, but the scale in itself may have been the sole contributor to the increase in accuracy. The other possibility, however, may have been that the additional step, represented by the scale, may have caused a change in the thought processes of the participants. As discussed by Kahneman (2011), simple comparison of two images may have involved only system one thinking, whereas system two may have had to be engaged with the introduction of the additional information, in the form of the scale, and thus enhanced the accuracy of the results. While there was no difference in the confidence levels after implementation of a 5 point scale, a 3 point scale was found to be associated with and increase in confidence of the participants, (Z = -5.029, p = 0.001).

It is clear that the implementation of a scale is beneficial in maximising classification accuracy and dividing the scale to an appropriate number of reference levels, 3 in this case, is an effective way to improve participant confidence.

Hypothesis 5: The difference between machine learning and human models in colour classification accuracy *There is difference between the accuracy of a machine learning model compared to a human system in colour classification. (The null hypothesis of which is; there is no significant difference between the two systems)*

The latter portion of chapter 7 explores this hypothesis and concludes that the described machine learning programme more accurately classifies the images than a human system, the results of the Wilcoxon test used for this analysis are presented in appendix H. As these findings were statistically significant for both pathologies (p=0.001) we can reject the null hypothesis stated above.

It is possible, however, that the imperfections of the current system open a window of opportunity for the development of a complex and sophisticated, computer based auto-evaluation system, which would be more objective in its assessments and consistent in its reporting. Given that it is possible for a computer programme to modify images to degrees indistinguishable to the human eye, then it should be within the powers of the technology companies to engineer a system, which would be capable of extracting and identifying similar subtleties. The elementary auto evaluation system created for the purposes of this thesis was tested using the same images as those presented in the experiments. The results were generally quite encouraging and comparable, in accuracy, to the answers of the human participants. It must be stressed, once again, that this was a simple model and, as such, suffered from similar problems to those observed particularly in non-doctors answers, in that its inability to extract areas of interest resulted in the evaluations being skewed by contextual colour. This system of evaluation relied entirely on comparison of an image to those held in the training folder, the number of which was limited by the overall number available from the experiment database. The greater the number of training images available, the more accurate will be the system. For this reason, it is essential that a co-operative association is formed with health services and other interested parties in order to maximise the number of training images available. A potential limitation of this system would be the infinite variety of eye shapes and colours associated with different ethnic groups, and the varying colours of contextual skin. For this reason, it is essential that advanced work should be carried out on extraction of the areas if interest in order to remove the variables which could lead to inaccurate evaluation. In an ideal world, the perfect system would be for every individual to have reference images of their skin and eyes, in normal health, kept within their own medical records for direct comparison. This would obviously involve a huge effort on the part of international medical agencies, and this, together with the enormous amount of data storage required, may make such a target unachievable. It is probable, therefore, that an extensive training database would be the best compromise. Nonetheless, it appears that an auto evaluation system of this type, with considerable refinement by appropriate experts, has the potential to make a significant contribution to the advancement of telehealth.

8.2 Research Originality and Contributions

The work presented in this thesis resulted in the following significant contributions to knowledge in the area of telehealth and its associated research methods.

8.2.1 Research contribution 1: Introducing cellulitis and conjunctivitis to telehealth systems

The study investigates the potential application of SAF teleconsultation to the common medical conditions cellulitis and conjunctivitis. These two pathologies are not commonly used in telehealth system when the current study started.

8.2.2 Research contribution 2: Introducing and testing the use of colour scales in telehealth

Colour scales are commonly used in various areas of healthcare, such as ophthalmology. This study proposed their use in telehealth systems, more specifically in teleconsultation where they can be used as a common communication language or guide tool in classifying colours when using these to diagnose pathologies. Further potentially valuable areas of research suggested by this study would involve closer investigation and confirmation of the results uncovered by the experimental work reported in this thesis.

8.2.3 Research contribution 3: Introducing and testing the use of image quality scales in telehealth

The study proposed the use of an image quality scale showing standardised samples of digital images at different resolutions, for use as a guide when assessing the suitability of an image received for diagnosis.

8.2.4 Research contribution 4: Introducing and testing confidence scales in telehealth

The study introduced a confidence scale of 0 to 10 as a comparative measurement tool to help analyse the level of accuracy that humans have when perceiving colour. The potential influences of overconfidence or lack of confidence were taken into account.

8.2.5 Research contribution 5: Combined design for measuring colour perception

The combination in the experimental design in this study linked three related procedures. The first procedure was to measure image quality using a scale of 0-9, representing high, medium and low quality. The second was to use a colour intensity scale as a guide when performing tasks that involved ranking and differentiating colours. The third procedure was to measure the confidence level of participants by using a confidence scale as a tool to indicate the accuracy of performance. The combination of these three elements represented a new approach to testing colour perception in a telehealth context, as well as in evaluating image quality more generally. Similar concepts have, however, already been applied in other contexts, as was reported in the literature review.

8.2.6 Research contribution 6: The involvement of non-healthcare professionals in telehealth

The experiments in this study were about colour evaluation and not diagnosis. The evidence indicates that colour perception appears to be generic in its mechanism. The accuracy of the performance of individuals was generally consistent across all the participants in the experiments. This was particularly interesting because it suggested that non-doctors may be able to work in telehealth roles previously filled only by qualified doctors. This could include communication with patients, receiving and rating image quality and categorising colour intensity. This could constitute a considerable saving in resources and costs. The final diagnoses would necessarily remain the job of doctors, due to their expertise, as well as legal and ethical considerations, but it should be possible for persons qualified to a lower level to be involved in the inspection and classification of submitted images, referring the necessary cases to a medical doctor for diagnosis and treatment.

8.2.7 Research contribution 7: A guide for an auto-evaluation system

The study proposed a set of requirements specifications for an auto-evaluation system for early stage classification of redness in images of pathologies. The concept was that considerable savings, in terms of time and money, could be made if redness levels could be identified and prioritised by an automated system prior to referral of appropriate cases to doctors. A simplistic, single algorithm based model was tested alongside the experimental results.

8.3 Limitations and Reflections

The multidisciplinary nature and originality of the research made the integration of subjects such as cognitive engineering, human-computer interaction, information technology, human colour perception, telehealth, medicine, image processing, and colour science, a complex and a time-consuming process. The literature review needed continuous reviews and amendments in order to link and integrate all the different sections together. However, such a developed literature review is a potentially valuable asset for future studies in this area of research. Let this word be the same with the rest paragraph

The complexity of the knowledge gap, research hypotheses and problem definition made the planning, design and timing of the project very challenging. This, together with the continual need to review and redesign the experiments, provided excellent experience in this area of work.

The examination of human perception is a challenge as there is no direct or straightforward way to measure and quantify it. However, previous studies were consulted, and knowledge gaps identified, in order to formulate and test the experimental design whereby participants could perform tasks involving colours. Such tasks included describing, grouping, ranking and rating colours in digital images. The study used the accuracy level of the answers as an indication of the colour perception of the participants and their levels of confidence as a further indication of the accuracy of result for the tasks undertaken. The experimental design of the research proceeded through several developmental stages. The working model was first tested in a set of pilot experiments, and, after necessary modification of the methodology, work then concentrated upon experiments that used actual medical data. These developmental stages of the experimental work reflected learning development occurring during the research. This development fostered a deep understanding of the topic, but made comparisons among the data sets of different experiments complex to analyse.

The cooperation of the Scottish Centre for Telehealth, now part of NHS24, was vital to the success of the study due to their providing the required medical data and images. This simplified the data acquisition process, which was a core part of the study. However, dealing with NHS patients and doctors requires ethical approval, which is a long and detailed procedure. The process was facilitated by the direct involvement and support of the medical team of the A&E department of the ARI, who collected data and administrated the ethical approval procedures. The sensitive nature of the data, particularly in the second stage of the experimental work that involved conjunctivitis, required lengthy applications processes, interviews, meetings and training to seek ethical approval. Securing such approval and meeting all legal requirements occupied a much greater part of the project time than was expected. However, the study designed an alternative plan for obtaining high quality images from different sources and had the images verified for use in this study, by two experts in the area of conjunctivitis.

The field of telehealth is developing rapidly in parallel with advances in telecommunications technologies, and is becoming more important in its contribution to healthcare. Telehealth is important, and indeed is sometimes urgently needed, for those who live in remote geographical areas or who work aboard ships or offshore production platforms. Telehealth can save lives in such situations as well as reduce costs and save time. The progress and developmental processes have become fast and the medical pressure in the usage of the technology has become challenging, in order to meet the requirements of the system and facilitate the medical activities, which can be critical for patient health and safety.

Many people still do not feel comfortable using telehealth systems because they are used to having personal contact with their doctor. They feel more assured with a traditional consultation model rather than with using machines. However, telehealth can sometimes be the only practical alternative to meet medical challenges in emergencies such as natural disasters, when it is typically difficult for medical teams to be on the ground. More generally, General practitioners (GPs) in remote areas might need the support of specialist consultants to diagnose or make decisions about specific pathologies and telemedicine may be the only practical solution in such cases. Much work remains to be done to assure both medical workers and patients that telehealth systems provide a safe and reliable alternative to traditional consultation methods. It is hoped that this thesis may go some way to increasing the understanding of human perception of colour and thus contribute to the future development and expansion of existing telemedicine operations. According to the available resources, no one has previously done any work of this nature in considering the integration between different subjects and using the same methodology covered by this study.

8.4 Recommendations for further work

8.4.1 Proposed auto-evaluation system for using digital images in telehealth

Whilst a simple model of such a system was generated and tested as part of this thesis, it is acknowledged that a true operational system would need to be more complex in order to come even close to the standard required for practical purposes. It is not intended that a computer based evaluation system should ever replace traditional diagnosis by a doctor, but there is little doubt that, given the increasing demands on doctors? time and healthcare budgets, any contribution from an auto-evaluation system would be welcomed. A huge amount of development would need to be carried to extract the areas of interest from images, and it is probable that a combined algorithm of complexity would be required in order to create a viable working system.

8.4.2 Mobile phone cameras experiment

A useful contribution could be made by research into the applicability of images captured using mobile phones for telehealth diagnostic purposes. Such a study should have a particular focus on the colour red as this colour has characteristics which may make its electronic transmission and representation problematic. As a part of this research, consideration should also be given to establishing the minimum image quality level which is still compatible with accurate diagnoses.

It is hoped that by improving the diagnostic accuracy of digital images, waiting lists will be reduced (hence reducing healthcare costs) and enabling better time-management for both healthcare professionals and patients. This project should aim to analyse images taken by mobile phones, and more specifically the clarity of their representation of the colour red in areas of the eye that would normally be white in a healthy specimen. The colour representations influence on human perception would be a central aim of such a study as perception and diagnostic accuracy are undoubtedly interlinked. This project would assess the ability of healthcare professionals in; describing, grouping, and rating the colour red. Different image capture devices vary in technical specifications and produce images of varying qualities. The suggested project, therefore, should study the impact on diagnostic applicability of images captured by a number of different mobile phones. It is suggested that comparison should be made using, and not using, the colour scale guide that is currently in use at ARI.

Images should only show the patients' eyes, but not the whole face. A total of at least 30 patients should be recruited to the study. Participation should be voluntary and should involve informed consent. Images should be anonymous but associated with a confirmed diagnosis. The images could then form a data set that would be used in an experimental study asking healthcare professionals to describe the characteristics of redness. The verbal descriptions would be recorded and used for qualitative analysis. Healthcare professionals would also be asked to rate, group and rank the images using scales indicating perceived image quality such as ?redness?. Participants would also be asked to rate their own confidence in their decisions. The results should be statistically analysed to ascertain the minimum level of quality, and the nature of that quality, needed for accurate diagnosis.

The researcher of this study obtained ethical approval to conduct such an experiment. The approval included the capturing of digital images by doctors in the eye unit of ARI using mobile phones. This experiment might form part of a postdoctoral research project. All the documents of the proposed project are available for perusal via the Dropbox link at http://bit.ly/1l5WyTy.

The following are the main research questions that are proposed.

- **Research Question 1-Colour Characteristics**
  What are the required characteristics and qualities of the colour red, in medical images, that permit accurate diagnosis using telehealth techniques?

- **Research Question 2-Minimum Image Quality**
  What is the minimum image quality required in a mobile phone image to ensure an accurate diagnosis when used in telehealth?

- **Research Question 3-Colour Description**
  How do health workers describe the colour red in medical images used in telehealth?

- **Research Question 4-Impact of degradation**
  What is the impact of degrading the mobile phone images captured by healthcare professionals, on the accuracy level of diagnoses based on these images, in comparison with diagnoses derived using normal face to face diagnostic protocols?

- **Research Question 5-Impact of using colour scales**
  What is the difference in the diagnostic applicability of mobile telephone images showing redness in the front part of the eye (conjunctivitis) when using or not using the colour scale that is already used in ARI during normal face to face consultations.

8.4.3 Image eye tracking

There is a need for a study with the aim of analysing the eye movements of doctors when diagnosing digital images. The study should use eye-tracking technologies to record the number of saccades (unordered random rapid movements of the eyes within points of a scene or image) and fixations (stopping and focusing on particular points) made by doctors during a diagnosis. Tracking these eye movements would indicate the significant areas in the images. These areas of significant interest, it is assumed, would map to the most affected areas. This information might then be applicable to the recognition of symptoms when designing telehealth diagnostic systems.

8.4.4 Testing the results of this study in real use

It would also be useful to dedicate research to testing the results of this study in real time, on one of the existing telehealth systems such as the one employed in ARI. This evaluation study would be essential for decision makers in the field of telehealth.

8.5 Conclusion

Telehealth utilises modern technology to augment the distribution and scope of healthcare. This quickly adapting advancement has already seen widespread use providing healthcare to those without access, aiding in education programmes disseminating learning materials with noted economic benefits allowing information to be transported digitally. There is still much untapped potential for telehealth such its role in automation as well as the implementation of ubiquitous mobile phones. The current study explored the use of digital images of two pathologies, cellulitis and conjunctivitis, as a diagnostic and assessment medium assessing accuracy, confidence and time evaluating also the effect of numeric and image based scales as well as medical experience. It was found that digital images were indeed a viable medium by which to diagnose these conditions allowing participants to accurately identify the conditions based on the images alone. Implementation of a digital system may allow patients in rural or impoverished areas to gain an accurate diagnosis without the need to travel while also reducing the cost burden on both the patients and the health service. Additionally this highlights the potential of this system to be used in other conditions where colour is a key diagnostic feature.

Medical experience was found, in most cases, have no effect on the accuracy of colour classification indicating the potential for those with limited medical training to provide accurate colour judgment. Therefore it may be possible that individuals with limited medical knowledge may assist in the diagnostic process in situations where there is a shortage of medical professional. This would also provide economic benefits while easing the patient burden on medical staff. Numeric and image based scales with varying divisions were both found to increase diagnostic accuracy. Of the two, an image based scale was found to be most beneficial the optimal division level of which is one with three tiers as it maximises accuracy as well as confidence. It was also found that the use of a standardised image classification system such as that used in the conjunctivitis experiments resulted in higher diagnostic accuracy. The results here recommend the use of a low division scale using a standardised set of images for digital classification of pathology. Finally the system was reanalysed using an automated classification system developed with a machine learning programme. The ML system was found to be significantly more accurate than a human system in diagnosing and classifying the pathologies based on the digital images. This finding shows that this process can be automated allowing for quicker, more efficient and cost effective diagnosis and classification pathologies. This may be especially useful for quick, accurate sorting of hundreds or thousands of images which would impose a temporal and economic burden on a human.

Chapter 9

References

ABELSON, M.B., 2010. Code red: The key features of hyperemia. [online] Newtown Square, PA 19073, USA: Jobson Medical Information LLC. Available from: http://www.revophth.com/content/d/therapeutic_topics/i/1205/c/22716/ [Accessed 9/7/2013 2013]

ABRAHAMSSON, P. et al., 2002.Agile Software Development Methods: Review and Analysis (VTT publications).

ABRAMOFF, M.D. and SUTTORP-SCHULTEN, M.S., 2005. Web-based screening for diabetic retinopathy in a primary care population: the EyeCheck project. Telemedicine Journal & e-Health, 11(6), pp. 668-674

ABRAMOV, I. et al., 2012. Sex and vision II: color appearance of monochromatic lights. Biology of sex differences, 3(1), pp. 1-15

ADAMS, J.K., 1957. A confidence scale defined in terms of expected percentages. The American Journal of Psychology, 70(3), pp. 432-436

A guide to understanding colour communication. 2007. [online] x-rite. Available from: http://www.xrite.com/documents/literature/en/L10-001_Understand_Color_en.pdf [Accessed 9/7/2013 2013]

AFLAKI, S. AND MEMARZADEH, M., 2011. Interpreting SuperPAVE PG test results with confidence intervals.Construction and Building Materials,25(6), pp.2777-2784.

AHMED, F. et al., 2012. Classification of crops and weeds from digital images: A support vector machine approach. Crop Protection, 40, pp. 98-104

Air products healthcare report [Online] Available from: WWW.airproducts.co.uk/homecare [Accessed on 15/ 06/ 2013].

ALEXANDER, T., 1995. [BOOK REVIEW] THE PRIMARY COLORS, THREE ESSAYS. New Statesman and Society, 8(380), pp. 46

AL-HINDAWE, J., 1996. Considerations when constructing a semantic differential scale. La Trobe working papers in linguistics, 9

Allergan grading guide for bulbar conjunctiva hyperaemia: eye and speciality pharmaceutical products United States of America [Online] Available from http://www.allergan.com, [Accessed on 8/9/2013]

AL-RAWI, M., QUTAISHAT, M. and ARRAR, M., 2007. An improved matched filter for blood vessel detection of digital retinal images. Computers in biology and medicine, 37(2), pp. 262-267

AMES, D.R. et al., 2010. Not so fast: The (not-quite-complete) dissociation between accuracy and confidence in thin-slice impressions. Personality and Social Psychology Bulletin, 36(2), pp. 264-277

ANAND, H., MIR, R. and SAXENA, R., 2009. Hemoglobin color scale a diagnostic dilemma.

BERGMAN, R.A.,2015, Anatomy atlases: Anatomy of first aid: A case study approach: The eye [online] Available from: http://www.anatomyatlases.org/firstaid/Eye.shtml [Accessed 31/10/ 2015]

ARMSTRONG, A.W. et al., 2012a. Teledermatology Operational Considerations, Challenges, and Benefits: The Referring Providers' Perspective. Telemedicine and e-Health, 18(8), pp. 580-584

ARMSTRONG, A.W. et al., 2012b. State of teledermatology programs in the United States. Journal of the American Academy of Dermatology, 67(5), pp. 939-944

ARNSTEIN, F., 1997. Catalogue of human error. British journal of anaesthesia, 79(5), pp. 645-656
ATKINSON, R.C. and SHIFFRIN, R.M., 1968. Human memory: A proposed system and its control processes. The psychology of learning and motivation, 2, pp. 89-195

BADDOUR, L.M., 2000. Cellulitis syndromes: an update. International journal of antimicrobial agents, 14(2), pp. 113-116

BERETTA, G. and MORONEY, N., 2008. Cognitive aspects of color. Hewlett-Packard Laboratories Technical Report, 109

BERGMAN, H. et al., 2009. Remote assessment of acne: the use of acne grading tools to evaluate digital skin images. TELEMEDICINE and e-HEALTH, 15(5), pp. 426-430

BERNDT, R. et al., 2012. Development of a mobile teledermatology system. Telemedicine and e-Health, 18(9), pp. 668-673

BIEMANS, M., SWAAK, J., HETTINGA, M. and SCHUURMAN, J.G., 2005. Involvement matters: the proper involvement of users and behavioural theories in the design of a medical teleconferencing application. Proceedings of the 2005 international ACM SIGGROUP conference on Supporting group work. ACM. pp. 304-312

BISANTZ, A.M., 2008. Cognitive engineering applications in health care. The Bridge, , pp. 39-47

BISANTZ, A.M. and BURNS, C.M., 2009. Applications of cognitive work analysis. Taylor & Francis US.

BISTA, S.R., SRNDI, I., DOGAN, S., ASTVATSATOUROV, A., MSGES, R. and DESERNO, T.M., 2013. Automatic conjunctival provocation test combining Hough circle transform and self-calibrated color measurements. SPIE Medical Imaging. International Society for Optics and Photonics. pp. 86702J-86702J-10

BITTORF, A. et al., 1997. Resolution requirements for digital images in dermatology. Journal of the American Academy of Dermatology, 37(2), pp. 195-198

BLOJ, M. and HEDRICH, M., 2012. Color Perception. Handbook of Visual Display Technology. Springer. pp. 171-178

BLUM, A., ZALAUDEK, I. and ARGENZIANO, G., 2008. Digital image analysis for diagnosis of skin tumors. Seminars in cutaneous medicine and surgery. Elsevier. pp. 11-15

BORNSTEIN, B.H. and ZICKAFOOSE, D.J., 1999. " I know I know it, I know I saw it": The stability of the confidence?accuracy relationship across domains. Journal of Experimental Psychology: Applied, 5(1), pp. 76

BREAR, M., 2006. Evaluating telemedicine: lessons and challenges. Health Information Management Journal, 35(2), pp. 23-31

BRENNER, E., GRANZIER, J.J. and SMEETS, J.B., 2007. Perceiving colour at a glimpse: The relevance of where one fixates. Vision research, 47(19), pp. 2557-2568

BREWER, N., KEAST, A. and RISHWORTH, A., 2002. The confidence-accuracy relationship in eyewitness identification: The effects of reflection and disconfirmation on correlation and calibration. Journal of Experimental Psychology: Applied, 8(1), pp. 44

BRIGNELL, M., WOOTTON, R. and GRAY, L., 2007. The application of telemedicine to geriatric medicine. Age and Ageing, 36(4), pp. 369-374

BRILAKIS, I. and SOIBELMAN, L., 2005. Content-based search engines for construction image databases. Automation in Construction, 14(4), pp. 537-550

BUCKLEY, K.M., ADELSON, L.K. and AGAZIO, J.G., 2009. Reducing the risks of wound consultation: adding digital images to verbal reports. Journal of Wound Ostomy & Continence Nursing, 36(2), pp. 163-170

BURNINGHAM, N., PIZLO, Z. and ALLEBACH, J.P., 2002. Image quality metrics. Encyclopedia of imaging science and technology,

BYRNE, A., 2011. Error in clinical measurement. Anaesthesia & Intensive Care Medicine, 12(12), pp. 578-580

CACERES, C. et al., 2006. An integral care telemedicine system for HIV/AIDS patients. International journal of medical informatics, 75(9), pp. 638-642

CAMPOS, C. et al., 2012. Setting up a telemedicine service for remote real-time video-EEG consultation in La Rioja (Spain). International journal of medical informatics, 81(6), pp. 404-414

CAVALCANTI, P.G. and SCHARCANSKI, J., 2011. Automated prescreening of pigmented skin lesions using standard cameras. Computerized Medical Imaging and Graphics, 35(6), pp. 481-491

CHANDLER, D.M., 2013. Seven challenges in image quality assessment: past, present, and future research. ISRN Signal Processing, 2013

CHANG, C.C., CHUANG, J.C. AND HU, Y.S., 2004. Retrieving digital images from a JPEG compressed image database. Image and Vision Computing, 22(6), pp.471-484.

CHAUDHARI, D.K. and SHAW, J.S., 2012. Pathophysiology of altered color Perception.

CHENG, Y. and CHEN, S., 2003. Image classification using color, texture and regions. Image and Vision Computing, 21(9), pp. 759-776

CHODOSH, J., CHINTAKUNTLAWAR, A.V. and ROBINSON, C.M., 2008. Human Eye Infections. In: B. MAHY and M. VAN REGENMORTEL, eds. Encyclopedia of Virology. 3rd ed. London: Academic Press. pp. 491-497

CHOONG, M.K., LOGESWARAN, R. and BISTER, M., 2007. Cost-effective handling of digital medical images in the telemedicine environment. International journal of medical informatics, 76(9), pp. 646-654

CIOCCA, G., MARINI, F. and SCHETTINI, R., 2009. Image quality assessment in multimedia applications. IS&T/SPIE Electronic Imaging. International Society for Optics and Photonics. pp. 72550A-72550A-8

COLLIS, J. and HUSSEY, R., 2009. Business research: A practical guide for undergraduate and postgraduate students. Hampshire Palgrave Macmillan.

COLVEN, R. et al., 2011. Dermatological diagnostic acumen improves with use of a simple telemedicine system for underserved areas of South Africa. Telemedicine and e-Health, 17(5), pp. 363-369

CORSO, J.J. and HAGER, G.D., 2009. Image description with features that summarize. Computer Vision and Image Understanding, 113(4), pp. 446-458

CRAIG, J. and PATTERSON, V., 2005. Introduction to the practice of telemedicine. Journal of telemedicine and telecare, 11(1), pp. 3-9

CRITCHLEY, J. and BATES, I., 2005. Haemoglobin colour scale for anaemia diagnosis where there is no laboratory: A systematic review. International journal of epidemiology, 34(6), pp. 1425-1434

CURRELL, R. et al., 2000. Telemedicine versus face to face patient care: effects on professional practice and health care outcomes. Cochrane Database Syst Rev, 2(2),

D'AVERSA, G. et al., 1995. Diagnosis and treatment of eye infections. Primary care update for Ob/Gyns, 2(4), pp. 138-142

DAVIES, I.R. et al., 1998. Cross-cultural differences in colour vision: Acquired ?colour-blindness? in Africa. Personality and Individual Differences, 25(6), pp. 1153-1162

DAWSON, C.W., 2005. Projects in computing and information systems: a student's guide. Pearson Education.

DE BORTOLI, M. AND MAROTO, J., 2001. Colours across cultures: Translating colours in interactive marketing communications. In Elicit 2001: Proceedings of the European Languages and the Implementation of Communication and Information Technologies (Elicit) conference (pp. 3-4). UK: Paisley University Language Press.

DEKKER, S., 2006. The field guide to human error. Bedford, UK: Cranfield University Press.

DEMPSEY, S.E. and BURR, M., 2009. The level of confidence and responsibility accepted by Australian radiation therapists in developing plans and implementing treatment. Radiography, 15(2), pp. 139-145

DESHPANDE, A. et al., 2009. Asynchronous Telehealth: a scoping review of analytic studies. Open Medicine, 3(2), pp. e69

DIXON, B.E., HOOK, J.M. AND MCGOWAN, J.J., 2008. Using telehealth to improve quality and safety: Finding from the AHRQ Portfolio (Prepared by the AHRQ National Resource Center for Health IT under contract No. 290-04-0016).

DHEER, A. and CHATURVEDI, R., 2005. Embracing a revolution?Telemedicine. Medical Journal Armed Forces India, 61(1), pp. 51-56

DOMINGOS, P., 2012. A few useful things to know about machine learning.Communications of the ACM,55(10), 78-87.

DRAGAN, D. and IVETI?, D., 2012. Request redirection paradigm in medical image archive implementation. Computer methods and programs in biomedicine, 107(2), pp. 111-121

EDISON, K.E. et al., 2008. Diagnosis, diagnostic confidence, and management concordance in live-interactive and store-and-forward teledermatology compared to in-person examination. TELEMEDICINE and e-HEALTH, 14(9), pp. 889-895

EEDY, D. and WOOTTON, R., 2001. Teledermatology: a review. British Journal of Dermatology, 144(4), pp. 696-707

EIKELBOOM, R.H. et al., 2000. Methods and limits of digital image compression of retinal images for telemedicine. Investigative ophthalmology & visual science, 41(7), pp. 1916-1924

ENGELDRUM, P.G., 2004. A theory of image quality: The image quality circle. Journal of Imaging Science and Technology, 48(5), pp. 447-457

ENGELKE, U., MAEDER, A. and ZEPERNICK, H., 2012. Human observer confidence in image quality assessment. Signal Processing: Image Communication,

ENGELKE, U., MAEDER, A. and ZEPERNICK, H., 2009. On confidence and response times of human observers in subjective image quality assessment. Multimedia and Expo, 2009. ICME 2009. IEEE International Conference on. IEEE. pp. 910-913

ENOMOTO, Y., BURNS, C.M., MOMTAHAN, K. and CAVES, W., 2006. Effects of Visualization Tools on Cardiac Telephone Consultation Processes. Proceedings of the Human Factors and Ergonomics Society Annual Meeting. SAGE Publications. pp. 1044-1048

FABBROCINI, G. et al., 2011. Teledermatology: from prevention to diagnosis of nonmelanoma and melanoma skin cancer. International journal of telemedicine and applications, 2011, pp. 1

FEI, X. et al., 2012. Perceptual image quality assessment based on structural similarity and visual masking. Signal Processing: Image Communication, 27(7), pp. 772-783

FIEGUTH, P. and SIMPSON, T., 2002. Automated measurement of bulbar redness. Investigative ophthalmology & visual science, 43(2), pp. 340-347

FIGTREE, M. et al., 2010. Risk stratification and outcome of cellulitis admitted to hospital. Journal of Infection, 60(6), pp. 431-439

FLAVELL, R. and HEATH, A., 1992. Further investigations into the use of colour coding scales. Interacting with Computers, 4(2), pp. 179-199

FLEISHER, G. et al., 1981. Cellulitis: initial management. Annals of Emergency Medicine, 10(7), pp. 356-359

FLEISHER, G. and LUDWIG, S., 1980. Cellulitis: a prospective study. Annals of Emergency Medicine, 9(5), pp. 246-249

FOLEY, J.D., 1996. Computer graphics: Principles and practice, in C. Addison-Wesley Professional.

FOLEY, J.D. and VAN DAM, A., 1982. Fundamentals of interactive computer graphics. Addison-Wesley Systems Programming Series, Reading, Mass.: Addison-Wesley, 1982, 1

GAGNON, M.P. et al., 2012. Using a modified technology acceptance model to evaluate healthcare professionals' adoption of a new telemonitoring system. Telemedicine and e-Health, 18(1), pp. 54-59

GANESHALINGAM, A. et al., 2010. Effectiveness of asynchronous tele-endoscopy. Gastrointestinal endoscopy, 71(3), pp. 461-467. e2

GERSH, J.R., MCKNEELY, J.A. and REMINGTON, R.W., 2005. Cognitive engineering: Understanding human interaction with complex systems. Johns Hopkins APL Technical Digest, 26(4), pp. 377-382

GILCHRIST, A. and NOBBS, J.H., 2001. Colour Science: proceedings of the International Conference & Exhibition, Harrogate, April 1-3, 1998. Colour physics. University of Leeds, Department of Colour Chemistry.

GMEZ, E.J. et al., 2001. Optimisation and evaluation of an asynchronous transfer mode teleradiology co-operative system: the experience of the EMERALD and the BONAPARTE projects. Computer methods and programs in biomedicine, 64(3), pp. 201-214

GMEZ, E.J. et al., 2002. Telemedicine as a tool for intensive management of diabetes: the DIABTel experience. Computer methods and programs in biomedicine, 69(2), pp. 163-177

GONZALEZ, R.C. AND WOODS, R.E., 2008. Digital image processing: Pearson prentice hall. Upper Saddle River, NJ.

GREEN, A. and MARTIN, N.G., 1990. Measurement and perception of skin colour in a skin cancer survey. British Journal of Dermatology, 123(1), pp. 77-84

GRUBAUGH, A.L. et al., 2008. Attitudes toward medical and mental health care delivered via Telehealth applications among rural and urban primary care patients. The Journal of nervous and mental disease, 196(2), pp. 166-170

GUARNERI, F., VACCARO, M. and GUARNERI, C., 2008. Digital image compression in dermatology: Format comparison. Telemedicine and e-Health, 14(7), pp. 666-670

Guidelines on the Management of Cellulitis in adults, CREST (Clinical Resource Efficiency Support Team, Northern Ireland) [online] Available from:http://www.crestni.org.uk/publications/cellulitis-guide.pdf, June 2005[Accessed on 15/ 09/ 2013].

GUNDERSON, C.G., 2011. Cellulitis: definition, etiology, and clinical features. The American Journal of Medicine, 124(12), pp. 1113-1122

HAEGHEN, Y.V. et al., 2000. An imaging system with calibrated color image acquisition for use in dermatology. Medical Imaging, IEEE Transactions on, 19(7), pp. 722-730

HALL, K., 2011. Telemedicine in the NHS: The benefits and costs of implementing telecare services. Computer Weekly, 5

HARTSON, H.R. and HIX, D., 1989. Human-computer interface development: concepts and systems for its management. ACM Computing Surveys (CSUR), 21(1), pp. 5-92

HAWKINS, J., 1991. Definition of confidence. The Oxford Encyclopedic English Dictionary. Oxford University Press, USA.

HAYES, N., 2000. Doing psychological research. Taylor & Francis Group.

HEDRICK, J., 2003. Acute bacterial skin infections in pediatric medicine: current issues in presentation and treatment. Pediatric Drugs, 5(Supplement 1), pp. 35-46

HEISE, D.R., 1970. The semantic differential and attitude research. Attitude measurement, , pp. 235-253

HENNON, V.A., 1910. Sex differences and variability in color perception. University of Colorado Studies, (7), pp. 207-214

HERSH, W.R. et al., 2006. Diagnosis, access and outcomes: Update of a systematic review of telemedicine services. Journal of telemedicine and telecare, 12(suppl 2), pp. 3-31

HIGHSMITH, J.A., 2002. Agile software development ecosystems (Vol. 13). Addison-Wesley Professional.

HINGORANI, M. et al., 2006. Conjunctivitis.

HSUEH, M.T. et al., 2012. Teledermatology patient satisfaction in the pacific northwest. Telemedicine and e-Health, 18(5), pp. 377-381

HUANG, J., KUMAR, S.R. and ZABIH, R., 2003. Automatic hierarchical color image classification. EURASIP Journal on Applied Signal Processing, 2003, pp. 151-159

Ishihara plates for colour blindness test online. [online] Available from: http://colourblind.freeservers.com[Accessed on 7th of Feb.2010]

JACOBS, D.E., GOLDMAN, D.B. and SHECHTMAN, E., 2010. Cosaliency: Where people look when comparing images. Proceedings of the 23nd annual ACM symposium on User interface software and technology. ACM. pp. 219-228

JAIN, N. et al., 2010. Gender based alteration in color perception.

JAMESON, K.A., HIGHNOTE, S.M. and WASSERMAN, L.M., 2001. Richer color experience in observers with multiple photopigment opsin genes. Psychonomic bulletin & review, 8(2), pp. 244-261

PREECE, J., SHARP, H., & ROGERS, Y. (2015). Interaction Design-beyond human-computer interaction. John Wiley & Sons.

PREECE, J. et al., 1994. Human-computer interaction. Addison-Wesley Longman Ltd.

JIJI, G.W., 2011. Colour texture classification for human tissue images. Applied Soft Computing, 11(2), pp. 1623-1630

JOSEPH, A., 2004. Trauma and common infections of the eye. Surgery (Oxford), 22(8), pp. 191-193

JOSEPH, V. et al., 2011. Key challenges in the development and implementation of Telehealth projects. Journal of telemedicine and telecare, 17(2), pp. 71-77

KAHNEMAN, D., 2011.Thinking, fast and slow. Macmillan.

KAHNEMAN, D., and TVERSKY, A. (1996). On the reality of cognitive illusions.

KAHNEMAN, D. AND KLEIN, G., 2009. Conditions for intuitive expertise: a failure to disagree. American psychologist, 64(6), p.515.

KHAN HA, Leibowitz H, Ganley JP, Kini M, Colton T, Nickerson R, Dawber TR. 1975 Randomized controlled clinical trial. National Eye Institute workshop for ophthalmologists. Standardizing diagnostic procedures. Am J Ophthalmol 1975;79:768?75.

KANTHRAJ, G., 2009. Classification and design of teledermatology practice: What dermatoses? Which technology to apply? Journal of the European Academy of Dermatology and Venereology, 23(8), pp. 865-875

KAPLAN, B. and LITEWKA, S., 2008. Ethical challenges of telemedicine and Telehealth. Cambridge Quarterly of Healthcare Ethics, 17(04), pp. 401-416

KARATZAS, D. and ANTONACOPOULOS, A., 2007. Colour text segmentation in web images based on human perception. Image and Vision Computing, 25(5), pp. 564-577

KEBBELL, M.R., EVANS, L. and JOHNSON, S.D., 2010. The influence of lawyers' questions on witness accuracy, confidence, and reaction times and on mock jurors' interpretation of witness accuracy. Journal of Investigative Psychology and Offender Profiling, 7(3), pp. 262-272

KEEN, N., 2005. Color Moments. Instructor, 501(0341091),

KENET, R.D., 1995. Digital imaging in dermatology. Clinics in dermatology, 13(4), pp. 381-392

KEPECS, A. and MAINEN, Z.F., 2012. A computational framework for the study of confidence in humans and animals. Philosophical Transactions of the Royal Society B: Biological Sciences, 367(1594), pp. 1322-1337

KIM, H.M. et al., 2003. Accuracy of a web-based system for monitoring chronic wounds. Telemedicine Journal and e-Health, 9(2), pp. 129-140

KING, T.D., 2005. Human color perception, cognition, and culture: why red is always red. Electronic Imaging 2005. International Society for Optics and Photonics. pp. 234-242

KLAYMAN, J. et al., 1999. Overconfidence: It depends on how, what, and whom you ask. Organizational behavior and human decision processes, 79(3), pp. 216-247

KLUGE, E.W., 2011. Ethical and legal challenges for health telematics in a global world: Telehealth and the technological imperative. International journal of medical informatics, 80(2), pp. e1-e5

KOTSIANTIS, S. B., ZAHARAKIS, I. and PINTELAS, P., 2007. Supervised machine learning: A review of classification techniques.

KRACHMER, J., MANNIS, M. and HOLLAND, E., 2011. CORNEA. Third ed. China: Elsevier.

KREPS, G.L. and NEUHAUSER, L., 2010. New directions in eHealth communication: Opportunities and challenges. Patient education and counseling, 78(3), pp. 329-336

KROSHINSKY, D., GROSSMAN, M.E. and FOX, L.P., 2007. Approach to the patient with presumed cellulitis. Seminars in cutaneous medicine and surgery. Elsevier. pp. 168-178

KRUG, K., 2007. The relationship between confidence and accuracy: Current thoughts of the literature and a new area of research. Applied Psychology in Criminal Justice, 3(1), pp. 7-41

KRUPINSKI, E.A. et al., 1999. Diagnostic accuracy and image quality using a digital camera for teledermatology. Telemedicine Journal, 5(3), pp. 257-263

KRUPINSKI, E.A., SILVERSTEIN, L.D., HASHMI, S.F., GRAHAM, A.R., WEINSTEIN, R.S. and ROEHRIG, H., 2012. Influence of LCD color reproduction accuracy on observer performance using virtual pathology slides. SPIE Medical Imaging. International Society for Optics and Photonics. pp. 83180P-83180P-6

KRUPINSKI, E. et al., 2008. American telemedicine association?s practice guidelines for teledermatology. Telemedicine and e-Health, 14(3), pp. 289-302

KUEHNI, R.G., 2001. Determination of unique hues using Munsell color chips. Color Research & Application, 26(1), pp. 61-66

KUNTALP, M. and AKAR, O., 2004. A simple and low-cost Internet-based teleconsultation system that could effectively solve the health care access problems in underserved areas of developing countries. Computer methods and programs in biomedicine, 75(2), pp. 117-126

LANDRIGAN, C.P. and FRIEDMAN, J., 2007. PatientSafety and Medical Errors.

LASIERRA, N. et al., 2012. Lessons learned after a three-year store and forward teledermatology experience using internet: Strengths and limitations. International journal of medical informatics, 81(5), pp. 332-343

LATINO, R.J., 2007. Defining and reducing human error. Briefings on Patient Safety, pp.6-7.

LAWLER, E.K., HEDGE, A. and PAVLOVIC-VESELINOVIC, S., 2011. Cognitive ergonomics, socio-technical systems, and the impact of healthcare information technologies. International Journal of Industrial Ergonomics, 41(4), pp. 336-344

LEE, S. et al., 2000. Telemedicine: challenges and opportunities. Journal of High Speed Networks, 9(1), pp. 15-30

LEE, S. et al., 2011. The Relations between Time, Confidence and Accuracy: Using a Judgment Questionnaire. International Journal of Humanities and Social Science 1(15).

LEPISTO, L., LAUNIAINEN, A. and KUNTTU, I., 2009. Red eye detection using color and shape. Local and Non-Local Approximation in Image Processing, 2009. LNLA 2009. International Workshop on. IEEE. pp. 153-157

LI, H.K., 1999. Telemedicine and ophthalmology. Survey of ophthalmology, 44(1), pp. 61-72

LICHTENSTEIN, S. and FISCHHOFF, B., 1977. Do those who know more also know more about how much they know? Organizational behavior and human performance, 20(2), pp. 159-183

LIM, E.W., CELLER, B.G., BASILAKIS, J. and TAUBMAN, D., 2006. A Novel Image Capture System for Use in Telehealth Applications. Engineering in Medicine and Biology Society, 2006. EMBS'06. 28th Annual International Conference of the IEEE. IEEE. pp. 4743-4746

LIN, C., 2012. Mobile telemedicine: a survey study. Journal of medical systems, 36(2), pp. 511-520

LITTLE, K.B., 1961. Confidence and reliability. Educational and Psychological Measurement,

LIU, B.J. et al., 2007. International Internet2 connectivity and performance in medical imaging applications: Bridging the Americas to Asia. Journal of High Speed Networks, 16(1), pp. 5-20

LONG, F., YANG, Z. and PURVES, D., 2006. Spectral statistics in natural scenes predict hue, saturation, and brightness. Proceedings of the National Academy of Sciences, 103(15), pp. 6013-6018

LPEZ, C. et al., 2011. A telephone survey of patient satisfaction with realtime telemedicine in a rural community in Colombia. Journal of telemedicine and telecare, 17(2), pp. 83-87

LOPEZ, F.A. and LARTCHENKO, S., 2006. Skin and soft tissue infections. Infectious disease clinics of North America, 20(4), pp. 759-772.

LOTTO, R.B. and PURVES, D., 2004. Perceiving colour. Review of Progress in Coloration and Related Topics, 34(1), pp. 12-25

LU, D. and WENG, Q., 2007. A survey of image classification methods and techniques for improving classification performance. International journal of Remote sensing, 28(5), 823-870.

LYONS, K., 2004. The agile approach. Technical report, Conoco Phillips Australia Pty Ltd.MENP, T. and PIETIKINEN, M., 2004. Classification with color and texture: jointly or separately? Pattern Recognition, 37(8), pp. 1629-1640

MANN, D., 1993. The Relationship between Diagnostic Accuracy and Confidence in Medical Students.

MASSONE, C., BRUNASSO, A.M., CAMPBELL, T.M. and SOYER, H.P., 2009. Mobile teledermoscopy?melanoma diagnosis by one click? Seminars in cutaneous medicine and surgery. Elsevier. pp. 203-205

MASSONE, C. and SCHETTINI, A.P.M., 2012. Teledermatology. Leprosy. Springer. pp. 371-373

MATHER, G., 2009. Foundations of sensation and perception. Psychology Press.

MATVEEV, N.V. and KOBRINSKY, B.A., 2006. Automatic colour correction of digital skin images in teledermatology. Journal of telemedicine and telecare, 12(suppl 3), pp. 62-63

MAUSFELD, R. and HEYER, D., 2003. Colour Perception: Mind and the physical world. Oxford University Press Oxford.

MENP, T. AND PIETIKINEN, M., 2004. Classification with color and texture: jointly or separately?. Pattern recognition, 37(8), pp.1629-1640.

MCGUINNESS, D. & BRABYN, L., 1984. In pursuit of visuo-spatial ability part 1: Visual systems. Journal of Mental Imagery, (8), pp. 1-12

MCILVAINE, W.B., 2006. Human error and its impact on anesthesiology. Seminars in Anesthesia, Perioperative Medicine and Pain. Elsevier. pp. 172-179

MCMONNIES, C.W. and CHAPMAN-DAVIES, A., 1987. Assessment of conjunctival hyperemia in contact lens wearers. Part I. American Journal of Optometry and Physiological Optics, 64(4), pp. 246-250

MCNEILL, K. et al., 2002. ; Color Image in Medicine. Practical Methods of Color Quality Assurance for Telemedicine Systems. Journal Code: X0353A, 20(2), pp. 111-116

MCNIEL, D.E., SANDBERG, D.A. and BINDER, R.L., 1998. The relationship between confidence and accuracy in clinical assessment of psychiatric patients' potential for violence. Law and human behavior, 22(6), pp. 655-669

MENGELKAMP, C. and BANNERT, M., 2010. Accuracy of confidence judgments: Stability and generality in the learning process and predictive validity for learning outcome. Memory & cognition, 38(4), pp. 441-451

MILITELLO, L.G. et al., 2010. The role of cognitive systems engineering in the systems engineering design process. Systems Engineering, 13(3), pp. 261-273

MILLER, E.A., 2007. Solving the disjuncture between research and practice: Telehealth trends in the 21st century. Health Policy, 82(2), pp. 133-141

MILLIS, E., 1997. Contact lenses and the red eye. Contact Lens and Anterior Eye, 20, pp.S5-S10.

MISRA, S. C., KUMAR, V., and KUMAR, U., 2006. Success Factors of Agile Software Development. InSoftware Engineering Research and Practice(pp. 233-239).

MITCHELL, T. M. (1997). Does machine learning really work?.AI magazine,18(3), 11.

MITCHELL, T. M., 2006.The discipline of machine learning(Vol. 17). Carnegie Mellon University, School of Computer Science, Machine Learning Department.

MOORE, M., 1999. The evolution of telemedicine. Future Generation Computer Systems, 15(2), pp. 245-254

MORENO-RAMIREZ, D. et al., 2007. Store-and-forward teledermatology in skin cancer triage: experience and evaluation of 2009 teleconsultations. Archives of Dermatology, 143(4), pp. 479

MOUMTZOGLOU, A. and KASTANIA, A., 2013. An Expository Discourse of E-Health. Systems Analysis Tools for Better Health Care Delivery. Springer. pp. 49-63

MUFIT FERMAN, A., 2008. Automatic detection of red-eye artifacts in digital color photos. Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on. IEEE. pp. 617-620

MURPHY, K. P., 2012.Machine learning: a probabilistic perspective. MIT press.

MURPHY, P. et al., 2006. How red is a white eye? Clinical grading of normal conjunctival hyperaemia. Eye, 21(5), pp. 633-638

MURRAY, I.J. et al., 2012. Sex-related differences in peripheral human color vision: A color matching study. Journal of vision, 12(1),

MAGLOGIANNIS, I. et al., 2001. Utilizing artificial intelligence for the characterization of dermatological images. In 4th International Conference Neural Networks and Expert Systems in Medicine and Healthcare, Milos Island, Greece (pp. 362-368).

NAKAZATO, M., MANOLA, L. and HUANG, T.S., 2003. ¡ i¿ ImageGrouper¡/i¿: a group-oriented user interface for content-based image retrieval and digital image arrangement. Journal of Visual Languages & Computing, 14(4), pp. 363-386

NASSAU, K., 2001. The physics and chemistry of color: the fifteen causes of color. The Physics and Chemistry of Color: The Fifteen Causes of Color, 2nd Edition, by Kurt Nassau, pp.496.ISBN 0-471-39106-9.Wiley-VCH, July 2001., 1

NORMAN, D.A., 1983. Design rules based on analyses of human error. Communications of the ACM, 26(4), pp. 254-258

OATES, B.J., 2005. Researching information systems and computing. Sage.

OBSTFELDER, A., ENGESETH, K.H. and WYNN, R., 2007. Characteristics of successfully implemented telemedical applications. Implement Sci, 2(25), pp. 1-11

ODINOT, G., WOLTERS, G. and VAN KOPPEN, P.J., 2009. Eyewitness memory of a supermarket robbery: A case study of accuracy and confidence after 3 months. Law and human behavior, 33(6), pp. 506

OIKONOMOU, S., GWYNEDD, Y. and TRUST, N.W.N., 2009. Teledermatology: digital revolution in the management of skin disease. Proceedings of the 2nd WSEAS international conference on Biomedical electronics and biomedical informatics. World Scientific and Engineering Academy and Society (WSEAS). pp. 117-121

OLSSON, N. and JUSLIN, P., 2002. Calibration of confidence among eyewitnesses and earwitnesses. Metacognition. Springer. pp. 203-218

OLSSON, N., JUSLIN, P. and WINMAN, A., 1998. Realism of confidence in earwitness versus eyewitness identification. Journal of Experimental Psychology: Applied, 4(2), pp. 101

ORRUO, E. et al., 2011. Evaluation of teledermatology adoption by health-care professionals using a modified Technology Acceptance Model. Journal of telemedicine and telecare, 17(6), pp. 303-307

OUNI, S., CHAMBAH, M., HERBIN, M. and ZAGROUBA, E., 2008. Are existing procedures enough? Image and video quality assessment: review of subjective and objective metrics. Electronic Imaging 2008. International Society for Optics and Photonics. pp. 68080Q-68080Q-11

PANDEY, A., DEEN, A. J., and PANDEY, R., 2013. Color and Shape Content Based Image Classification using RBF Network and PSO Technique: A Survey.arXiv preprint arXiv:1311.6881.

PANDURAGAN, S.L. et al., 2011. Level of confidence among nursing students in the clinical setting. Procedia-Social and Behavioral Sciences, 18, pp. 404-407

PAONE, S. and SHEVCHIK, G., 2013. Making a Business Case for eHealth and Teleservices. Telerehabilitation. Springer. pp. 297-309

PAPPAS, Y. and SEALE, C., 2009. The opening phase of telemedicine consultations: An analysis of interaction. Social science & medicine, 68(7), pp. 1229-1237

PARDO, P.J., PEREZ, A. and SUERO, M., 2007. An example of sex?linked color vision differences. Color Research & Application, 32(6), pp. 433-439

PASCHOS, G. and PETROU, M., 2003. Histogram ratio features for color texture classification. Pattern Recognition Letters, 24(1), pp. 309-314

PASS, G. and ZABIH, R., 1999. Comparing images using joint histograms. Multimedia systems, 7(3), pp. 234-240

PATHIPATI, A. and ARMSTRONG, A., 2011. Teledermatology: Outcomes and Economic Considerations.

PATRICOSKI, C. and FERGUSON, A.S., 2009. Selecting a digital camera for telemedicine. TELEMEDICINE and e-HEALTH, 15(5), pp. 465-475

PEIRCE, C.S. and JASTROW, J., 1885. On small differences of sensation. US Government Printing Office.

PEREDNIA, D.A., 1991. What dermatologists should know about digital imaging. Journal of the American Academy of Dermatology, 25(1), pp. 89-108

PETERSON, R.C. and WOLFFSOHN, J.S., 2009. Objective grading of the anterior eye. Optometry & Vision Science, 86(3), pp. 273-278

PHILIP, B.R., 1938. Sex differences in the perception of color mass. The American Journal of Psychology, 51(2), pp. 398-404

PFAUTZ, J. AND ROTH, E., 2006. Using cognitive engineering for system design and evaluation: A visualization aid for stability and support operations. International Journal of Industrial Ergonomics, 36(5), pp.389-407.

PICCOLO, D. et al., 1999. Face-to-face diagnosis vs telediagnosis of pigmented skin tumors: a teledermoscopic study. Archives of Dermatology, 135(12), pp. 1467

PICKFORD, R.W., 1951. Individual differences in colour vision.

PLESKAC, T.J. and BUSEMEYER, J.R., 2010. Two-stage dynamic signal detection: a theory of choice, decision time, and confidence. Psychological review, 117(3), pp. 864

POPPER, K., 2002. The logic of scientific discovery. Routledge.

PORTER, J. et al., 2010. Ecological image databases: from the webcam to the researcher. Ecological Informatics, 5(1), pp. 51-58

PREECE, J. and BENYON, D., 1993. A guide to usability: Human factors in computing. Addison-Wesley Longman Publishing Co., Inc.

PREECE, J., SHARP, H. AND ROGERS, Y., 2015. Experimental design. Interaction Design: Beyond Human-Computer Interaction, p.486.

PULFORD, B.D., 1996. Overconfidence in human judgement. University of Leicester.

PUNCHARD, N.A., WHELAN, C.J. and ADCOCK, I., 2004. Journal of inflammation, 1(1), pp. 1

RTSCH, G., 2004. A brief introduction into machine learning. In21st Chaos Communication Congress.

REASON, J., 2006.Human error. Cambridge university press.

REASON, J., 2000. Human error: models and management. BMJ: British Medical Journal, 320(7237), pp. 768

RHEINGANS, P.L., 2000. Task-based color scale design. 28th AIPR Workshop: 3D Visualization for Data Exploration and Decision Making. International Society for Optics and Photonics. pp. 35-43

RICHTER, G.M. et al., 2009. Telemedicine for retinopathy of prematurity diagnosis: evaluation and challenges. Survey of ophthalmology, 54(6), pp. 671-685

RIMNER, T. et al., 2010. Digital skin images submitted by patients: an evaluation of feasibility in store-and-forward teledermatology. European Journal of Dermatology, 20(5), pp. 606-610

ROS-YUIL, J., 2012. Correlation Between Face-to-Face Assessment and Telemedicine for the Diagnosis of Skin Disease in Case Conferences. Actas Dermo-Sifiliogrficas (English Edition), 103(2), pp. 138-143

RODRIGUEZ, J.D. et al., 2013. Automated grading system for evaluation of ocular redness associated with dry eye. Clinical Ophthalmology, 7, pp. 1197-1204

RODRIGUEZ-CARMONA, M. et al., 2008. Sex-related differences in chromatic sensitivity. Visual neuroscience, 25(3), pp. 433

ROGERS, M.L. et al., 2004. Barriers to implementing wrong site surgery guidelines: a cognitive work analysis. Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on, 34(6), pp. 757-763

ROMERO, G., CORTINA, P. and VERA, E., 2008. Telemedicine and teledermatology (II): current state of research on dermatology teleconsultations. Actas Dermo-Sifiliogrficas (English Edition), 99(8), pp. 586-597

ROMERO, G., GARRIDO, J. and GARCA-ARPA, M., 2008. Telemedicine and teledermatology (I): concepts and applications. Actas Dermo-Sifiliogrficas (English Edition), 99(7), pp. 506-522

ROTGER, V., SOLARZ, P., RUIZ, L., SALAS, A., MENA, M.G. and OLIVERA, J., 2011. Teledermatology: an experience in Tucumn. Journal of Physics: Conference Series. IOP Publishing. pp. 012053

SALIBA, V. et al., 2012. Telemedicine across borders: A systematic review of factors that hinder or support implementation. International journal of medical informatics,

SAMUEL, A. L., 1959. Some studies in machine learning using the game of checkers.IBM Journal of research and development,3(3), 210-229.

SANDER, P. and SANDERS, L., 2003. Measuring confidence in academic study: A summary report. Electronic Journal of Research in Educational Psychology and Psychopedagogy, 1(1), pp. 1-17

SASAKI, H. et al., 2007. Right hemisphere specialization for color detection. Brain and cognition, 64(3), pp. 282-289

SAUNDERS, M.N. et al., 2011. Research Methods For Business Students, 5/e. Pearson Education India.

SCHULZ, C. et al., 2011. Visual attention of anaesthetists during simulated critical incidents. British journal of anaesthesia, 106(6), pp. 807-813

SCHULZE, M.M., HUTCHINGS, N. and SIMPSON, T.L., 2009. The perceived bulbar redness of clinical grading scales. Optometry & Vision Science, 86(11), pp. E1250-E1258

SCHULZE, M.M., JONES, D.A. and SIMPSON, T.L., 2007. The development of validated bulbar redness grading scales. Optometry & Vision Science, 84(10), pp. 976-983

SCHULZE-WOLLGAST, P., TOMINSKI, C. and SCHUMANN, H., 2005. Enhancing Visual Exploration by Appropriate Color Coding. WSCG (Full Papers). pp. 203-210

SERENER, A., KAVALCIOGLU, C. and CYPRUS, N., Teledermatology based medical images with AWGN Channel in Wireless Telemedicine System. Proceedings of the 1st WSEAS International Conference on Manufacturing Engineering, Quality and Production Systems. pp. 145-150

SERGYN, S., 2007. Color content-based image classification. In5th Slovakian-Hungarian Joint Symposium on Applied Machine Intelligence and Informatics(pp. 427-434).

SHANNON, G.W. and BUKER, C.M., 2010. Determining accessibility to dermatologists and teledermatology locations in Kentucky: demonstration of an innovative geographic information systems approach. Telemedicine and e-Health, 16(6), pp. 670-677

SHAO-ZHEN, Y. andXIAN-DONG, Y., 2010. Medical Image Retrieval Based on Extraction of Region of Interest. Bioinformatics and Biomedical Engineering (iCBBE), 2010 4th International Conference on. IEEE. pp. 1-4

SHAPIRO, M. and CROASDALE, C., 1997. The red eye: a clinical guide to a rapid and accurate diagnosis. Cornea.St.Louis: Mosby, , pp. 438-445

SHIH, J. and CHEN, L., 2002. Colour image retrieval based on primitives of colour moments. IEE Proceedings-Vision, Image and Signal Processing, 149(6), pp. 370-376

SIKKA, N. et al., 2012. The Use of Mobile Phone Cameras in Guiding Treatment Decisions for Laceration Care. Telemedicine and e-Health, 18(7), pp. 554-557

SINGH, S.M. and HEMACHANDRAN, K., 2012. Content-Based Image Retrieval using Color Moment and Gabor Based Image Retrieval using Color Moment and Gabor Texture Feature Texture Feature.

SMITH, S.E. et al., 2013. Use of Telemedicine to Diagnose Tinea in Kenyan Schoolchildren. Telemedicine and e-Health, 19(3), pp. 166-168

SONG, M. et al., 2010. Color to gray: visual cue preservation. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 32(9), pp. 1537-1552

SPALDING, J., 1999. Colour vision deficiency in the medical profession. The British Journal of General Practice, 49(443), pp. 469

STANBERRY, B., 2000. Telemedicine: barriers and opportunities in the 21st century. Journal of internal medicine, 247(6), pp. 615-628

STANKOV, L. et al., 2012. Confidence: A better predictor of academic achievement than self-efficacy, self-concept and anxiety? Learning and Individual Differences, 22(6), pp. 747-758

STOECKER, W.V. and MOSS, R.H., 1992. Editorial: digital imaging in dermatology. Computerized Medical Imaging and Graphics, 16(3), pp. 145-150

STOWE, S. and HARDING, S., 2010. Telecare, Telehealth and telemedicine. European Geriatric Medicine, 1(3), pp. 193-197

STRICKER, M.A. andORENGO, M., 1995. Similarity of color images. IS&T/SPIE's Symposium on Electronic Imaging: Science & Technology. International Society for Optics and Photonics. pp. 381-392

STRUBER, J., 2004. An introduction to telemedicine and email consultations. The Internet Journal of Allied Health Sciences and Practice, 2(3)

STULBERG, D.L., PENROD, M.A. and BLATNY, R.A., 2002. Common bacterial skin infections. American Family Physician, 66(1), pp. 119-128

TARABISHY, A.B. and JENG, B.H., 2008. Bacterial conjunctivitis: a review for internists. Cleveland Clinic journal of medicine, 75(7), pp. 507-512

TAYLOR, S.C., 2002. Skin of color: biology, structure, function, and implications for dermatologic disease. Journal of the American Academy of Dermatology, 46(2), pp. S41-S62

TAI, M., YANG, R.Y.H. AND LIAO, S., 2010, June. A study of color description and cognition. In 3rd International Conference on Information Sciences and Interaction Sciences (ICIS) (pp. 138-141).

TERRY, R. et al., 1995. Variability of clinical investigators in contact lens research. (CL-359). Optometry & Vision Science, 72(12), pp. 16

THEROUX, A., 1995. [BOOK REVIEW] THE PRIMARY COLORS, THREE ESSAYS. New Statesman and Society, 8(380), p.46.

TOROK, M. and CONLON, C., 2009. Skin and soft tissue infections. Medicine, 37(11), pp. 603-609

TRAN, K. et al., 2011. Mobile teledermatology in the developing world: implications of a feasibility study on 30 Egyptian patients with common skin diseases. Journal of the American Academy of Dermatology, 64(2), pp. 302-309

TSAI, C., 2007. Image mining by spectral features: A case study of scenery image classification. Expert Systems with Applications, 32(1), pp. 135-142

TSAI, H. et al., 2004. Teleconsultation by using the mobile camera phone for remote management of the extremity wound: a pilot study. Annals of Plastic Surgery, 53(6), pp. 584-587

TVERSKY, A., and KAHNEMAN, D., 1974. Judgment under uncertainty: Heuristics and biases.science,185(4157), 1124-1131.

VAN DER HEIJDEN, JOB P et al., 2010. A pilot study on tertiary teledermatology: feasibility and acceptance of telecommunication among dermatologists. Journal of telemedicine and telecare, 16(8), pp. 447-453

VIOPIO, V. and LAMMINEN, H., 2002. Lighting and colour in digital photography. Teledermatology. London: Royal Society of Medicine Press.

VICKERS, D. AND PACKER, J., 1982. Effects of alternating set for speed or accuracy on response time, accuracy and confidence in a unidimensional discrimination task. Acta psychologica, 50(2), pp.179-197.

VU, B.L. et al., 2003. Development of a clinical severity score for preseptal cellulitis in children. Pediatric emergency care, 19(5), pp. 302-307

WALLACE, D. et al., 2012. A systematic review of the evidence for telemedicine in burn care: with a UK perspective. Burns, 38(4), pp. 465-480

WARSHAW, E.M. et al., 2011. Teledermatology for diagnosis and management of skin conditions: a systematic review. Journal of the American Academy of Dermatology, 64(4), pp. 759-772. e21

WARSHAW, E.M. et al., 2009. Accuracy of teledermatology for nonpigmented neoplasms. Journal of the American Academy of Dermatology, 60(4), pp. 579-588

WHITED, J.D., 2010. Economic analysis of telemedicine and the teledermatology paradigm. Telemedicine and e-Health, 16(2), pp. 223-228

WHITED, J.D. et al., 1999. Reliability and accuracy of dermatologists? clinic-based and digital image consultations. Journal of the American Academy of Dermatology, 41(5), pp. 693-702

WHITTEN, P.S., 2003. Teledermatology delivery modalities: real time versus store and forward. Curr Probl Dermatol, 32, pp. 24-31.

WIRBELAUER, C., 2006. Management of the red eye for the primary care physician. The American Journal of Medicine, 119(4), pp. 302-306

WITKAMP, L., 2009. Teledermatology Helps Doctors and Hospitals to Serve Their Clients. Electronic Healthcare. Springer. pp. 98-105

WOLFFSOHN, J., 2004. Incremental nature of anterior eye grading scales determined by objective image analysis. British Journal of Ophthalmology, 88(11), pp. 1434-1438

WOOTTON, R., CRAIG, D.J. and PATTERSON, V., 2006. Introduction to telemedicine. Royal Society of Medicine Press.

WRIGHT, G. and AYTON, P., 1989. Judgemental probability forecasts for personal and impersonal events. International Journal of Forecasting, 5(1), pp. 117-125

WURM, E.M., CAMPBELL, T.M. and SOYER, H.P., 2008. Teledermatology: how to start a new teaching and diagnostic era in medicine. Dermatologic clinics, 26(2), pp. 295-300

WURM, E.M. and SOYER, H.P., 2012. Mobile Teledermatology. Telemedicine in Dermatology. Springer. pp. 79-85

XIE, Q. and LIU, J., 2010. Mobile phone based biomedical imaging technology: A newly emerging area. Recent Patents on Biomedical Engineering, 3(1), pp. 41-53

YAN, S., SAYAD, S. and BALKE, S.T., 2009. Image quality in image classification: Design and construction of an image quality database. Computers & Chemical Engineering, 33(2), pp. 421-428

ZHANG, Q. andXIAO, H., 2008. Extracting regions of interest in biomedical images. Future BioMedical Information Engineering, 2008. FBIE'08. International Seminar on. IEEE. pp. 3-6

ZUKOSKI, M.J., BOULT, T. and IYRIBOZ, T., 2006. A novel approach to medical image compression. International journal of bioinformatics research and applications, 2(1), pp. 89-1

Appendix A

A.1 Appendix 1: Proposed specification for auto evaluation system using digital images in telehealth

The following section provides a brief top-down description of a proposed system as an example of one of the possible applications of the current study.

Purpose of the system
The proposed red eye tele-evaluation system is an auto enrolment and classification diagnostic system that supports doctors in judging the degree of redness in infected eyes using digital images. The system features a database of stored image data, an image quality scale, a redness scale, and a confidence scale in order to assess their values Data gathered by such a system is sent to doctors in order to decide whether patients needed further treatment, a follow up consultation or referral to a specialist for more detailed investigation.

Overall description
The proposed system focuses on the intensity of red as a key characteristic in the diagnosis of eye infection. There are three main levels of the redness intensity in the proposed system; normal, moderate, and severe. The system has an image database (or atlas) which contains training images. These images are ranked and labelled based on their redness intensity and are used to classify and rank the newly received images. The system accepts only images of sufficiently high resolution before starting the classification. It confirms the accuracy of the classification by testing the confidence level of the system operator. The system accepts only subjective confidence scores of 70% or more and sends feedback and reports to all concerned users, such as the ophthalmologist, the patient's GP, the patient and the system operator.

Area of application
This system is designed as a diagnostic tool for conjunctivitis (commonly called "red

eye"), however, it might also be applied to diagnosing other pathologies where colour is one of the key symptoms, for example skin infections such as cellulitis.

Telehealth method
This system is envisaged as an interactive system to classify digital images of conjunctivitis cases based on their degree of severity using a store and forward consultation model. Users of the system There are four main user groups of the system:

- Image senders: this includes patients, carers and nurses.
- Doctors: this includes ophthalmologist and GPs.
- Image receivers: this includes mainly the system operators, and doctors.
- Technical support staff: this includes the IT and maintenance staff.

System Interaction
The proposed system interacts with the following:

- The system interacts with the four main categories of users listed above.
- The system interacts with the existing NHS medical database in order to create, store, retrieve and update patient information.
- The system interacts with the proposed image atlas. This is a database storing images of red eyes for comparative and system training purposes.
- Image capture devices such as mobile phones.

Components of the system
The proposed system has the following components:

- Image data acquisition which is mentioned above as the capture devices.
- Image quality scale measurement
- Redness intensity scale measurement
- Operator confidence scale measurement
- Image atlas

The following section describes the components above.
Image data acquisition
The users capture and acquire a digital image using a mobile phone camera, or web cam or digital camera. The images are then sent via telecommunications links (either radio frequency mobile phone signals or internet, depending on system configuration) to the telehealth system in the hospital.

Image quality scale measurement

The scale was developed for the original study described in this thesis in order to standardise the quality of images acquired from different devices. This is required in the case of the auto-evaluation system to ensure a minimum level of quality. The scale consists of eleven scale divisions from 1 to 11. Each of these divisions represents image resolution in percentages from 0% to 100% and matching image quality levels from poor to high. This scale will be able to classify images as poor, low, medium, or high quality, which values will be used as indicators for the acceptance or rejection of an image. If the image quality is classified as poor, low or medium (less than 70% resolution), this indicates that the image quality is below the minimum required level and will be rejected by the system. If the image quality is classified as high, equal or above 70% resolution, this indicates suitable quality and the image will be accepted by the system. The following table A.1 shows the proposed eleven divisions along with their respective resolution percentages, and image quality classification descriptors.

Table A.1: Image quality scale measurement

Scale division	1	2	3	4	5	6	7	8	9	10	11
Resolution%	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
Image quality	poor	poor	poor	low	low	medium	medium	high	high	high	high

Red eye scale measurement

The proposed scale is a modified version of a scale that is already in use with the NHS in the UK (Figure 2.13 and Figure 7.1) for diagnosing conjunctivitis during traditional face-to-face consultations. The scale represents the degree of redness in 3 main categories (normal, moderate and severe), apportioned over 100 scale divisions that are expressed as percentage of redness. Normal, moderate and severe are defined as follows:

1. Normal: This is an indication of no redness in the front part of the eye. This category includes eyes with redness between 0% and 10%
2. Moderate: This includes cases of trace, mild and moderate eye redness with redness between 11% and 50%.
3. Severe: This level of classification includes severe and very severe eye redness with redness between 51% and 100%.

This scale classifies redness as shown in the scale below in table A.2.

Table A.2: Red eye scale measurement

Scale division	1	2	3
Redness %	0% to 10%	11% to 50%	51% to 100%
Severity levels	normal/none	moderate (including trace, mild)	severe (including very severe)

Operator confidence scale measurement

A confidence scale is used by the system to confirm the accuracy of the redness classification. The system displays a questionnaire of the confidence scale shown below in table A.3 and then asks the user to choose the level of confidence that they have in the matches that have been reported by the system. The scale has 10 divisions (0 to 9) with corresponding confidence measures between 0% and 100% distributed over 3 levels (low, medium and high). If the confidence is less than 70%, the user is requested to repeat the classification process. The result is accepted if the confidence is 70% or more.

Table A.3: Operator confidence scale measurement

Scale division	0	1	2	3	4	5	6	7	8	9
Confidence %	<10%	<20%	<30%	<40%	<50%	<60%	<70%	<80%	<90%	≤100%
Confidence level	low			medium				high		

Image atlas database

The image atlas is a database that contains a set of training images that together represent all divisions on the redness scale. The database must therefore include at least 100 images. Each image will be processed and stored according to its redness intensity by isolating the region of interest using edge detection and colour extraction and matching techniques. Images will be matched using image classification techniques, like those mentioned earlier in the literature review in chapter 2. Image properties stored will include a unique image identifier and name, image ranking, image display, image division, redness description, and redness intensity.

Table A.4 shows example image properties as stored in the database. The table gives one example of each of the three possible degrees of redness.

Table A.4: Examples of image atlas

Label	Example 1	Example2	Example3
Image ID	212	343	454
Image name	red-eye_image1	red-eye_image2	red-eye_image3
Image ranking	10	45	80
Image display	display1	display2	display3
Image division	1	2	3
Redness description	normal/none	moderate (including trace, mild)	severe (including very severe)
Redness intensity	10%	45%	80%

The database allows every image created in the system to be used as a query against the training images that are already in the database in order to classify a new image within one of the divisions. The new image may then be displayed along with a listing of the properties mentioned above. Figure A.1 shows how the proposed system processes and compares both types of images (training and query images)

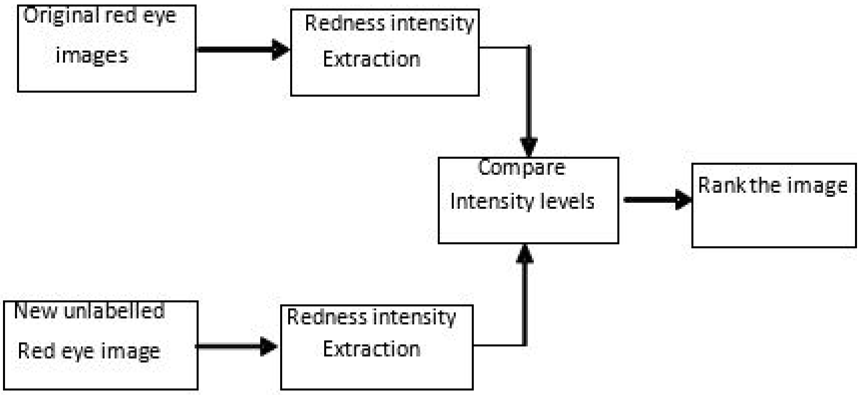

Original red eye images
Redness intensity Extraction
Compare Intensity levels
Rank the image
New unlabelled Red eye image
Redness intensity Extraction

Figure A.1: colour intensity matching method

A new image is processed by measuring its level of intensity using the same procedures that will have been used initially when processing the original images that are now in the database. The system classifies the intensity level of the colour red in the new image then compares it with the classification data stored for the images in the database. The system then identifies the closest-matching image from the database, copies the salient properties to the new image and ranks it appropriately before committing it to storage.

To ensure optimum classificatory accuracy, there are key steps which need to be considered and well-integrated when designing any objective automatic diagnostic protocol in telehealth. The same steps can also be applied and integrated into automatic, semi-automatic and manual systems. In the manual method, the processes are usually carried out by a panel of three to five consultants. Qualitative measures would be used to approve the image before ranking and matching to the scales. Comparison might not be entirely objective but could be based on clinical practice and the experience of the operator. The automatic and semi-automatic systems include steps such as isolating regions of interest, colour feature extraction, colour feature measurement, image comparison, image matching, and query image classification.

System flow diagram 1 (The key stages of the system)
The following diagram in figure A.2 shows the seven main operational stages of the proposed system.

Figure A.2 gives only a general indication of the main processes involved in the proposed system. A more detailed flowchart may be inspected in chapter 2 (Fig 2.16)

The following section provides a brief explanation of the flow diagram figure (2.16) in

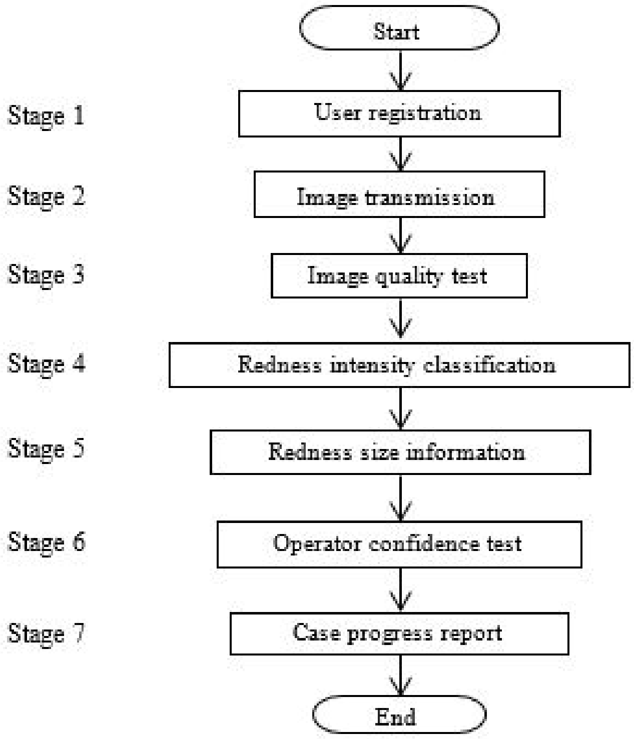


Figure 8.2: Red eye tele- diagnosis system key stages

Figure A.2: Red eye telediagnosis system key stages

chapter 2.

System interactive flow
All the interactions in the proposed auto-evaluation system are further explained as follows:

Stage 1: User registration

- Users register and log in order to use the system. Registration will proceed using protocols that are outwith the scope of the present discussion.
- The system sets up log in details for registered users
- The system links the user with NHS records, from which their medical history can be retrieved.

Stage 2: Image transmission

- Users capture digital images of their affected eyes using a suitable mobile phone.
- Users send the digital images to the system in the hospital using the mobile phone

Stage 3: Image quality test

- Auto-evaluation system runs an image quality test using the image quality scale.
- Auto-evaluation system accepts the user's image if its quality is calculated to be 70% or more.
- Auto-evaluation system rejects the user's image if its quality is lower than 70%.
- Auto-evaluation system sends feedback to the user as "Image accepted" if it is satisfactory.
- Auto-evaluation system sends feedback to the user as "Image is not accepted. Please send a new image with higher quality/resolution", if it is unsatisfactory.
- Auto-evaluation system stores the image in the system database with a unique identifier and name.

Stage 4: Red eye classification Auto-evaluation system conducts comparisons between the user's image and the images in the database. The system runs the classification for the user's image in the following order:

- Auto-evaluation system identifies and stores the calculated intensity level of the redness in the image, assigning it as a percentage.
- Auto-evaluation system retrieves image(s) from the image atlas with the same degree of redness intensity.
- Auto-evaluation system displays the two images while announcing "Redness intensity match found".
- Auto-evaluation system copies the properties of the matched image to a new record created for the user's image, ranks it among the other database images and commits it to storage.
- Auto-evaluation system displays the two images along with a readout of their image properties, including their ranking in the scale, division number and redness description.

Stage 5: Redness size information

- The system counts the pixels that have redness in the front part of the eye
- The system records the number
- The system displays the information as part of the progress report
- If the number of pixels increases in the follow up report, this means that the redness is spreading and covering more areas.

Stage 6: Operator confidence test

When the image quality test and redness classification are completed, the system conducts a confidence test to check the level of confidence that the operator of the system has in the classification result on inspection of the user's image and the matched image displayed together. The system does the following:

- Auto-evaluation system checks if the operator saw the displayed new image.
- Auto-evaluation system checks if the operator saw the displayed matched images.
- Auto-evaluation system checks if the operator saw the result of image quality, redness classification.
- Auto-evaluation system checks if the operator accepts the results (yes or no).
- If the operator rejects the results, a support enquiry is made with the IT team and the system repeats the procedure.
- If the operator accepts the results then the Auto-evaluation system checks the level of confidence that the operator has in the accuracy of the result by display a question with words to the effect of the following: Please, highlight below the level of your confidence in the accuracy of the result presented by the system.

Table A.5: Operator confidence scale measurement

Scale division	0	1	2	3	4	5	6	7	8	9
Confidence %	<10%	<20%	<30%	<40%	<50%	<60%	<70%	<80%	<90%	≤100%
Confidence levels	low			medium				high		

- If the operator indicates a confidence level that is between 0% and 70% in table A.5 then the system notifies all users of the system that the results have been rejected.
- If the operator indicates a confidence level that is over 70%, then the system notifies all users that the results have been accepted.

Stage 7: Case progress report

- When the results are accepted the system notifies the users. These include the patient, the system operator and the GP of the patient. The notification encloses information about the user's image, now processed and stored in the system's database. This information consists of the image's unique identifier, the name attributed to the image by the system, the assigned division number, its calculated image ranking, redness description text, and redness intensity values.
- If the newly labelled image does not pass the confidence test, which happens if the result of the operator confidence test is less than 70%, then the system does not recognise the classification and asks the concerned users to repeat the procedure again.
- The system repeats the same procedure whenever a patient uploads new images. These may be ordered by the medical staff in order to investigate the progression of the condition.
- The system records any follow up procedures that are requested by the medical staff.
- If, the system's auto-analysis of subsequently uploaded patient's images perceives increased redness then the patient's condition is deemed to have worsened. The system then sends an alert email, tagged as "highly important", to the medical staff. This email reports the increase in redness and provides other salient information about the progression of the patient's condition. Alternatively, if the patient's condition is deemed to have improved, then the system sends an email to the medical staff reporting this accordingly.
- If the patient's condition persists for seven days then the system sends an email to report this also.
- The doctor will decide if further action is required. The information received from the system is available as a guide to assist them with this.
- The system automatically updates the patient's medical records in the NHS databases.

The system can be fully automatic (not semi-automatic) at the receiving end, where the patient's images are evaluated. The absence of human involvement in this part of the system is critically important to ensure a high level of accuracy. Figure A.3 shows how the system works fully independently.

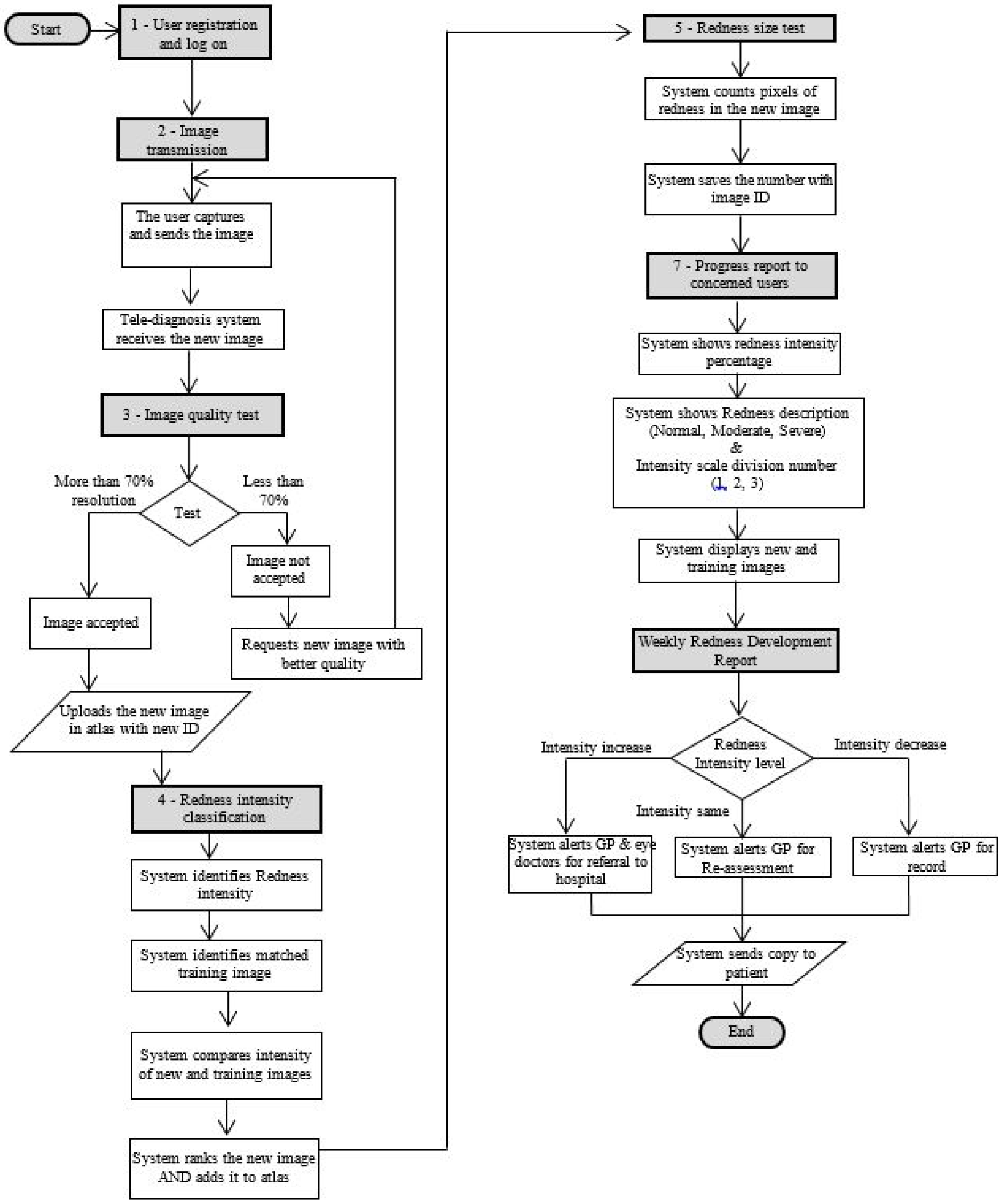

Start
1 - User registration and log on
2 - Image transmission
The user captures and sends the image
Tele-diagnosis system receives the new image
3 - Image quality test
More than 70% resolution
Test
Less than 70%
Image not accepted
Image accepted
Requests new image with better quality
Uploads the new image in atlas with new ID
4 - Redness intensity classification
System identifies Redness intensity
System identifies matched training image
System compares intensity of new and training images
System ranks the new image AND adds it to atlas
5 - Redness size test
System counts pixels of redness in the new image
System saves the number with image ID
7 - Progress report to concerned users
System shows redness intensity percentage
System shows Redness description (Normal, Moderate, Severe) & Intensity scale division number (1, 2, 3)
System displays new and training images
Weekly Redness Development Report
Redness Intensity level
Intensity increase
Intensity same
Intensity decrease
System alerts GP & eye doctors for referral to hospital
System alerts GP for Re-assessment
System alerts GP for record
System sends copy to patient
End

Figure A.3: Conjunctivitis auto-evaluation system (Eye redness intensity classification)

This study proposes the use of the above automated system in classifying redness as part of the diagnostic protocol when investigating cases of conjunctivitis, as well as other conditions that also involve redness as a key indicator, such as cellulitis. The motivation is to save time and expense and to ease the difficulties of healthcare provision in remote and inaccessible locations. The system would augment and enhance the current set up of teleconsultation in NHS24 and A&E at ARI in terms of increasing objectivity and obviating errors caused by inconsistencies in human colour perception. Typical errors that can occur with the current manual system are mismatching or disordering the images (for example, tiredness or lack of concentration can lead to a query image being erroneously matched with dissimilar images from the database, or the colour intensity of a query image might be inaccurately assessed.

This can be done if the system is built, tested and its results medically approved. This requires further studies and more evaluation of the current teleconsultation system used at ARI.

A.2 Appendix 2: Results and Analysis of Task1A (Colour Description)

See Figure A.4 and Figure A.5 for this Appendix

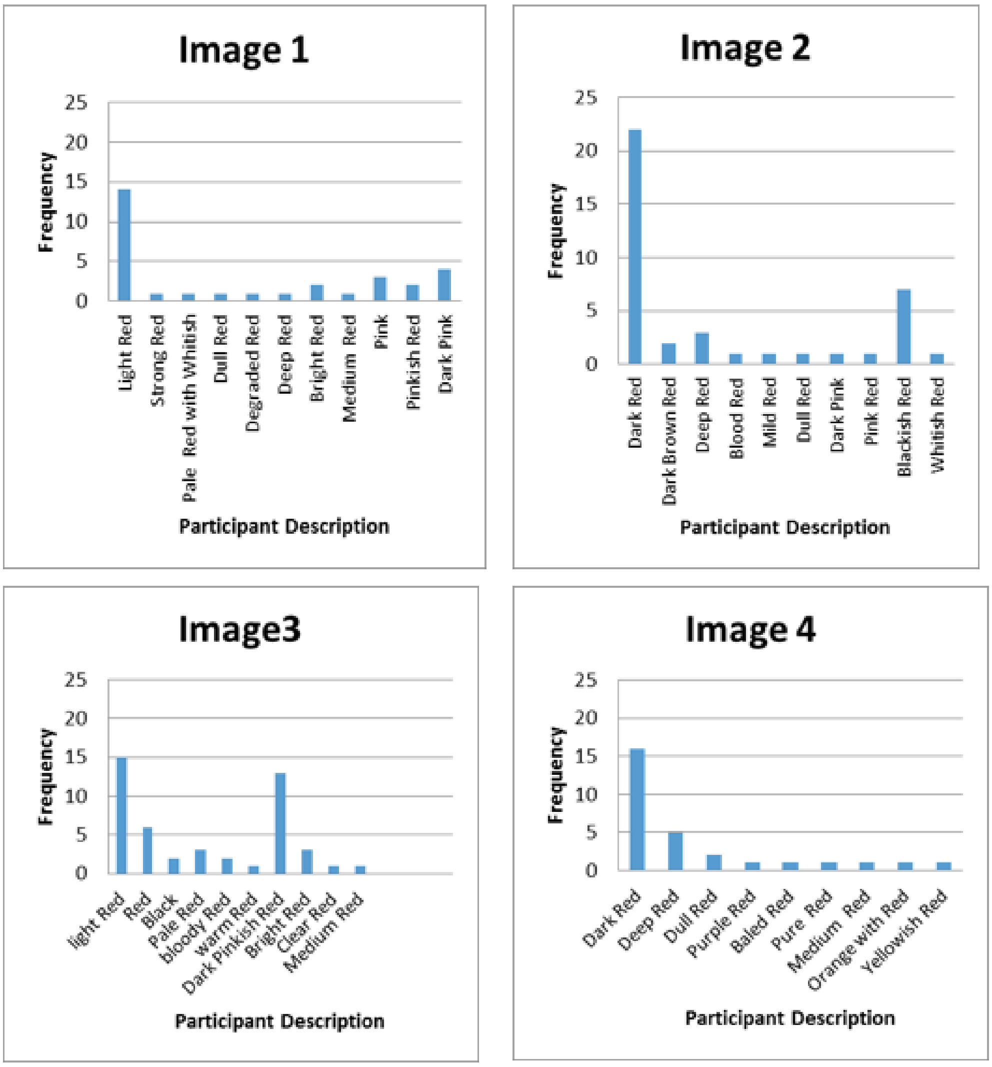


Figure A.4: The participant descriptions of the colour in images 1 to 4

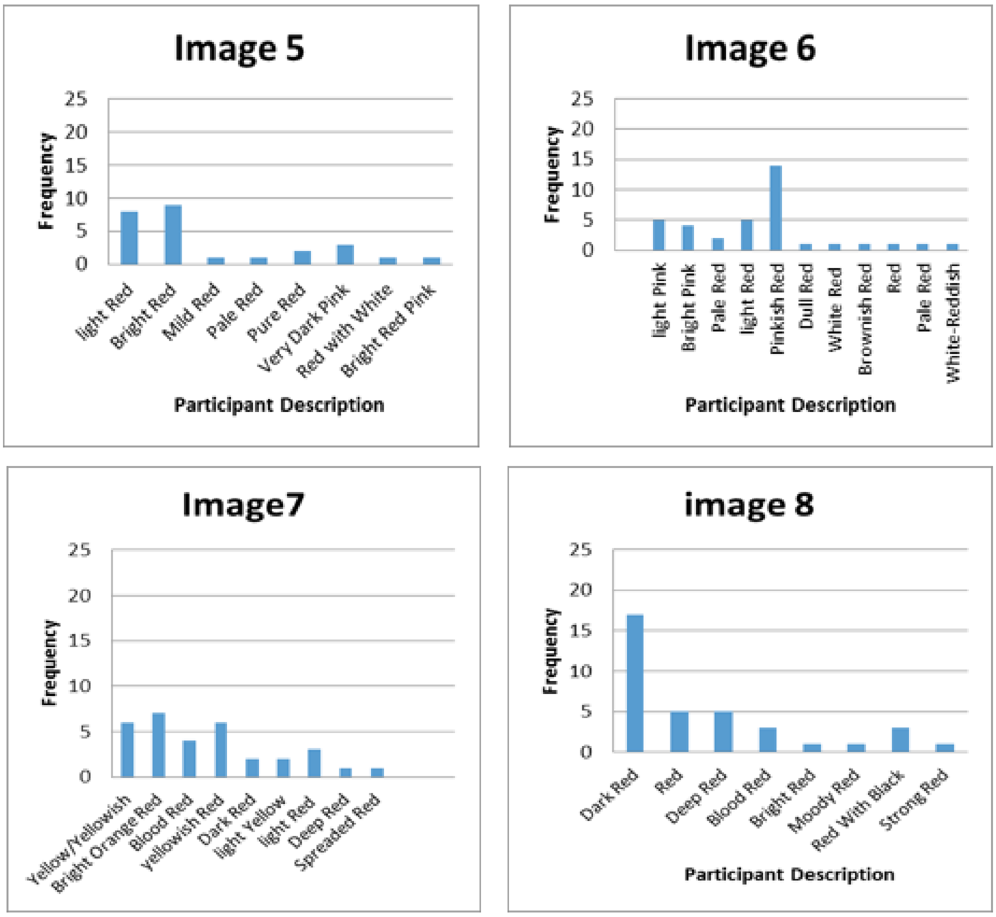


Figure A.5: The participant descriptions of the colour in images 5 to 8

A.3 Appendix 3: Image Classification. Python source code

```
# This is Python code for image classification based on colour.
# Image classification is done using the OpenCV library.
# A Nearest Neighbour algorithm is used to compare HSV
# (Hue Saturation Value) colour histograms of images.
#
# Running Experiment;
#
# Required Tools;
#   - Anaconda command prompt
#   - IPython Notebook
#   - Open CV
#
#  To run;
```

```
#
#  1- Open Anaconda command prompt
#  2- Type "ipython notebook" (without the ‘‘)
#  3- Browser opens containing files
#  4- Select file of code to run
#  5- Code editor opens
#  6- Click the "run" button

# SECTION 1: Package import and Path Definition
# ------------------------------------------------------------------------
import cv2       #import opencv
from matplotlib import pyplot as plt
import os

path = "C:/images/experiments/"
# ------------------------------------------------------------------------
# SECTION 2: Containers Declaration
# ------------------------------------------------------------------------
for folder in os.listdir(path):
    train_images = {}   #container for training images
    train_histograms = {} #container for trainig image histograms
    train_path = path+folder+"/train/" #path to folder where training images are stored
# ------------------------------------------------------------------------
# SECTION 3: Reading Images & Conversion
# ------------------------------------------------------------------------
  for f in os.listdir(train_path):  #for each file in the folder
       im = cv2.imread(train_path+f)  #read the image file
       im = cv2.cvtColor(im, cv2.COLOR_BGR2HSV) #convert image color format from BGR to RGB
# ------------------------------------------------------------------------
# SECTION 4: Storing Images & Histogram
# ------------------------------------------------------------------------
       train_images[f] = im #store training image
       hist = cv2.calcHist([im], [0], None, [8], [0,255]) #calculate histogram
       hist = cv2.normalize(hist).flatten() #normalise histogram
       train_histograms[f] = hist #store histogram
# ------------------------------------------------------------------------
# SECTION 5: Repeating steps 2,3, and 4
# ------------------------------------------------------------------------
    test_images = {}
    test_histograms = {}
    test_path = path+folder+"/test/"
    for f in os.listdir(test_path):
       im = cv2.imread(test_path+f)
       im = cv2.cvtColor(im,cv2.COLOR_BGR2HSV)
       test_images[f] = im
```

```
        hist = cv2.calcHist([im], [0], None, [8], [0,255])
        hist = cv2.normalize(hist).flatten()
        test_histograms[f] = hist

# -----------------------------------------------------------------------
# SECTION 6: Perform KNN by comparing distances between training and Test Sets
# -----------------------------------------------------------------------
    img_results = {} #container for saving results in memory
    for (k1, hist1) in test_histograms.items(): #For each test histogram
        results = {}
        for (k, hist) in train_histograms.items(): #For each training image histogram
            d = cv2.compareHist(hist1, hist, cv2.cv.CV_COMP_BHATTACHARYYA) #calculate ↩
                ....
            #  distance between test image histogram and training image histogram
            results[k] = d; #save distance
        results = sorted([(v,k) for (k,v) in results.items()], reverse=False) #sort  ↩
            results..
        # according to distance
# -----------------------------------------------------------------------
SECTION 7: Save Results
# -----------------------------------------------------------------------
        img_results[k1] = results[0][1] # get and save nearest neighbour
    labels = {'s':'Severe', 'n':'Normal', 'm':'Moderate', 't':'Trace', 'i':'Mild'}
    print "========= Results of " + folder + "experiment ==============="

# SECTION 8: Print each test image with corresponding Nearest Neighbour
# -----------------------------------------------------------------------
    for i, (k,v) in enumerate(img_results.items()): #for each test image and nearest ↩
         neighbour
        label = ''
        if v[:1].lower() in labels:
            label = labels[v[:1].lower()]
        print "Patient "+str(i+1)+": " + k + " => " + v + " (" +label+")\n" #print  ↩
            test image and nearest neighbour
    print "\n\n"
```

A.4 Appendix 4: Image Classification Results

```
=========  Results of cellulitis experiment ============

Patient 1: MD0_M17_TS_image 5 copy 4 task 4 cellu.jpg => MD0_M60_TR_image 11 copy 1  ↩
    task 4 cellu.jpg (Moderate)
Moderate = Moderate
=======================================================================
Patient 2: MD0_L32_TS_image 8 copy 6 task 4 cellu.jpg => MD0_M40_TR_image 8 copy 8 task 4 cellu.jpg (Moderate)
Moderate = Moderate
==============================================================================
Patient 3: S0_M13_TS_image 4 copy 8 task 4 cellu.jpg => S0_H08_TR_image 4 copy 1 task 4 cellu.png (Severe)
Severe = Severe
==============================================================================
Patient 4: S0_L10_TS_Image 2 copy number 3 cellu task 4.jpg => S0_L28_TR_image 7 copy 2 cellu task 4.jpg (Severe)
Severe = Severe
==============================================================================
Patient 5: S0_H03_TS_Image 2 copy number 9 cellu task 4.jpg => S0_H04_TS_Image 2 copy number 12 cellu task 4.jpg (Severe)
Severe = Severe
==============================================================================
Patient 6: TR0_M33_TS_image 7 copy 6 cellu task 4.jpg => TR0_L09_TR_Image 2 copy number 2 cellu task 4.jpg (Trace)
Trace = Trace
==============================================================================
Patient 7: MD0_M16_TS_image 5 copy 3 task 4 cellu.jpg => MD0_M19_TR_image 5 copy 9 task 4 cellu.jpg (Moderate)
Moderate = Moderate
==============================================================================

Patient 8: MD0_M76_TS_image 12 copy 10 task 4 cellu.jpg => MD0_M77_TR_image 12 copy 11 task 4 cellu.jpg (Moderate)
Moderate = Moderate
==============================================================================
Patient 9: MD0_L05_TS_Image 1 copy 6 cellu task 4.jpg => MD0_M02_TR_Image 1 copy 7 cellu task 4.jpg (Moderate)
Moderate = Moderate
==============================================================================
Patient 10: S0_H01_TS_Image 1 copy 8 cellu task 4.jpg => S0_M44_TS_image 9 copy 5 cellu task 4.jpg (Severe)
Severe = Severe
==============================================================================
Patient 11: S0_L13_TS_image 3 copy 4 cellu task 4.jpg => S0_L16_TR_image 3 copy 9 cellu task 4.jpg (Severe)
Severe = Severe
==============================================================================
Patient 12: MD0_M70_TS_image 12 copy 1 task 4 cellu.jpg => MD0_M77_TR_image 12 copy 11 task 4 cellu.jpg (Moderate)
Moderate = Moderate
==============================================================================
Patient 13: S0_H15_TS_image 7 copy 8 cellu task 4.jpg => S0_L30_TR_image 7 copy 10 cellu task 4.jpg (Severe)
```

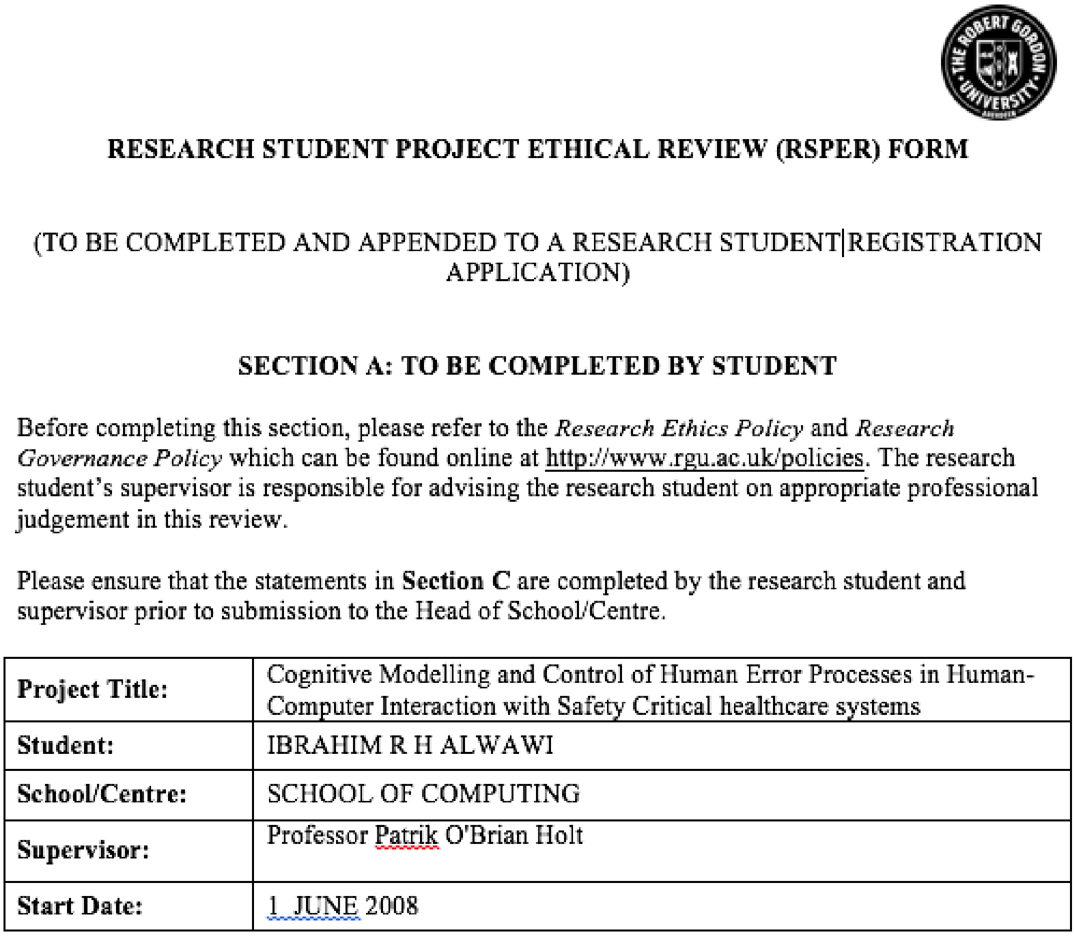
```
Severe = Severe
==============================================================================
Patient 14: MD0_M47_TS_image 9 copy 8 cellu task 4.jpg => MD0_M46_TR_image 9 copy 7 cellu task 4.jpg (Moderate)
Moderate = Moderate
==============================================================================
Patient 15: TR0_M35_TS_image 7 copy 12 cellu task 4.jpg => TR0_L09_TR_Image 2 copy number 2 cellu task 4.jpg (Trace)
Trace = Trace
==============================================================================
Patient 16: MD0_L35_TS_image 9 copy 2 cellu task 4.jpg => MD0_M46_TR_image 9 copy 7 cellu task 4.jpg (Moderate)
Moderate = Moderate
==============================================================================
Patient 17: S0_M01_TS_Image 1 copy 5 cellu task 4.jpg => S0_L07_TS_Image 1 copy 12 cellu task 4.jpg (Severe)
Severe = Severe
=============================================================================
Patient 18: MD0_M64_TS_image 11 copy 7 task 4 cellu.jpg => MD0_M62_TS_image 11 copy 4 task 4 cellu.jpg (Moderate)
Moderate = Moderate
==============================================================================
Patient 19: TR0_M31_TS_image 7 copy 3 cellu task 4.jpg => TR0_L09_TR_Image 2 copy number 2 cellu task 4.jpg (Trace)
Trace = Trace
==============================================================================
Patient 20: TR0_L23_TS_image 4 copy 11 task 4 cellu.jpg => TR0_L22_TR_image 4 copy 9 task 4 cellu.jpg (Trace)
Trace = Trace
==============================================================================
Patient 21: MD0_M27_TS_image6 copy 7 task 4 cellu.jpg => MD0_M30_TS_image 6 copy 11 task 4 cellu.jpg (Moderate)
Moderate = Moderate
==============================================================================

Patient 22: S0_M08_TS_Image 2 copy number 10 cellu task 4.jpg => S0_H17_TR_image 7 copy 11 cellu task 4.jpg (Severe)
Severe = Severe
==============================================================================
Patient 23: MD0_M11_TS_image 3 copy 5 cellu task 4.jpg => MD0_M12_TS_image 3 copy 10 cellu task 4.jpg (Moderate)
Moderate = Moderate
==============================================================================
Patient 24: S0_H11_TS_image 5 copy 8 task 4 cellu.jpg => S0_M24_TR_image 6 copy 4 task 4 cellu.jpg (Severe)
Severe = Severe
==============================================================================
Patient 25: MDO_H23_TS_image 11 copy 3 task 4 cellu.jpg => MDO_M66_TR_image 11 copy 9 task 4 cellu.jpg (Moderate)
Moderate = Moderate
==============================================================================
Patient 26: SO_L08_TS_image2 copy number 1 cellu task 4.jpg => SO_H17_TR_image 7 copy 11 cellu task 4.jpg (Severe)
Severe = Severe
==============================================================================
Patient 27: MDO_L11_TS_Image 2 copy number 8 cellu task 4.jpg => MDO_M07_TR_Image 2 copy number 7 cellu task 4.jpg (Moderate)
Moderate = Moderate
==============================================================================
Patient 28: MDO_M45_TS_image 9 copy 6 cellu rask 4.jpg => MDO_M43_TS_image 9 copy 4 cellu task 4.jpg (Moderate)
Moderate = Moderate
==============================================================================
Patient 29: MDO_M03_TS_Image 1 copy 9 cellu task 4.jpg => MDO_M02_TR_Image 1 copy 7 cellu task 4.jpg (Moderate)
Moderate = Moderate
==============================================================================
Patient 30: SO_H21_TS_image 9 copy 11 cellu task 4.jpg => SO_M50_TR_image 9 copy 12 cellu task 4.jpg (Severe)
Severe = Severe
==============================================================================
Patient 31: SO_M34_TS_image 7 copy 7 cellu task 4.jpg => SO_H17_TR_image 7 copy 11 cellu task 4.jpg (Severe)
Severe = Severe
==============================================================================
Patient 32: SO_L34_TS_image 9 copy 1 cellu task 4.jpg => SO_M44_TS_image 9 copy 5 cellu task 4.jpg (Severe)
Severe = Severe
==============================================================================
Patient 33: TRO_M72_TS_image 12 copy 5 task 4 cellu.jpg => TRO_M75_TR_image 12 copy 9 task 4 cellu.jpg (Trace)
Trace = Trace
==============================================================================
Patient 34: MDO_L33_TS_image 8 copy 9 task 4 cellu.jpg => MDO_M43_TS_image 9 copy 4 cellu task 4.jpg (Moderate)
Moderate = Moderate
==============================================================================


Patient 35: SO_M21_TS_image 5 copy 12 task 4 cellu.jpg => SO_M24_TR_image 6 copy 4 task 4 cellu.jpg (Severe)
Severe = Severe
==============================================================================

Patient 36: MD0_L25_TS_image 5 copy 7 task 4 cellu.jpg => MD0_M30_TS_image 6 copy 11 task 4 cellu.jpg (Moderate)
Moderate = Moderate
==============================================================================
Patient 37: S0_H12_TS_image 6 copy 3 task 4 cellu.jpg => S0_M22_TR_image 6 copy 1 task 4 cellu.jpg (Severe)
Severe = Severe
==============================================================================
Patient 38: S0_M37_TS_image 8 copy 2 task 4 cellu.png => S0_M41_TS_image 8 copy 10 task 4 cellu.png (Severe)
Severe = Severe
==============================================================================
Patient 39: TR0_L03_TS_Image 1 copy 3 cellu task 4.jpg => TR0_L01_TR_Image 1 copy 1 cellu task 4.jpg (Trace)
Trace = Trace
==============================================================================
Patient 40: MD0__M38_TS_image 8 copy 3 task 4 cellu.jpg => MD0_M40_TR_image 8 copy 8 task 4 cellu.jpg (Moderate)
Moderate = Moderate
==============================================================================
Patient 41: S0_L12_TS_image 3 copy 3 cellu task 4.jpg => S0_L16_TR_image 3 copy 9 cellu task 4.jpg (Severe)
Severe = Severe
==============================================================================
Patient 42: TR0_M36_TS_image 8 copy 1 task 4 cellu.png => TR0_L31_TS_image 8 copy 5 task 4 cellu.jpg (Trace)
Trace = Trace
==============================================================================
Patient 43: S0_M53_TS_image 10 copy 4 task 4 cellu.jpg => S0_H06_TR_image 3 copy 8 cellu task 4.jpg (Severe)
Severe = Severe
==============================================================================
Patient 44: S0_H10_TS_image 4 copy 5 task 4 cellu.jpg => S0_H09_TR_image 4 S_copy 2 task 4 cellu.png (Severe)
Severe = Severe
==============================================================================
Patient 45: MD0_M39_TS_image 8 copy 7 task 4 cellu.jpg => MD0_M40_TR_image 8 copy 8 task 4 cellu.jpg (Moderate)
Moderate = Moderate
==============================================================================
Patient 46: TR0_L18_TS_image 4 copy 3 task 4 cellu.jpg => TR0_L19_TR_image 4 copy 4 task 4 cellu.jpg (Trace)
Trace = Trace
==============================================================================
Patient 47: S0_M32_TS_image 7 copy 5 cellu task 4.jpg => S0_H16_TR_image 7 copy 9 cellu task 4.jpg (Severe)
Severe = Severe
```

```
===============================================================================


Patient 48: S0_L36_TS_image 10 copy 3 task 4 cellu.jpg => S0_H14_TR_image 6 copy 12 task 4 cellu.jpg (Severe)
Severe = Severe
===============================================================================
Patient 49: S0_L40_TS_image 12 copy 6 task 4 cellu.jpg => S0_M74_TR_image 12 copy 8 task 4 cellu.jpg (Severe)
Severe = Severe
===============================================================================

Patient 50: S0_L29_TS_image 7 copy 4 cellu task4.jpg => S0_H17_TR_image 7 copy 11 cellu task 4.jpg (Severe)
Severe = Severe
===============================================================================
Patient 51: TR0_H18_TS_image 8 copy 4 task 4 cellu.jpg => TR0_L31_TS_image 8 copy 5 task 4 cellu.jpg (Trace)
Trace = Trace
===============================================================================
Patient 52: TR0_M65_TS_image 11 copy 8 task 4 cellu.jpg => TR0_M26_TS_image 6 copy 6 task 4 cellu.jpg (Trace)
Trace = Trace
===============================================================================
Patient 53: S0_H02_TS_Image 1 copy 11 cellu task 4.jpg => S0_M44_TS_image 9 copy 5 cellu task 4.jpg (Severe)
Severe = Severe
===============================================================================
Patient 54: S0_M48_TS_image 9 copy 9 cellu task 4.jpg => S0_M44_TS_image 9 copy 5 cellu task 4.jpg (Severe)
Severe = Severe
===============================================================================
Patient 55: MD0_L38_TS_image 11 copy 5 task 4 cellu.jpg => MD0_M66_TR_image 11 copy 9 task 4 cellu.jpg (Moderate)
Moderate = Moderate
===============================================================================
Patient 56: S0_L27_TS_image 7 copy 1 cellu task 4.jpg => S0_H17_TR_image 7 copy 11 cellu task 4.jpg (Severe)
Severe = Severe
===============================================================================
Patient 57: S0_M51_TS_image 10 copy 1 task 4 cellu.jpg => S0_H06_TR_image 3 copy 8 cellu task 4.jpg (Severe)
Severe = Severe
===============================================================================
Patient 58: S0_M19_TS_image 5 copy 6 task 4 cellu.jpg => S0_M14_TR_image 5 copy 1 task 4 cellu.jpg (Severe)
Severe = Severe
```

```
==============================================================================
Patient 59: S0__H19_TS_image 8 copy 11 task 4 cellu.png => S0_M41_TS_image 8 copy 10 task 4 cellu.png (Severe)
Severe = Severe
==============================================================================
Patient 60: S0_M18_TS_image 5 copy 5 task 4 cellu.jpg => S0_M15_TR_image 5 copy 2 task 4 cellu.jpg (Severe)
Severe = Severe
==============================================================================


Patient 61: TR0_H24_TS_image 12 copy 3 task 4 cellu.jpg => TR0_M73_TR_image 12 copy 7 task 4 cellu.jpg (Trace)
Trace = Trace
==============================================================================
Patient 62: TR0_L04_TS_Image 1 copy 4 cellu task 4.jpg => TR0_L02_TR_Image 1 copy 2 cellu task 4.jpg (Trace)
Trace = Trace
==============================================================================
Patient 63: S0_H05_TS_image 3 copy 1 cellu task 4.jpg => S0_H14_TR_image 6 copy 12 task 4 cellu.jpg (Severe)
Severe = Severe
==============================================================================

Patient 64: TR0_M29_TS_image 6 copy 10 task 4 cellu.jpg => TR0_M56_TS_image 10 copy 7 task 4 cellu.jpg (Trace)
Trace = Trace
==============================================================================
Patient 65: TR0_L39_TS_image 12 copy 4 task 4 cellu.jpg => TR0_M73_TR_image 12 copy 7 task 4 cellu.jpg (Trace)
Trace = Trace
==============================================================================
Patient 66: S0_L37_TS_image 10 copy 8 task 4 cellu.jpg => S0_MD0_M25_TR_image 6 copy 5 task 4 cellu.jpg (Severe)
Severe = Severe
==============================================================================



=========  Results of eye experiment ==========

Patient 1: S014_TS.jpg => S019_TR_exp2A image1 Q1_MODERATE.jpg (Severe)
Severe = Severe
==============================================================================
Patient 2: N07_TS_exp2A image3 Q3_NONE.jpg => N010_TR_ exp1A Q7_MILD.jpg (Normal)
Normal = Normal
==============================================================================
Patient 3: S09_TS.jpg => S015_TR_ exp1D Q2_SEVERE.jpg (Severe)
Severe = Severe
==============================================================================
Patient 4: s05_TS.jpg => S04_TR.jpg (Severe)
Severe = Severe
==============================================================================
Patient 5: MD02_TS_exp2A image4 Q4_MODERATE.jpg => MD011_TR_ exp1D Q8_MODERATE.jpg (Moderate)
Moderate = Moderate
==============================================================================
Patient 6: s04_TS.jpg => S011_TR.jpg (Severe)
Severe = Severe
==============================================================================
Patient 7: I05_TS_exp2B image12 Q12_MILD.jpg => N010_TR_ exp1A Q7_MILD.jpg (Normal)
Mild = Normal
==============================================================================

Patient 8: S016_TS_ exp1D Q3_MODERATE.jpg => S019_TR_exp2A image1 Q1_MODERATE.jpg (Severe)
Severe = Severe
==============================================================================
Patient 9: N017_TS_ exp1A Q1_NONE.jpg => N04_TR.png (Normal)
Normal = Normal
==============================================================================
Patient 10: TR011_TS (3).jpg => N011_TR.jpg (Normal)
Trace = Normal
==============================================================================
Patient 11: TR03_TS_exp2D image2 Q2_TRACE.jpg => I02_TR_exp2A image2 Q2_MILD.jpg (Mild)
Trace = Mild
==============================================================================
Patient 12: S08_TS_ exp1E Q5_SEVERE.jpg => S018_TR_ exp1C Q5_SEVERE.jpg (Severe)
Severe = Severe
==============================================================================
Patient 13: N08_TS_exp2A image5 Q5_NONE.jpg => N04_TR.png (Normal)
Normal = Normal
==============================================================================
Patient 14: I011_TS (2).jpg => N011_TR_exp2B image10 Q10_NONE.jpg (Normal)
Mild = Normal
==============================================================================
Patient 15: N018_TS_ exp1A Q5_NONE.jpg => No1_TR.jpg (Normal)
Normal = Normal
==============================================================================
Patient 16: MD011_TS.jpg => MD011_TR_ exp1D Q8_MODERATE.jpg (Moderate)
Moderate = Moderate
==============================================================================
Patient 17: S014_TS_ exp1D Q6_SEVERE.jpg => s06_TR.png (Severe)
```

```
Severe = Severe
==============================================================================
Patient 18: S015_TS.jpg => S012_TR_exp2D image5 Q5_SEVERE.jpg (Severe)
Severe = Severe
==============================================================================
Patient 19: TR06_TS_exp2B image2 Q2_TRACE.jpg => I02_TR_exp2A image2 Q2_MILD.jpg (Mild)
Trace = Mild
==============================================================================
Patient 20: N09_TS_exp2B image1 Q1_NONE.jpg => S015_TR_ exp1D Q2_SEVERE.jpg (Severe)
None = Severe
==============================================================================
Patient 21: N021_TS_ exp1C Q3_NONE.jpg => N010_TR_ exp1A Q7_MILD.jpg (Normal)
Normal = Normal
==============================================================================
Patient 22: MD017_TS.jpg => MD04_TR_exp2B image4 Q4_MODERATE.jpg (Moderate)
Moderate = Moderate
==============================================================================
Patient 23: TR012_TS (2).jpg => No1_TR.jpg (Normal)
Trace = Normal
==============================================================================
Patient 24: TR09_TS.jpg => N011_TR.jpg (Normal)
Trace = Normal
==============================================================================
Patient 25: s01_TS.jpg => I02_TR_exp2A image2 Q2_MILD.jpg (Mild)
Severe = Mild

==============================================================================
```

A.5 Appendix 5: RGU Ethical Approval Form



RESEARCH STUDENT PROJECT ETHICAL REVIEW (RSPER) FORM

(TO BE COMPLETED AND APPENDED TO A RESEARCH STUDENT|REGISTRATION APPLICATION)

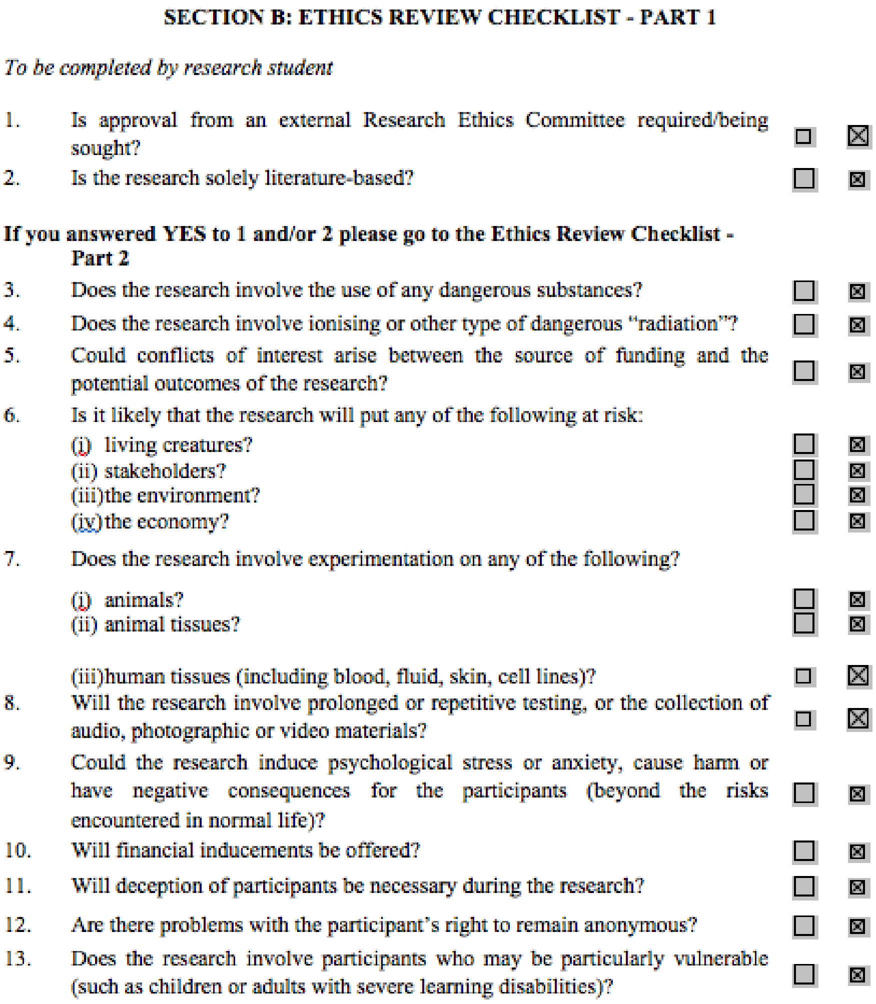
SECTION A: TO BE COMPLETED BY STUDENT

Before completing this section, please refer to the *Research Ethics Policy* and *Research Governance Policy* which can be found online at http://www.rgu.ac.uk/policies. The research student's supervisor is responsible for advising the research student on appropriate professional judgement in this review.

Please ensure that the statements in **Section C** are completed by the research student and supervisor prior to submission to the Head of School/Centre.

Project Title:	Cognitive Modelling and Control of Human Error Processes in Human-Computer Interaction with Safety Critical healthcare systems
Student:	IBRAHIM R H ALWAWI
School/Centre:	SCHOOL OF COMPUTING
Supervisor:	Professor Patrik O'Brian Holt
Start Date:	1 JUNE 2008

Figure A.6: Section A of RGU Ethical Approval Form



SECTION B: ETHICS REVIEW CHECKLIST - PART 1

To be completed by research student

1. Is approval from an external Research Ethics Committee required/being sought? □ ☒

2. Is the research solely literature-based? □ ☒

If you answered YES to 1 and/or 2 please go to the Ethics Review Checklist - Part 2

3. Does the research involve the use of any dangerous substances? □ ☒

4. Does the research involve ionising or other type of dangerous "radiation"? □ ☒

5. Could conflicts of interest arise between the source of funding and the potential outcomes of the research? □ ☒

6. Is it likely that the research will put any of the following at risk:

(i) living creatures? □ ☒
(ii) stakeholders? □ ☒
(iii)the environment? □ ☒
(iv)the economy? □ ☒

7. Does the research involve experimentation on any of the following?

(i) animals? □ ☒
(ii) animal tissues? □ ☒

(iii)human tissues (including blood, fluid, skin, cell lines)? □ ☒

8. Will the research involve prolonged or repetitive testing, or the collection of audio, photographic or video materials? □ ☒

9. Could the research induce psychological stress or anxiety, cause harm or have negative consequences for the participants (beyond the risks encountered in normal life)? □ ☒

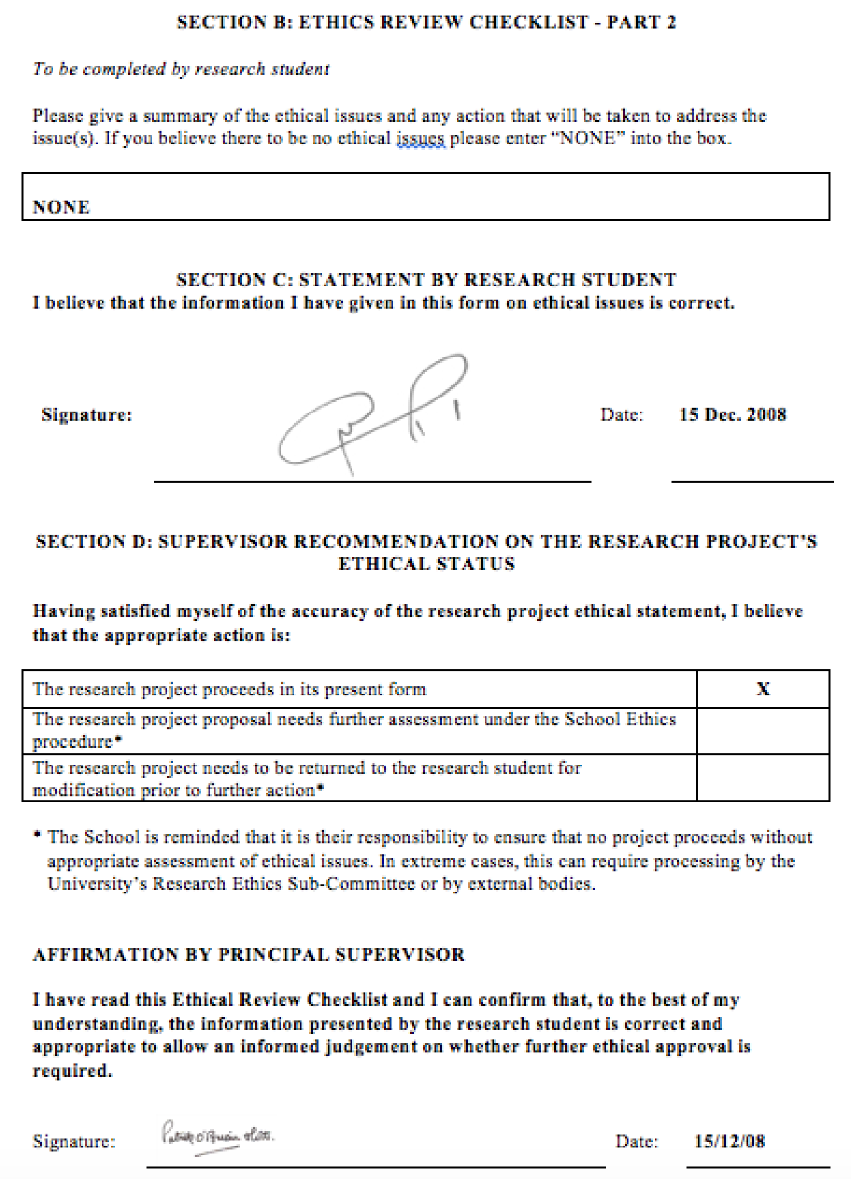
10. Will financial inducements be offered? □ ☒

11. Will deception of participants be necessary during the research? □ ☒

12. Are there problems with the participant's right to remain anonymous? □ ☒

13. Does the research involve participants who may be particularly vulnerable (such as children or adults with severe learning disabilities)? □ ☒

Figure A.7: Section B of RGU Ethical Approval Form



SECTION B: ETHICS REVIEW CHECKLIST - PART 2

To be completed by research student

Please give a summary of the ethical issues and any action that will be taken to address the issue(s). If you believe there to be no ethical issues please enter "NONE" into the box.

NONE

SECTION C: STATEMENT BY RESEARCH STUDENT

I believe that the information I have given in this form on ethical issues is correct.

Signature: Date: **15 Dec. 2008**

SECTION D: SUPERVISOR RECOMMENDATION ON THE RESEARCH PROJECT'S ETHICAL STATUS

Having satisfied myself of the accuracy of the research project ethical statement, I believe that the appropriate action is:

The research project proceeds in its present form	X
The research project proposal needs further assessment under the School Ethics procedure*	
The research project needs to be returned to the research student for modification prior to further action*	

* The School is reminded that it is their responsibility to ensure that no project proceeds without appropriate assessment of ethical issues. In extreme cases, this can require processing by the University's Research Ethics Sub-Committee or by external bodies.

AFFIRMATION BY PRINCIPAL SUPERVISOR

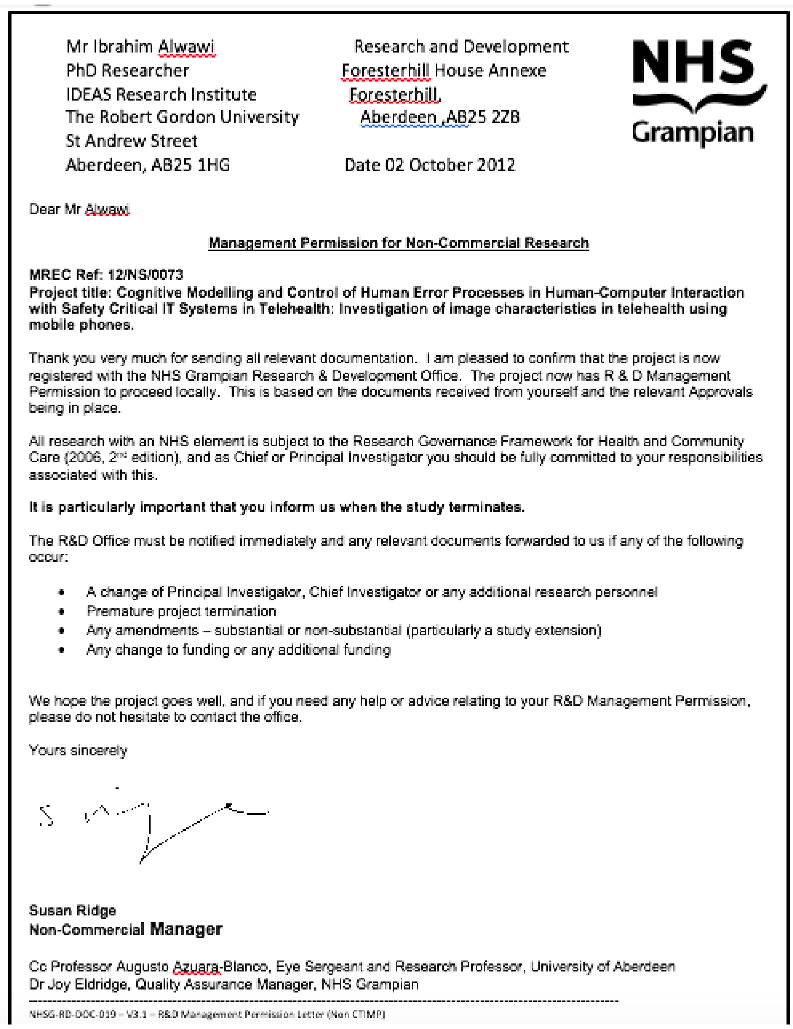
I have read this Ethical Review Checklist and I can confirm that, to the best of my understanding, the information presented by the research student is correct and appropriate to allow an informed judgement on whether further ethical approval is required.

Signature: Date: **15/12/08**

Figure A.8: Section C and D of RGU Ethical Approval Form

INSTRUCTIONS FOR RESEARCH STUDENT:

Once the School is satisfied with the ethical check surrounding your research work, please attach original signed copy of this form to your Registration Application Form (RDR). Once your RDR form is complete, signed and has all appropriate attachments, you should then forward it to the Research Degrees Office, AB44, and Schoolhill.



Mr Ibrahim Alwawi
PhD Researcher
IDEAS Research Institute
The Robert Gordon University
St Andrew Street
Aberdeen, AB25 1HG

Research and Development
Foresterhill House Annexe
Foresterhill,
Aberdeen ,AB25 2ZB

Date 02 October 2012

NHS
Grampian

Dear Mr Alwawi

Management Permission for Non-Commercial Research

MREC Ref: 12/NS/0073
Project title: Cognitive Modelling and Control of Human Error Processes in Human-Computer Interaction with Safety Critical IT Systems in Telehealth: Investigation of image characteristics in telehealth using mobile phones.

Thank you very much for sending all relevant documentation. I am pleased to confirm that the project is now registered with the NHS Grampian Research & Development Office. The project now has R & D Management Permission to proceed locally. This is based on the documents received from yourself and the relevant Approvals being in place.

All research with an NHS element is subject to the Research Governance Framework for Health and Community Care (2006, 2nd edition), and as Chief or Principal Investigator you should be fully committed to your responsibilities associated with this.

It is particularly important that you inform us when the study terminates.

The R&D Office must be notified immediately and any relevant documents forwarded to us if any of the following occur:

- A change of Principal Investigator, Chief Investigator or any additional research personnel
- Premature project termination
- Any amendments – substantial or non-substantial (particularly a study extension)
- Any change to funding or any additional funding

We hope the project goes well, and if you need any help or advice relating to your R&D Management Permission, please do not hesitate to contact the office.

Yours sincerely

Susan Ridge
Non-Commercial Manager

Cc Professor Augusto Azuara-Blanco, Eye Sergeant and Research Professor, University of Aberdeen
Dr Joy Eldridge, Quality Assurance Manager, NHS Grampian

NHSG-RD-DOC-019 – V3.1 – R&D Management Permission Letter (Non CTIMP)

Figure A.9: Management Permission

A.6 Appendix 6: Example instruction and answer sheets

Experiment No.1: “Testing Human Colour Perception”

Task 1.1 Describe the following images based on their colours:

Description:

____

____

____

A.7 Appendix 7.1: Consent Form (Content Only)

Following are the content of the Consent Form.

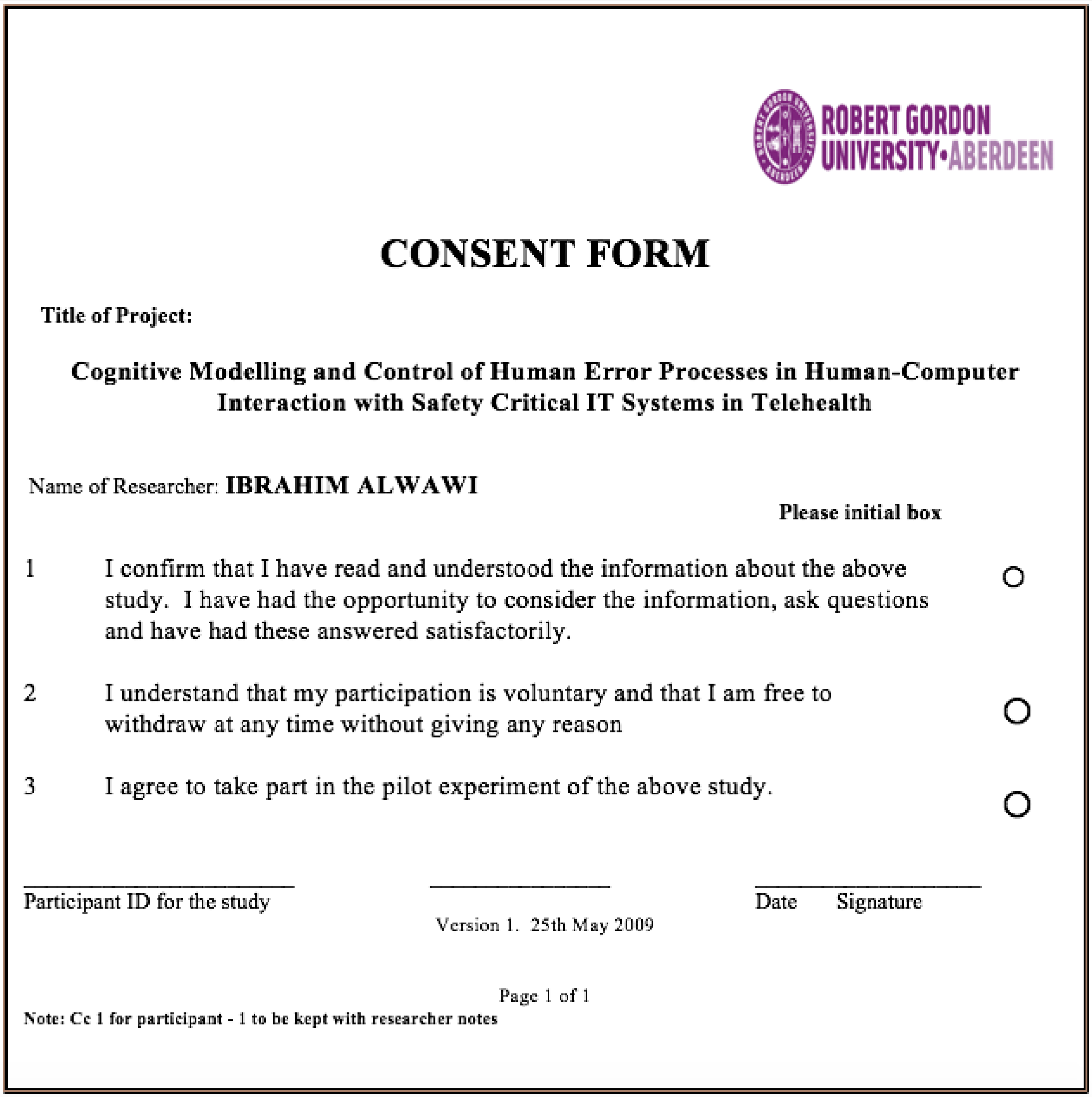
Title of Project: **Cognitive Modelling and Control of Human Error Processes in Human-Computer Interaction with Safety Critical IT Systems in Telehealth**

Name of Researcher: IBRAHIM ALWAWI.

There are three options in this consent form; as following:

1. I confirm that I have read and understood the information about the above study. I have had the opportunity to consider the information, ask questions and have had these answered satisfactorily.
2. I understand that my participation is voluntary and that I am free to withdraw at any time without giving any reason
3. I agree to take part in the pilot experiment of the above study.

The picture version of the Consent Form can be seen in Figure A.10



ROBERT GORDON UNIVERSITY•ABERDEEN

CONSENT FORM

Title of Project:

Cognitive Modelling and Control of Human Error Processes in Human-Computer Interaction with Safety Critical IT Systems in Telehealth

Name of Researcher: **IBRAHIM ALWAWI**

Please initial box

1 I confirm that I have read and understood the information about the above study. I have had the opportunity to consider the information, ask questions and have had these answered satisfactorily. ○

2 I understand that my participation is voluntary and that I am free to withdraw at any time without giving any reason ○

3 I agree to take part in the pilot experiment of the above study. ○

_________________________ _________________ _____________________

Participant ID for the study Date Signature

Version 1. 25th May 2009

Page 1 of 1

Note: Cc 1 for participant - 1 to be kept with researcher notes

Figure A.10: Consent Form

A.8 Appendix 7.2: NHS Consent Form (Content Only)

Following are the content of the NHS Consent Form.

Patient name: ____
Patient label: ____
Unit number: ____

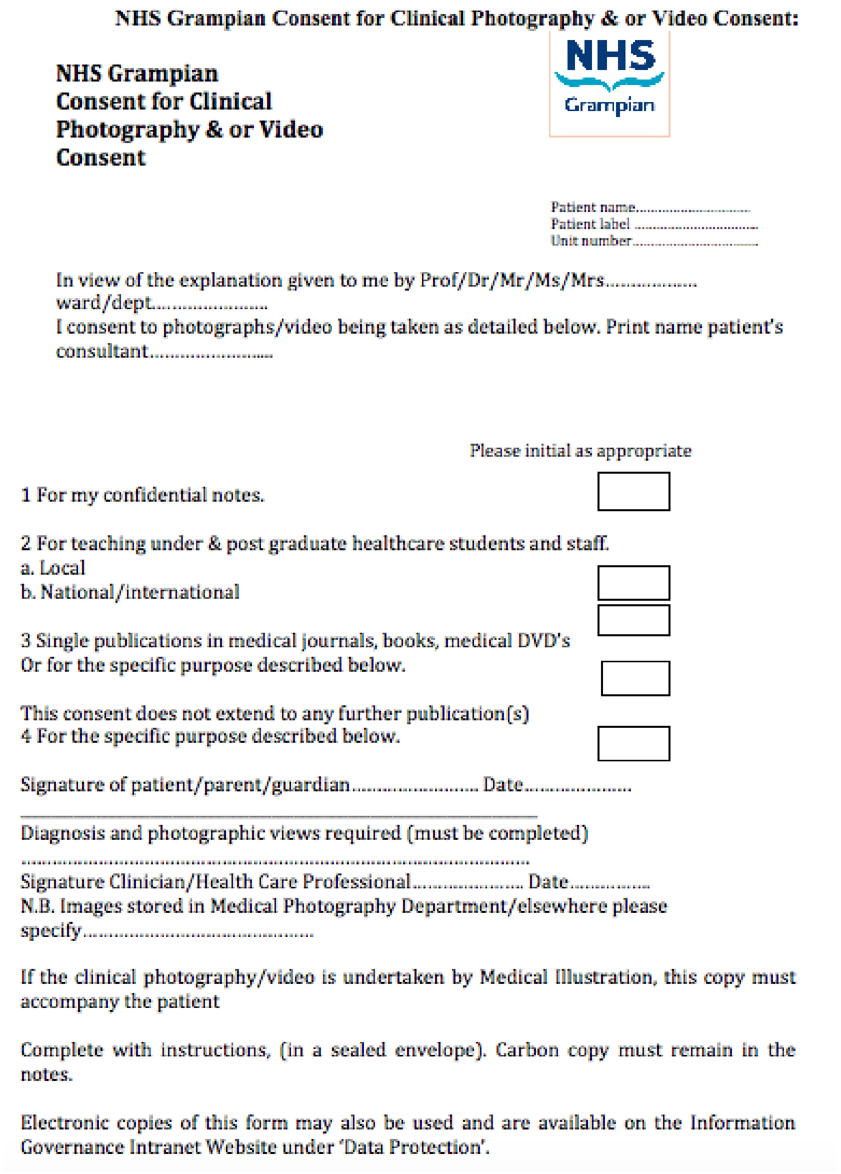
In view of the explanation given to me by Prof/Dr/Mr/Ms/Mrs_________________________

ward/dept_________________________ I consent to photographs/video being taken as detailed below.
Print name patient's consultant _________________________

Please initial as appropriate

1. For my confidential notes.
2. For teaching under & post graduate healthcare students and staff (Local or National/international)
3. Single publications in medical journals, books, medical DVD's or for the specific purpose described below.
4. This consent does not extend to any further publication(s) for the specific purpose described below.

The picture version of the NHS Consent Form can be seen in Figure A.11



NHS Grampian Consent for Clinical Photography & or Video Consent:

NHS Grampian
Consent for Clinical
Photography & or Video
Consent

NHS
Grampian

Patient name................................
Patient label ....
Unit number....................................

In view of the explanation given to me by Prof/Dr/Mr/Ms/Mrs..................
ward/dept......................
I consent to photographs/video being taken as detailed below. Print name patient's consultant.........................

Please initial as appropriate

1 For my confidential notes. □

2 For teaching under & post graduate healthcare students and staff.
a. Local □
b. National/international □

3 Single publications in medical journals, books, medical DVD's
Or for the specific purpose described below. □

This consent does not extend to any further publication(s)
4 For the specific purpose described below. □

Signature of patient/parent/guardian......................... Date.....................

Diagnosis and photographic views required (must be completed)
....
Signature Clinician/Health Care Professional...................... Date................
N.B. Images stored in Medical Photography Department/elsewhere please specify...

If the clinical photography/video is undertaken by Medical Illustration, this copy must accompany the patient

Complete with instructions, (in a sealed envelope). Carbon copy must remain in the notes.

Electronic copies of this form may also be used and are available on the Information Governance Intranet Website under 'Data Protection'.

Figure A.11: NHS Consent Form

A.9 Appendix 8: Colour Blind Test

Experiment No.2: "Testing Human Colour Perception" Colour Blindness Test: Write the Number that you can recognise from the given image A.12, page1

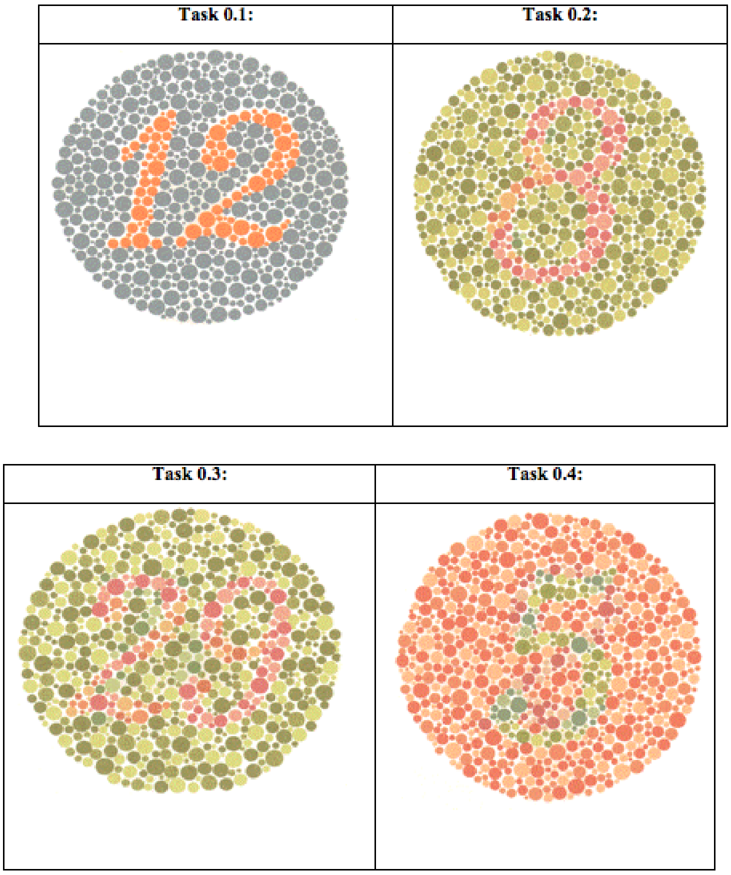
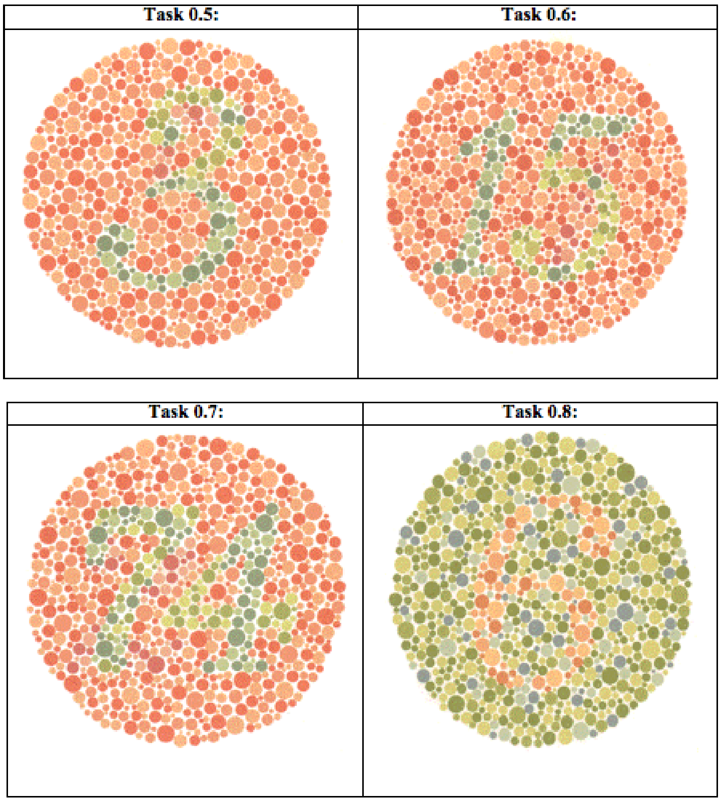


Figure A.12: Colour Blind Test Part 1

Write the Number that you can recognise from the given image A.13, page2



Task 0.5:

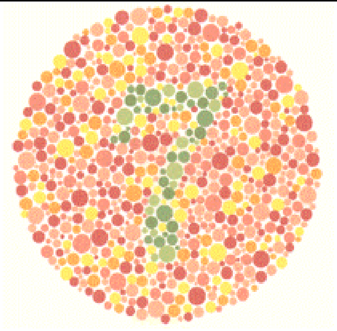
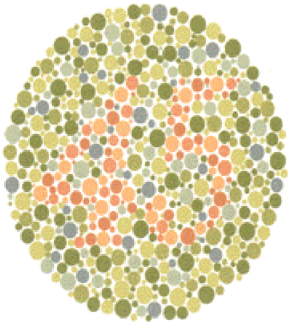
Task 0.6:

Task 0.7:

Task 0.8:

Figure A.13: Colour Blind Test Part 2

Write the Number that you can recognise from the given image A.14, page3

Task 0.9:	Task 0.10:
	

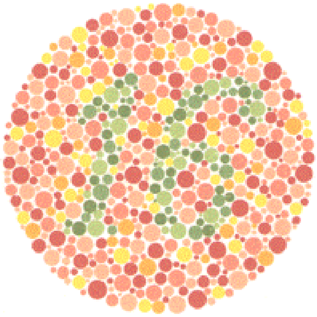
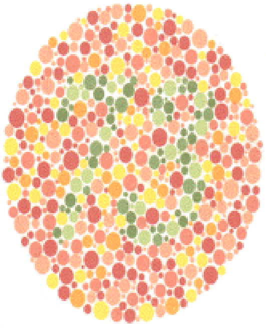
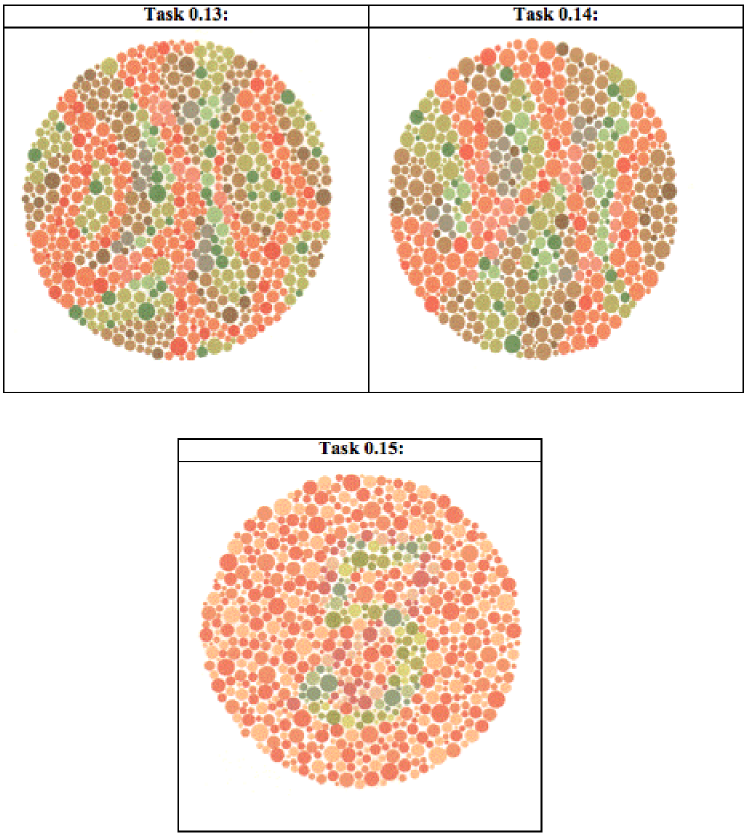
Task 0.11:	Task 0.12:
	

Figure A.14: Colour Blind Test Part 3

Write the Number that you can recognise from the given image A.15, page4



Task 0.13:

Task 0.14:

Task 0.15:

Figure A.15: Colour Blind Test Part 4

User Reference Number: ________________

Experiment No.1: “Testing Human Colour Perception”

Colour Blindness Test Answers Sheet:

Task 0: Write the Number that you can recognise from the given image.

Table A.6: Classification System

Plate number	Normal Vision (number seen)
1	
2	
3	
4	
5	
6	
7	
8	
9	
10	
11	
12	
13	
14	
15	

Answers for Colour Blindness Test Below is a table of what a person with normal vision should have seen.

Table A.7: Classification System

Plate number	Normal Vision (number seen)
1	12
2	8
3	29
4	5
5	3
6	15
7	74
8	6
9	7
10	45
11	16
12	73
13	None
14	None
15	5

Please note that this test should not be used as a diagnosis; please consult your doctor for further testing. Various factors could have affected the result you got such as the quality and the color settings of your monitor. Please only use this test as guidance. The test can be found online: [http://colourblind.freeservers.com/results.htm, accessed on January 2012]

A.10 Appendix 9: Instruction sheets

Study No.6: "Testing Human Colour Perception" **Dear user,**
Iwill be verygrateful if you could donate 5 minutes for me by doing myexperiment for my PhD project inTelehealth which you may find veryinteresting and a bit of fun. please click on the followinglink and to access to theexperiment youruser ID in the first page is http://www.comp.rgu.ac.uk/staff/iba/start.htm

Many thanks
Researcher: Ibrahim Alwawi
Principal supervisor: Professor Patrik O' Brian Holt
General Background
This online experiment is being conducted by Ibrahim Alwawi. I am a PhD Research

Student doing my project in Telehealth and Medical Imaging.
This project is in partnership with NHS24 and the A&E Department in Aberdeen Royal Infirmary ARI and the Robert Gordon University RGU.

The research focuses on human colour perception when diagnosing skin infection and cellulitis. This involves mainly doctors and any other related healthcare experts. Also it involves non-health workers for comparison and analysis.

The project requires data from at least 30 participants. You have been invited to contribute based on your experience and knowledge in colour matching. Telehealth involves medical diagnosis and treatment through telecommunication systems and devices typically involving images, video and sound. This is a special case of complex safety-critical human computer interaction where cognitive overload can result in errors of judgment.

This project involves analysis and comparing images to examine how these errors may occur. This Experiment (Image Quality Matching No.5)

This is a Telehealth experiment focusing on an Image Matching Test which consists of series of 25 images in pairs for you, the user, to compare then select your answers based on whether you feel the images are the SAME, DIFFERENT or you are NOT SURE within the shortest possible time.

You also get to indicate how confident you are in your decisions. In this experiment there is an interval page serving as a break in between the pages.

The Objective

To test the level of accuracy of users in image quality matching and make comparisons between data from health workers and data from non-health workers.

Remember

Your data is an important contribution to the academic research in the areas of Telehealth and Cognitive Engineering for improved healthcare services. Instructions

- Fill in the required details in the form below.
- Click on the START button to begin the experiment.
- Compare the pair of images displayed and select from the options below them 'SAME' if you think they are the same, 'DIFFERENT' if you think the images are different, or 'NOT SURE' if you are not sure if the images are the same or different.

- Then make a Confidence Level selection from 1 to 9 based on how sure you are of your answer above.
- Click on the NEXT button to move to the next page.
- You can take breaks only between the pages as the timing stops on the interval pages between images.
- Please click on the FINISH button at the end of the experiment in order to save your results.
- Please DO NOT click on the BACK or REFRESH buttons on your browser during the experiment as this will return you to the Start page.
- You have the right to withdraw from the experiment at any time.
- The user data is required and used only for research purposes.

JavaScript should not be turned off in your browser settings for this experiment. **End of Introduction** Thank You

A.11 Appendix 10: Sample of master data for image matching experiment-Chapter 6

User ID: 382_278
Gender: Female
Age: 16 - 20
Education: Undergraduate
Eyesight: Normal (not using glasses)
Colour Blindness: No
Occupation: Non-Dr
Year registered as a Doctor: Not Applicable
Experience with cellulitis/infections: Not Applicable

Table A.8: Participants master data-Non-Dr female, example1

Test No	Time Taken(secs)	User Result	System Result	Cogtool Time	Confidence
1	10.074	SAME	DIFFERENT	2.495	7
2	10.456	SAME	DIFFERENT	2.495	6
3	10.833	SAME	DIFFERENT	2.495	7
4	3.433	DIFFERENT	DIFFERENT	2.495	9
5	6.988	SAME	SAME	2.538	6
6	12.05	SAME	SAME	2.538	8
7	3.743	DIFFERENT	DIFFERENT	2.495	8
8	7.262	DIFFERENT	DIFFERENT	2.495	8
9	2.945	DIFFERENT	DIFFERENT	2.495	8
10	4.227	SAME	DIFFERENT	2.495	7
11	7.165	DIFFERENT	DIFFERENT	2.495	7
12	3.833	DIFFERENT	DIFFERENT	2.495	8
13	2.415	DIFFERENT	DIFFERENT	2.495	8
14	2.737	DIFFERENT	DIFFERENT	2.495	7
15	3.17	SAME	SAME	2.538	7
16	4.354	SAME	SAME	2.538	7
17	3.107	DIFFERENT	DIFFERENT	2.495	9
18	7.583	DIFFERENT	DIFFERENT	2.495	7
19	3.136	DIFFERENT	DIFFERENT	2.495	6
20	2.041	DIFFERENT	DIFFERENT	2.495	9
21	3.359	DIFFERENT	DIFFERENT	2.495	8
22	7.473	SAME	DIFFERENT	2.495	8
23	3.012	SAME	DIFFERENT	2.495	8
24	5.141	DIFFERENT	DIFFERENT	2.495	8
	3.334	SAME	SAME	2.538	7
Total Time = 133.87099999999998			Accuracy = 76 %		

User ID: 382_914

Gender: Male

Age: 31 - 40

Education: Postgraduate

Eyesight: Normal (not using glasses)

Colour Blindness: No

Occupation: Non-Dr

Year registered as a Doctor: Not Applicable

Experience with cellulitis/infections: Not Applicable

Table A.9: Participants master data-Non-Dr male, example2

Test No	Time Taken(secs)	User Result	System Result	Cogtool Time	Confidence
1	39.74	SAME	DIFFERENT	2.495	8
2	32.929	SAME	DIFFERENT	2.495	5
3	52.457	SAME	DIFFERENT	2.495	7
4	12.497	DIFFERENT	DIFFERENT	2.495	9
5	24.961	SAME	SAME	2.538	7
6	138.289	SAME	SAME	2.538	8
7	17.673	SAME	DIFFERENT	2.495	6
8	37.129	SAME	DIFFERENT	2.495	8
9	11.417	DIFFERENT	DIFFERENT	2.495	8
10	7.409	SAME	DIFFERENT	2.495	8
11	8.193	DIFFERENT	DIFFERENT	2.495	9
12	3.745	DIFFERENT	DIFFERENT	2.495	9
13	4.906	DIFFERENT	DIFFERENT	2.495	7
14	5.609	DIFFERENT	DIFFERENT	2.495	9
15	12.641	SAME	SAME	2.538	8
16	8.345	SAME	SAME	2.538	8
17	4.953	DIFFERENT	DIFFERENT	2.495	7
18	4.145	DIFFERENT	DIFFERENT	2.495	9
19	6.937	DIFFERENT	DIFFERENT	2.495	6
20	2.953	DIFFERENT	DIFFERENT	2.495	9
21	6.913	DIFFERENT	DIFFERENT	2.495	8
22	14.041	SAME	DIFFERENT	2.495	8
23	18.744	SAME	DIFFERENT	2.495	9
24	6.977	NOT SURE	DIFFERENT	2.495	5
25	7.192	NOT SURE	SAME	2.538	5
Total Time = 490.79499999999996			Accuracy = 60%		

User ID: 382_598

Gender: Male

Age: 21 - 30

Education: Undergraduate

Eyesight: Normal (not using glasses)

Colour Blindness: No

Occupation: Doctor

Year registered as a Doctor: 2010

Experience with cellulitis/infections: 2 years

Table A.10: Participants master data- Dr Male, example3

Test No	Time Taken(secs)	User Result	System Result	Cogtool Time	Confidence
1	20.408	DIFFERENT	DIFFERENT	2.495	6
2	5.344	DIFFERENT	DIFFERENT	2.495	9
3	8.829	DIFFERENT	DIFFERENT	2.495	7
4	2.922	DIFFERENT	DIFFERENT	2.495	9
5	23.455	SAME	SAME	2.538	9
6	10.72	DIFFERENT	SAME	2.538	7
7	11.282	DIFFERENT	DIFFERENT	2.495	9
8	3.923	SAME	DIFFERENT	2.495	9
9	3.704	DIFFERENT	DIFFERENT	2.495	9
10	3.141	DIFFERENT	DIFFERENT	2.495	9
11	2.562	DIFFERENT	DIFFERENT	2.495	9
12	2.219	DIFFERENT	DIFFERENT	2.495	9
13	2.39	DIFFERENT	DIFFERENT	2.495	9
14	1.704	DIFFERENT	DIFFERENT	2.495	9
15	2.516	DIFFERENT	SAME	2.538	9
16	4.297	DIFFERENT	SAME	2.538	9
17	2.078	DIFFERENT	DIFFERENT	2.495	9
18	2	DIFFERENT	DIFFERENT	2.495	9
19	8.188	SAME	DIFFERENT	2.495	7
20	3.172	DIFFERENT	DIFFERENT	2.495	9
21	2.469	DIFFERENT	DIFFERENT	2.495	9
22	3.062	DIFFERENT	DIFFERENT	2.495	9
23	8.36	SAME	DIFFERENT	2.495	9
24	15.11	SAME	DIFFERENT	2.495	8
25	2.938	DIFFERENT	SAME	2.538	8
Total Time = 156.793			Accuracy = 68%		

User ID: 382_422

Gender: Male

Age: 31 - 40

Education: Postgraduate

Eyesight: Normal (not using glasses)

Colour Blindness: No

Occupation: Doctor

Year registered as a Doctor: 2000

Experience with cellulitis/infections: 10 years

Table A.11: Participants master data- Dr Male, example4

Test No	Time Taken(secs)	User Result	System Result	Cogtool Time	Confidence
1	22.5	DIFFERENT	DIFFERENT	2.495	7
2	5.656	DIFFERENT	DIFFERENT	2.495	9
3	5.329	SAME	DIFFERENT	2.495	8
4	3.703	DIFFERENT	DIFFERENT	2.495	9
5	3.016	DIFFERENT	SAME	2.538	9
6	5.407	DIFFERENT	SAME	2.538	9
7	2.954	DIFFERENT	DIFFERENT	2.495	9
8	20.642	DIFFERENT	DIFFERENT	2.495	6
9	2.594	DIFFERENT	DIFFERENT	2.495	9
10	4.61	DIFFERENT	DIFFERENT	2.495	8
11	3.047	DIFFERENT	DIFFERENT	2.495	9
12	2.61	DIFFERENT	DIFFERENT	2.495	9
13	2.609	DIFFERENT	DIFFERENT	2.495	9
14	3.063	DIFFERENT	DIFFERENT	2.495	9
15	2.859	DIFFERENT	SAME	2.538	9
16	2.328	DIFFERENT	SAME	2.538	9
17	2.578	DIFFERENT	DIFFERENT	2.495	9
18	3.438	DIFFERENT	DIFFERENT	2.495	9
19	5.172	DIFFERENT	DIFFERENT	2.495	7
20	2.657	DIFFERENT	DIFFERENT	2.495	9
21	3.187	DIFFERENT	DIFFERENT	2.495	9
22	5.063	DIFFERENT	DIFFERENT	2.495	9
23	6.86	DIFFERENT	DIFFERENT	2.495	8
24	3.578	DIFFERENT	DIFFERENT	2.495	8
25	6.328	SAME	SAME	2.538	8
Total Time = 131.78799999999998			Accuracy = 80%		

A.12 Appendix 11: Conjunctivitis digital images

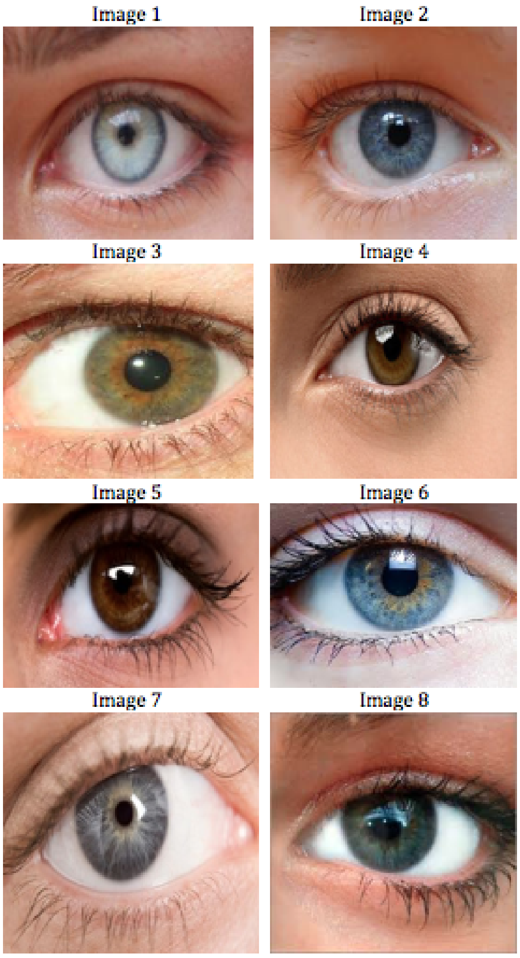


Figure A.16: Normal red (None)

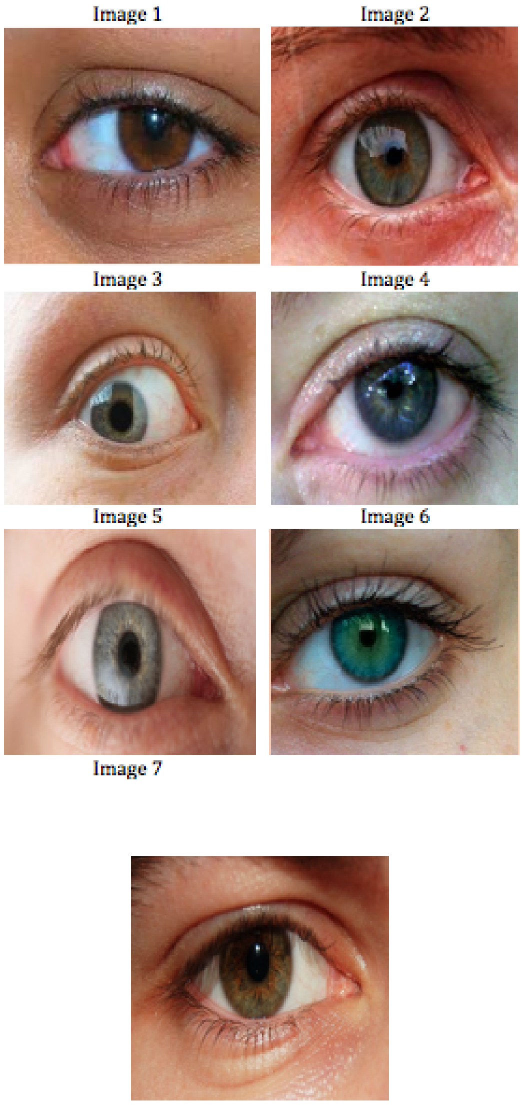


Figure A.17: Conjunctivitis digital images-Trace red

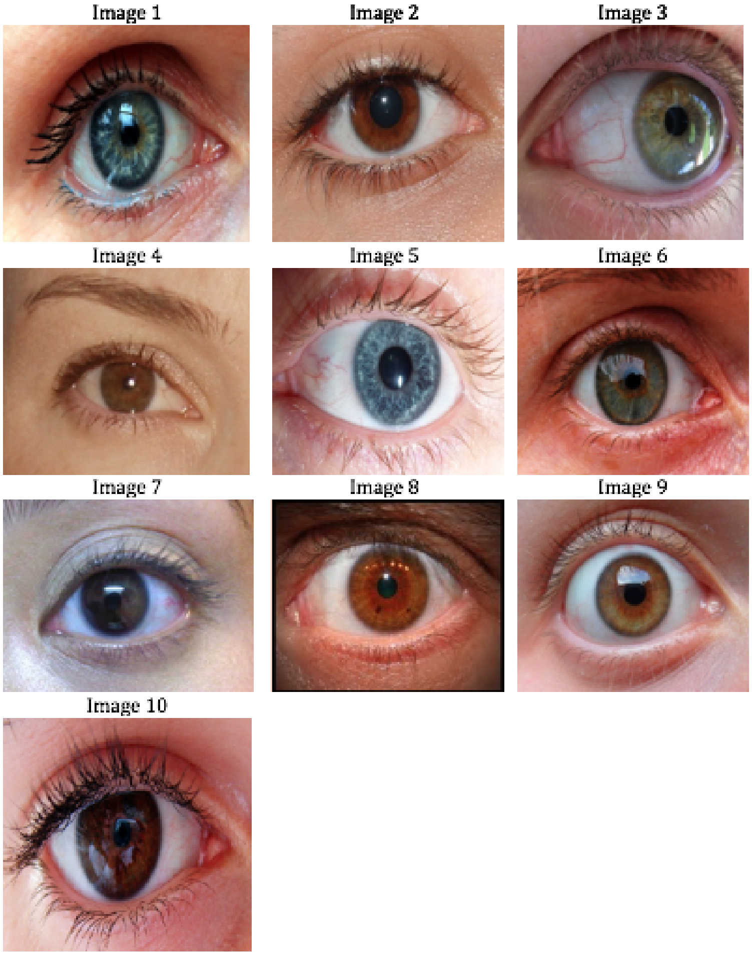


Figure A.18: Conjunctivitis digital images-Mild red

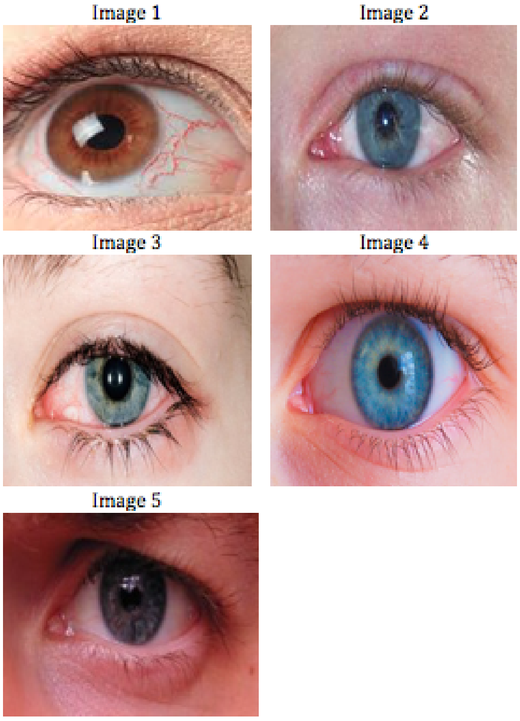


Figure A.19: Conjunctivitis digital images-Moderate red

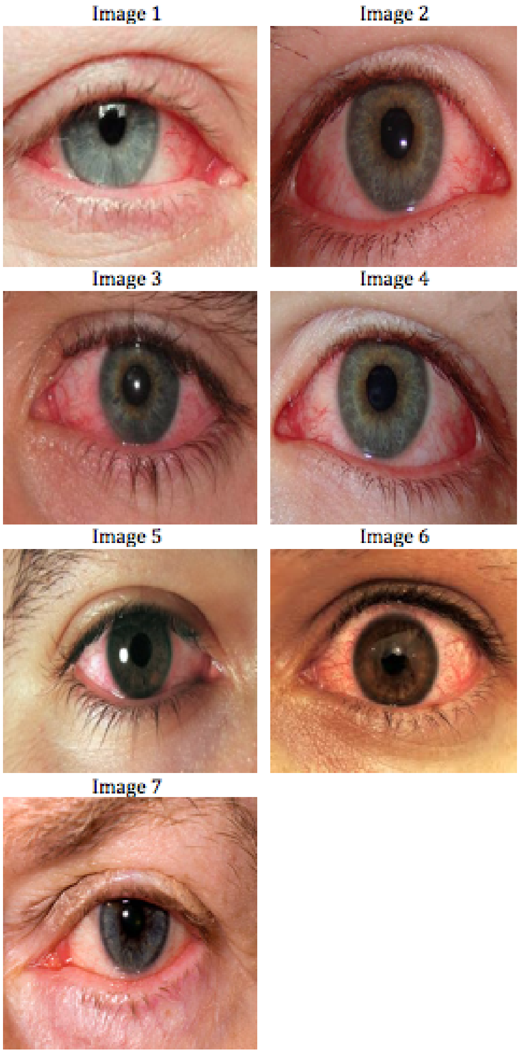


Figure A.20: Conjunctivitis digital images-Severe red

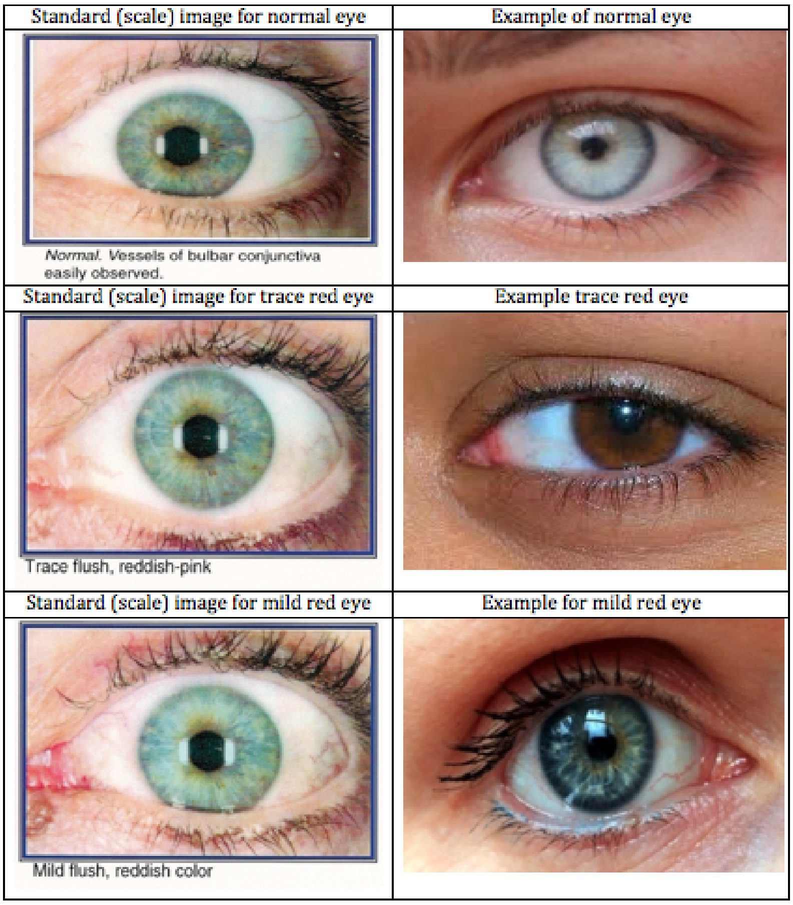

Standard (scale) image for normal eye
Example of normal eye
Normal. Vessels of bulbar conjunctiva easily observed.
Standard (scale) image for trace red eye
Example trace red eye
Trace flush, reddish-pink
Standard (scale) image for mild red eye
Example for mild red eye
Mild flush, reddish color

Figure A.21: Conjunctivitis digital images matching in the scale

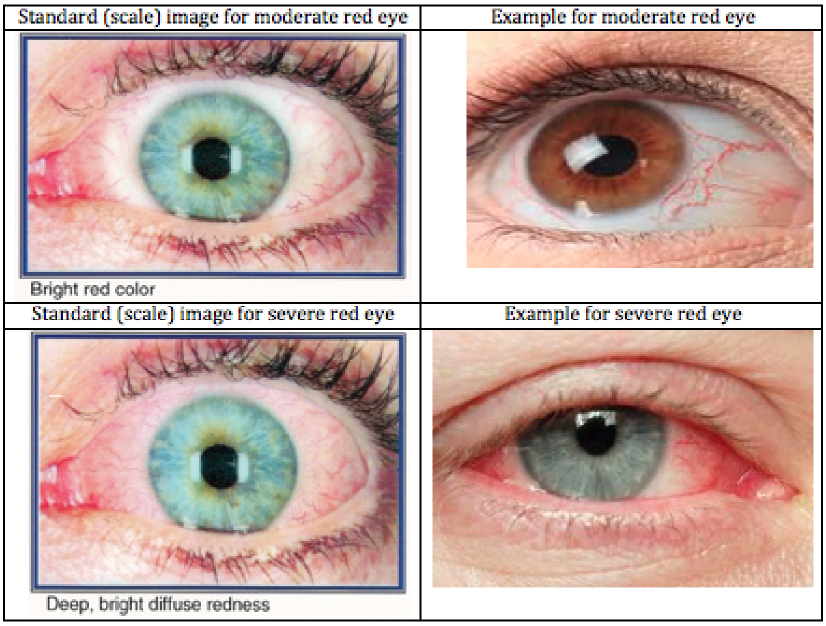

Standard (scale) image for moderate red eye
Example for moderate red eye
Bright red color
Standard (scale) image for severe red eye
Example for severe red eye
Deep, bright diffuse redness

Figure A.22: Conjunctivitis digital images matching in the scale

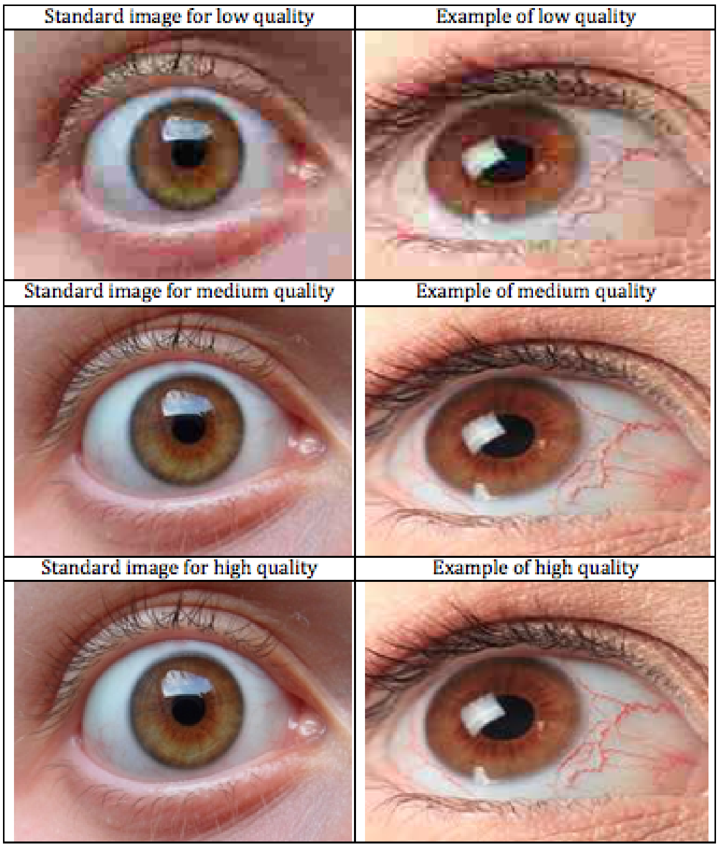

Standard image for low quality
Example of low quality
Standard image for medium quality
Example of medium quality
Standard image for high quality
Example of high quality

Figure A.23: Examples of image quality levels in the scale

A.13 Appendix 12: Red eye Online Instruction sheets-part1

General Background
This online experiment is being conducted by Ibrahim Alwawi. I am a PhD Research Student doing my project in Telehealth and Medical Imaging. This project is in partnership with NHS24 and the A&E Department in Aberdeen Royal Infirmary ARI and the Robert Gordon University RGU. The research focuses on human colour perception when diagnosing skin infection and cellulitis and red eye. This involves mainly doctors and any other related healthcare experts. Also it involves non-health workers for comparison and analysis purpose.

The project requires data from at least 30 participants. You have been invited to contribute based on your experience and knowledge in colour matching. Telehealth involves medical diagnosis and treatment through telecommunication systems and devices typically involving images, video and sound. This is a special case of complex safety-critical human computer interaction where cognitive overload can result in errors of judgment. This project involves analysis and comparing images to examine how these errors may occur.

The objective and tasks of red eye experimentset 1 (Image Matching)
To test the accuracy level of users in colour characteristics and image quality matching and make comparisons between data from healthcare professionals and data from non-health professionals.

This is a Telehealth experiment focusing on an Image Matching Test which consists of series of 5 tasks.

1. Task1:1 image showing a red eye. Please grade the redness of the eye using the provided Grading Guide.
2. Task2:2 images showing a red eye each for two patients. Please choose the correct statement that compares between their redness.
3. Task3:3 images showing a red eye. Please rank them based on the degree of their redness using the given grading guide.
4. Task4:1 image showing a red eye. Please grade the redness of the eye without using the any Grading Guide.
5. Task5:3 images showing a red eye. Please rank them based on the degree of their redness without using any grading guide.

The objective and tasks of red eye experimentset 2 (Image Matching)
To test the accuracy level of users in colour characteristics and image quality matching and make comparisons between data from healthcare professionals and data from non-health professionals.

This is a Telehealth experiment focusing on an Image Matching Test. You have only 6 screens in this short experiment. The experiment has 3 tasks on Image Matching.

1. Task6:5 images showing a red eye. Please rank them based on the degree of their redness using the given grading guide.
2. Task7:12 images showing a red eye. Please rank them based on the degree of their redness using the given grading guide.
3. Task8:5 images showing a red eye. Please rank them based on the degree of their redness without using any grading guide.
4. Task9:12 images showing a red eye. Please rank them based on the degree of their redness without using any grading guide.

The objective and tasks of red eye experimentset 3 (Image Matching)

This experiment focuses on the image quality and not the degree of redness. The study classifies the quality of an image to three levels (high, medium, and low).

High Quality image: In the high quality level, the image characteristics allow you to recognise the contents of the image clearly and easily and you will be able to have clear perception, and understanding of the image. Also you will be able to describe the image and make an accurate judgment about it and you confidence in your answer should be above 80

Medium Quality image: In the medium quality level, the image characteristics allow you to recognise the contents of the image but not as clear as the high quality image and you will be able to have perception and understanding of the image but not as clear as the high quality image. Also you will be able to describe the image and make judgment about it but this judgment is not as accurate as the high quality image and your confidence in your answer should normally be between (60-70%).

Low Quality image: In the low quality level, the image characterstics do not allow you to recognise the contents of the image clearly and you will not be able to have clear perception and understanding of the image. Also you will be struggling to describe the image and make judgment about it. Your confidence in your answer should normally be less than 50

The experiment has the following tasks on Image Matching.

1. Task10:1 image showing a red eye. Please rank it based on the image quality (Low-Medium-High).
2. Task11:2 images showing red eyes. Please state which one has the best quality. Please note that the degree of redness has no relevance.
3. Task12:3 images showing red eyes. Please state which one has the best quality. Please note that the degree of redness has no relevance.
4. Task13:5 images showing red eyes. Please state which one has the best quality. Please note that the degree of redness has no relevance.
5. Task14:12 images showing red eyes. Please state which one has the best quality. Please note that the degree of redness has no relevance.
6. Task15:12 images showing red eyes. Please state which of the images have acceptable quality or not. Please note that the degree of redness has no relevance.

The objective and tasks of red eye experimentset 4 (Image Matching)

To test the accuracy level of users in colour characteristics and image quality matching and make comparisons between data from healthcare professionals and data from non-health professionals.

This is a Telehealth experiment focusing on an Image Matching Test which consists of series of 5 tasks.

1. Task16:1 image showing a red eye. Please grade the redness of the eye using the provided Grading Guide.
2. Task17:2 images showing a red eye each for two patients. Please choose the correct statement that compares between their redness.
3. Task18:3 images showing a red eye. Please rank them based on the degree of their redness using the given grading guide.
4. Task19:1 image showing a red eye. Please grade the redness of the eye without using the any Grading Guide.
5. Task20:3 images showing a red eye. Please rank them based on the degree of their redness without using any grading guide.

The objective and tasks of red eye experimentset 5 (Image Matching)

To test the accuracy level of users in colour characteristics and image quality matching and make comparisons between data from healthcare professionals and data from non-health professionals.

This is a Telehealth experiment focusing on an Image Matching Test. You have only 6 screens in this short experiment. The experiment has 3 tasks on Image Matching.

1. Task21:5 images showing a red eye. Please rank them based on the degree of their redness using the given grading guide.
2. Task22:12 images showing a red eye. Please rank them based on the degree of their redness using the given grading guide.
3. Task23:5 images showing a red eye. Please rank them based on the degree of their redness without using any grading guide.
4. Task24:12 images showing a red eye. Please rank them based on the degree of their redness without using any grading guide.

The Confidence Level

You also get to indicate how confident you are in your decisions by using from the provided scale 0-9. If your confidence is low then it will be between 0-3 and if it is medium then it will be between 4-6 and if it is high it will be between 7-9.

Break Pages

In this experiment there is an interval page serving as a break in between the task pages.

Remember

Your data is an important contribution to the academic research in the areas of Telehealth and Cognitive Engineering for improved healthcare services.

A.14 Appendix 12.1: Red eye Online Instruction sheets and Consent-part2

Instructions:

- Fill in the required details in the form below.
- Click on the START button to begin the experiment.

- Answer the question by choosing the correct answer.
- You may not like some of the images but they are not offensive or scary.
- You may get more than one question which requires more than one answer.
- Then make a Confidence Level selection from 1 to 9 based on how sure you are of your answer above.
- Click on the NEXT button to move to the next page.
- You can take breaks only between the pages as the timing stops on the interval pages between images.
- Please click on the FINISH button at the end of the experiment in order to save your results.

Important Technical Note

- Please DO NOT click on the BACK or REFRESH buttons on your browser during the experiment as this will return you to the Start page.
- JavaScript should not be turned off in your browser settings for this experiment.

Data Protection and Confidentially

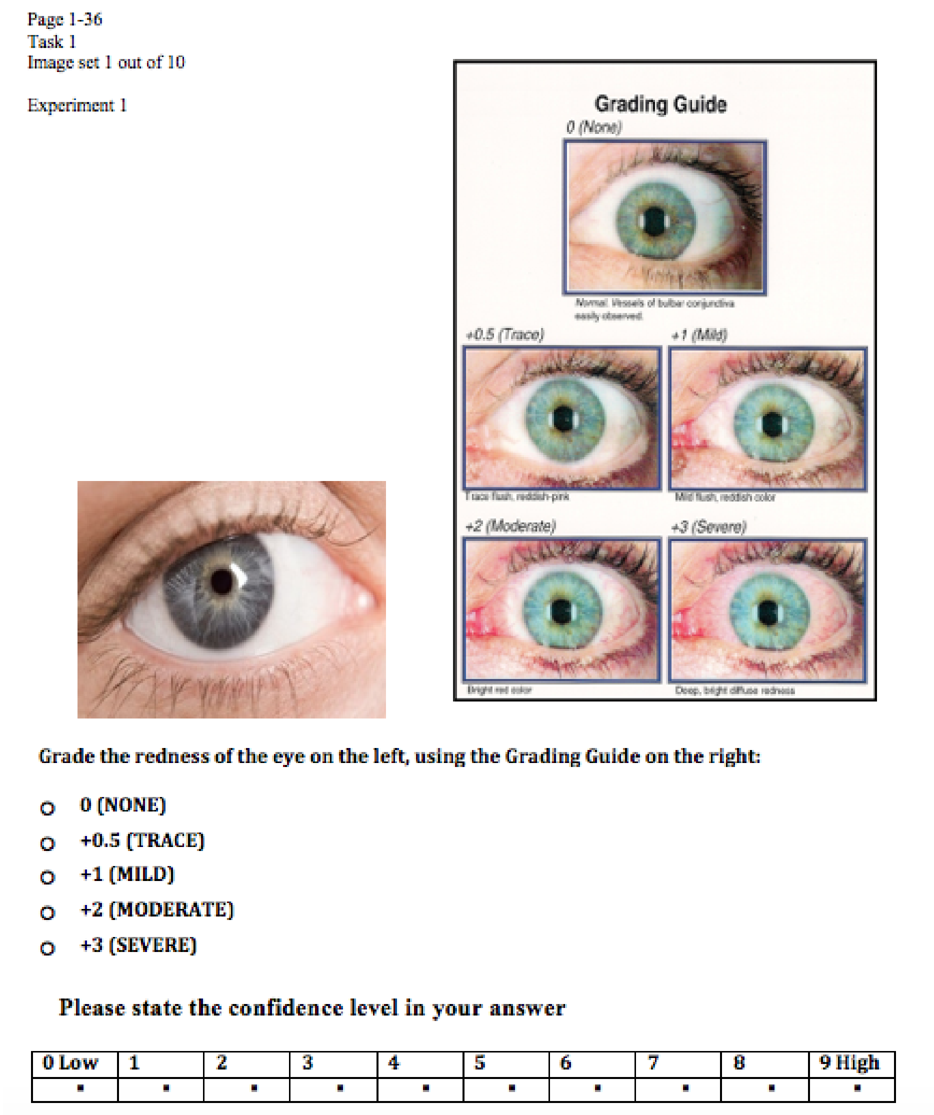
The user data is collected anonymously and will be treated confidentially and used only for research purposes.

Consent for Participation

Your participation in this study is voluntary and you have the right to withdraw from the experiment at any time, without giving any reason. After reading and understanding the above information about the study, if you are happy to participate in the experiment, then please log in via the form below using the password (ID) that was sent to you as part of the invitation email.

Thank you

A.15 Appendix 13: Examples of red eye online experiment



Page 1-36
Task 1
Image set 1 out of 10

Experiment 1

Grading Guide
0 (None)
Normal. Vessels of bulbar conjunctiva easily observed.
+0.5 (Trace)
+1 (Mild)
Trace flush, reddish-pink
Mild flush, reddish color
+2 (Moderate)
+3 (Severe)
Bright red color
Deep, bright diffuse redness

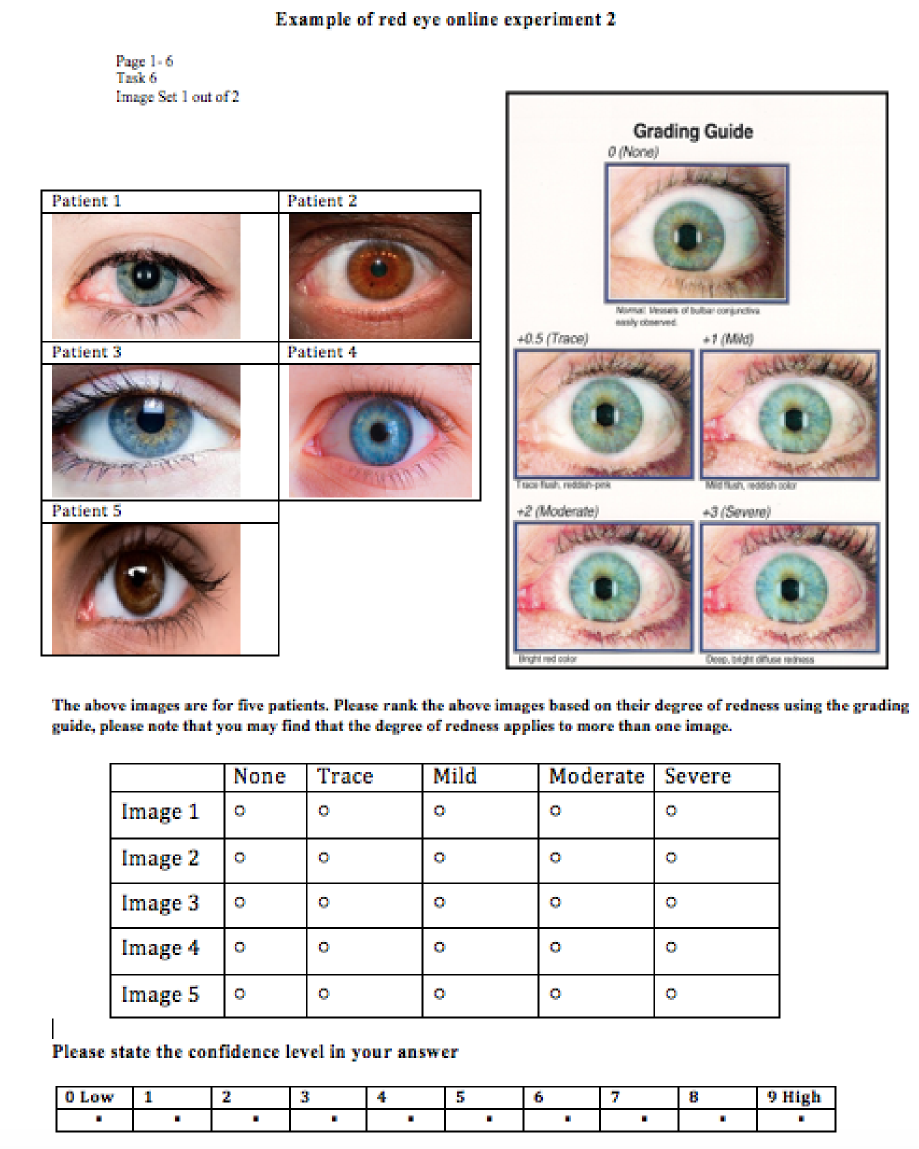
Grade the redness of the eye on the left, using the Grading Guide on the right:

- ○ **0 (NONE)**
- ○ **+0.5 (TRACE)**
- ○ **+1 (MILD)**
- ○ **+2 (MODERATE)**
- ○ **+3 (SEVERE)**

Please state the confidence level in your answer

0 Low	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9 High**
▪	▪	▪	▪	▪	▪	▪	▪	▪	▪

Figure A.24: Example of red eye online experiment 1



Example of red eye online experiment 2

Page 1- 6
Task 6
Image Set 1 out of 2

Patient 1

Patient 2

Patient 3

Patient 4

Patient 5

Grading Guide
0 (None)
Normal: Vessels of bulbar conjunctiva easily observed
+0.5 (Trace)
Trace flush, reddish-pink
+1 (Mild)
Mild flush, reddish color
+2 (Moderate)
Bright red color
+3 (Severe)
Deep, bright diffuse redness

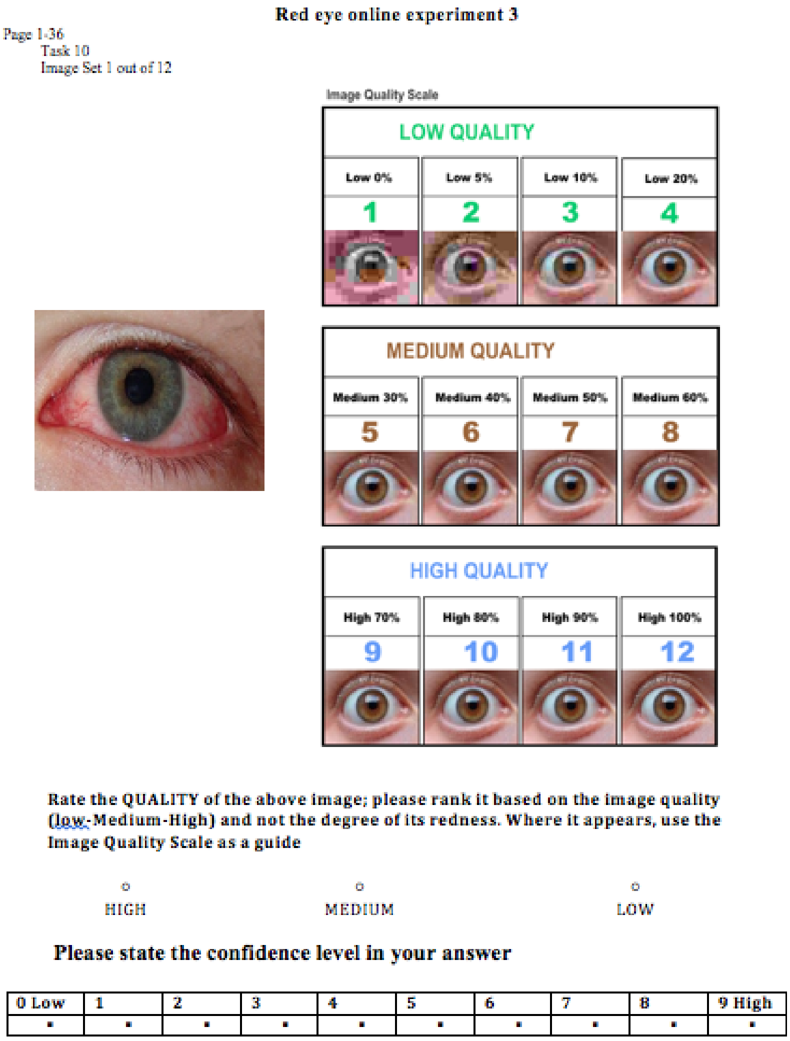
The above images are for five patients. Please rank the above images based on their degree of redness using the grading guide, please note that you may find that the degree of redness applies to more than one image.

	None	Trace	Mild	Moderate	Severe
Image 1	○	○	○	○	○
Image 2	○	○	○	○	○
Image 3	○	○	○	○	○
Image 4	○	○	○	○	○
Image 5	○	○	○	○	○

Please state the confidence level in your answer

0 Low	1	2	3	4	5	6	7	8	9 High
▪	▪	▪	▪	▪	▪	▪	▪	▪	▪

Figure A.25: Example of red eye online experiment 2



Red eye online experiment 3

Page 1-36
Task 10
Image Set 1 out of 12

Image Quality Scale

LOW QUALITY			
Low 0%	Low 5%	Low 10%	Low 20%
1	2	3	4

MEDIUM QUALITY			
Medium 30%	Medium 40%	Medium 50%	Medium 60%
5	6	7	8

HIGH QUALITY			
High 70%	High 80%	High 90%	High 100%
9	10	11	12

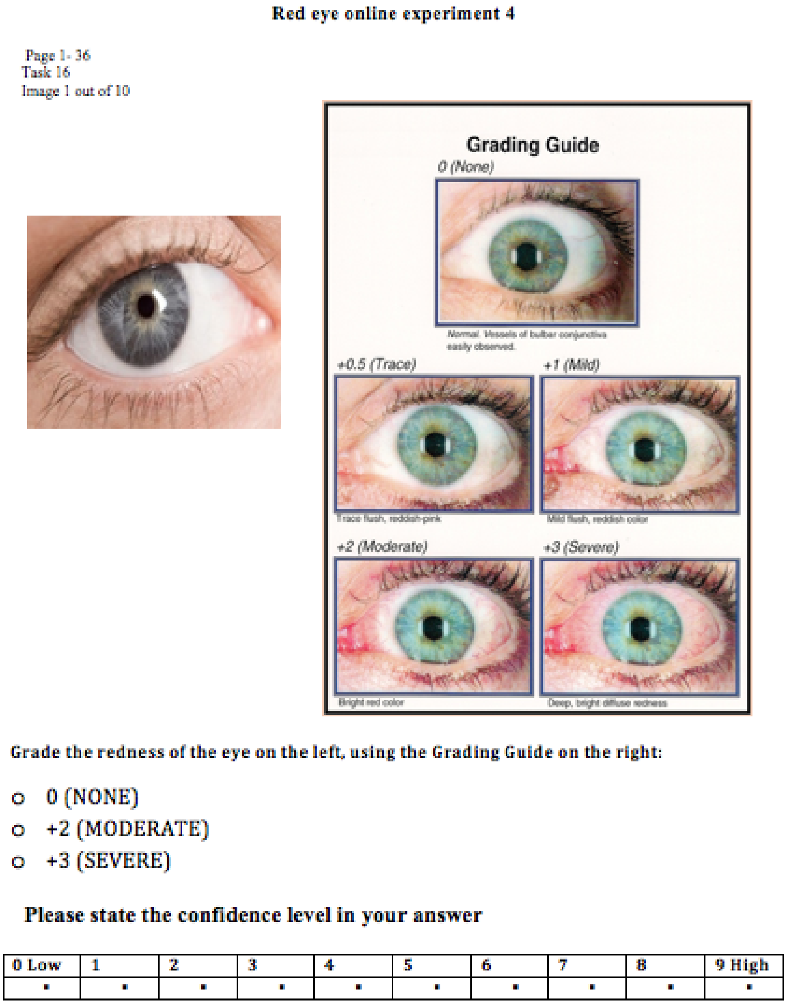
Rate the QUALITY of the above image; please rank it based on the image quality (low-Medium-High) and not the degree of its redness. Where it appears, use the Image Quality Scale as a guide

○ HIGH ○ MEDIUM ○ LOW

Please state the confidence level in your answer

0 Low	1	2	3	4	5	6	7	8	9 High
▪	▪	▪	▪	▪	▪	▪	▪	▪	▪

Figure A.26: Example of red eye online experiment 3



Red eye online experiment 4

Page 1- 36
Task 16
Image 1 out of 10

Grading Guide
0 (None)
Normal. Vessels of bulbar conjunctiva easily observed.
+0.5 (Trace)
Trace flush, reddish-pink
+1 (Mild)
Mild flush, reddish color
+2 (Moderate)
Bright red color
+3 (Severe)
Deep, bright diffuse redness

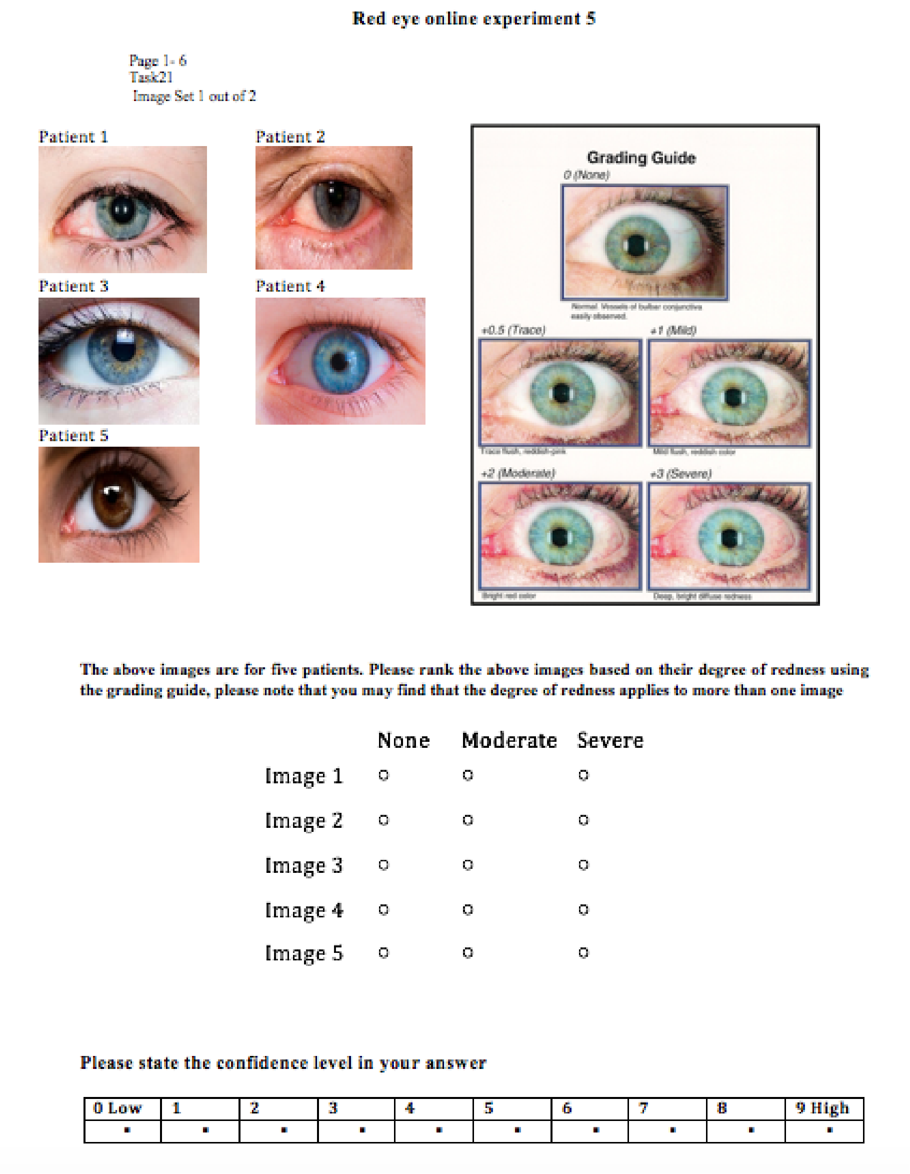
Grade the redness of the eye on the left, using the Grading Guide on the right:

- ○ 0 (NONE)
- ○ +2 (MODERATE)
- ○ +3 (SEVERE)

Please state the confidence level in your answer

0 Low	1	2	3	4	5	6	7	8	9 High
▪	▪	▪	▪	▪	▪	▪	▪	▪	▪

Figure A.27: Example of red eye online experiment 4



Patient 1
Patient 2
Patient 3
Patient 4
Patient 5
Grading Guide
0 (None)
Normal. Vessels of bulbar conjunctiva easily observed.
+0.5 (Trace)
Trace flush, reddish-pink
+1 (Mild)
Mild flush, reddish color
+2 (Moderate)
Bright red color
+3 (Severe)
Deep, bright diffuse redness

Red eye online experiment 5

Page 1- 6
Task21
Image Set 1 out of 2

The above images are for five patients. Please rank the above images based on their degree of redness using the grading guide, please note that you may find that the degree of redness applies to more than one image

	None	Moderate	Severe
Image 1	○	○	○
Image 2	○	○	○
Image 3	○	○	○
Image 4	○	○	○
Image 5	○	○	○

Please state the confidence level in your answer

0 Low	1	2	3	4	5	6	7	8	9 High
▪	▪	▪	▪	▪	▪	▪	▪	▪	▪

Figure A.28: Example of red eye online experiment 5

A.16 Appendix 14: Mann-Whitney test results for joint experiments 5a and 5b

Table A.12: Descriptive Statistics for experiments 5a and 5b

	N	Mean	Std.Deviation	Minimum	Maximum	Percentiles		
						25th	50th (Median)	75th
Total accuracy for Exp. 5a and 5b	172	6.8547	2.3086	1	12	6	7	9
Doctors and non-Doctors	172	0.4012	0.49156	0	1	0	0	1

Table A.13: Ranks

	Doctors and non-Doctors	N	Mean Rank	Sum of Ranks
Total accuracy for Exp. 5a and 5b	0	103	84.16	8668.5
	1	69	89.99	6209.5
	Total	172		

Table A.14: Test Statistics

	Total accuracy for Exp. 5a and 5b
Mann-Whitney U	3312.5
Wilcoxon W	8668.5
Z	-0.761
Asymp. Sig. (2-tailed)	0.447
a. Grouping Variable: Dr_or_Non	

The severity levels were denoted as N-Normal, TR-Trace, I-Mild, MD-Moderate and S-Severe. 0 denotes a test image while an X denotes a closest matched training image.

Table A.15: Classification System 1

	Normal	Trace	Mild	Moderate	Severe	Error Level
	Normal	RED				
Patient No.						
1					ox	NONE
2	ox					NONE
3					ox	NONE
4					ox	NONE
5				ox		NONE
6					ox	NONE
7	x		o			Level 2
8					ox	NONE
9	ox					NONE
10	x	o				Level 1
11		o	x			NONE
12					ox	NONE
13	ox					NONE
14	x		o			Level 2
15	ox					NONE
16				ox		NONE
17					ox	NONE
18					ox	NONE
19		o	x			NONE
20	o				x	Level 4
21	ox					NONE
22				ox		NONE
23	x	o				Level 1
24	x	o				Level 1
25			x		o	NONE

Table A.16: Classification System 2

	Normal	Trace	Mild	Moderate	Severe	Error Level
	Normal	Moderate			Severe	
Patient No						
1					ox	NONE
2	ox					NONE
3					ox	NONE
4					ox	NONE
5				ox		NONE
6					ox	NONE
7	x		o			Level 2
8					ox	NONE
9	ox					NONE
10	x	o				Level 1
11		o	x			NONE
12					ox	NONE
13	ox					NONE
14	x		o			Level 2
15	ox					NONE
16				ox		NONE
17					ox	NONE
18					ox	NONE
19		o	x			NONE
20	o				x	Level 4
21	ox					NONE
22				ox		NONE
23	x	o				Level 1
24	x	o				Level 1
25			x		o	Level 2

Table A.17: Classification System 3

	Normal	Trace	Mild	Moderate	Severe	Error Level
Patient No.						
1					ox	NONE
2	ox					NONE
3					ox	NONE
4					ox	NONE
5				Ox		NONE
6					ox	NONE
7	x		o			Level 2
8					ox	NONE
9	ox					NONE
10	x	o				Level 1
11		o	x			Level 1
12					ox	NONE
13	ox					NONE
14	x		o			Level 2
15	ox					NONE
16				Ox		NONE
17					ox	NONE
18					ox	NONE
19		o	x			Level 1
20	o				x	Level 4
21	ox					NONE
22				Ox		NONE
23	x	o				Level 1
24	x	o				Level 1
25			x		o	Level 2

A.17 Appendix B-Cellulitis Results

Table A.18: Classification System 4

	Normal	Trace	Mild	Moderate	Severe	Error Level
	Normal				Severe	
Patient No						
1	ox					NONE
2				Ox		NONE
3	ox					NONE
4				Ox		NONE
5					ox	NONE
6					ox	NONE
7					ox	NONE
8		o		X		Level 2
9				Ox		NONE
10				Ox		NONE
11	ox					NONE
12					ox	NONE
13					ox	NONE
14				Ox		NONE
15					ox	NONE
16				Ox		NONE
17		ox				NONE
18	ox					NONE
19				Ox		NONE
20	ox					NONE
21					ox	NONE
22				Ox		NONE
23		ox				NONE
24		ox				NONE
25				Ox		NONE
26					ox	NONE
27				Ox		NONE
28					ox	NONE
29				Ox		NONE
30	ox					NONE
31					ox	NONE
32				Ox		NONE
33				Ox		NONE

Table A.19: Classification System 5

	Normal	Trace	Mild	Moderate	Severe	Error Level
	Normal				Severe	
Patient No						
34				O	x	Level 1
35					ox	NONE
36					ox	NONE
37					ox	NONE
38		ox				NONE
39	ox					NONE
40	ox					NONE
41					ox	NONE
42				Ox		NONE
43				Ox		NONE
44					ox	NONE
45					ox	NONE
46		ox				NONE
47				Ox		NONE
48	ox					NONE
49		ox				NONE
50					ox	NONE
51					ox	NONE
52				Ox		NONE
53	ox					NONE
54					ox	NONE
55					ox	NONE
56					ox	NONE
57					ox	NONE
58					ox	NONE
59		ox				NONE
60					ox	NONE
61					ox	NONE
62				Ox		NONE
63					ox	NONE
64				Ox		NONE
65					ox	NONE
66		ox				NONE
67					ox	NONE
68					ox	NONE
69					ox	NONE

Table A.20: Classification System 6

	Normal	Trace	Mild	Moderate	Severe	Error Level
	Normal				Severe	
Patient No						
70		ox				NONE
71		ox				NONE
72					ox	NONE
73		ox				NONE
74		ox				NONE
75					ox	NONE
76		ox				NONE

Table A.21: Classification System 7

	Normal	Trace	Mild	Moderate	Severe	Error Level
	Normal			Moderate	Severe	
Patient No						
1	ox					NONE
2				ox		NONE
3	ox					NONE
4				ox		NONE
5					Ox	NONE
6					Ox	NONE
7					Ox	NONE
8		o		x		Level 2
9				ox		NONE
10				ox		NONE
11	ox					NONE
12					Ox	NONE
13					Ox	NONE
14				ox		NONE
15					Ox	NONE
16				ox		NONE
17		ox				NONE
18	ox					NONE
19				ox		NONE
20	ox					NONE
21					Ox	NONE
22				ox		NONE
23		ox				NONE
24		ox				NONE
25				ox		NONE
26					Ox	NONE
27				ox		NONE
28					Ox	NONE
29				ox		NONE
30	ox					NONE

Table A.22: Classification System 8

	Normal	Trace	Mild	Moderate	Severe	Error Level
	Normal			Moderate	Severe	
Patient No						
31					Ox	NONE
32				ox		NONE
33				ox		NONE
34				o	X	Level 1
35					Ox	NONE
36					Ox	NONE
37					Ox	NONE
38		ox				NONE
39	ox					NONE
40	ox					NONE
41					Ox	NONE
42				ox		NONE
43				ox		NONE
44					Ox	NONE
45					Ox	NONE
46		ox				NONE
47				ox		NONE
48	ox					NONE
49		ox				NONE
50					Ox	NONE
51					Ox	NONE
52				ox		NONE
53	ox					NONE
54					Ox	NONE
55					Ox	NONE
56					Ox	NONE
57					Ox	NONE
58					Ox	NONE
59		ox				NONE
60					Ox	NONE
61					Ox	NONE
62				ox		NONE
63					Ox	NONE
64				ox		NONE
65					Ox	NONE
66		ox				NONE
67					Ox	NONE
68					Ox	NONE
69					Ox	NONE
70		ox				NONE
71		ox				NONE
72					Ox	NONE
73		ox				NONE
74		ox				NONE
75					Ox	NONE
76		ox				NONE

Table A.23: Classification System 9

	Normal	Trace	Mild	Moderate	Severe	Error Level
Patient No						
1	ox					NONE
2				Ox		NONE
3	ox					NONE
4				Ox		NONE
5					ox	NONE
6					ox	NONE
7					ox	NONE
8		o		X		Level 2
9				Ox		NONE
10				Ox		NONE
11	ox					NONE
12					ox	NONE
13					ox	NONE
14				Ox		NONE
15					ox	NONE
16				Ox		NONE
17		ox				NONE
18	ox					NONE
19				Ox		NONE
20	ox					NONE
21					ox	NONE
22				Ox		NONE
23		ox				NONE
24		ox				NONE
25				Ox		NONE
26					ox	NONE

Table A.24: Classification System 10

	Normal	Trace	Mild	Moderate	Severe	Error Level
Patient No						
27				Ox		NONE
28					ox	NONE
29				Ox		NONE
30	ox					NONE
31					ox	NONE
32				Ox		NONE
33				Ox		NONE
34				O	x	Level 1
35					ox	NONE
36					ox	NONE
37					ox	NONE
38		ox				NONE
39	ox					NONE
40	ox					NONE
41					ox	NONE
42				Ox		NONE
43				Ox		NONE
44					ox	NONE
45					ox	NONE
46		ox				NONE
47				Ox		NONE
48	ox					NONE
49		ox				NONE
50					ox	NONE
51					ox	NONE
52				Ox		NONE
53	ox					NONE
54					ox	NONE
55					ox	NONE
56					ox	NONE
57					ox	NONE
58					ox	NONE
59		ox				NONE
60					ox	NONE
61					ox	NONE
62				Ox		NONE
63					ox	NONE
64				Ox		NONE
65					ox	NONE

Table A.25: Classification System 11

	Normal	Trace	Mild	Moderate	Severe	Error Level
Patient No						
66		ox				NONE
67					ox	NONE
68					ox	NONE
69					ox	NONE
70		ox				NONE
71		ox				NONE
72					ox	NONE
73		ox				NONE
74		ox				NONE
75					ox	NONE
76		ox				NONE

A.18 Appendix D-Cellulitis Classification Result

Table A.26: Classification Error - Conjunctivitis two classes

Patient No.	Error Level 1	Error Level 2	Error Level 3	Error Level 4	Misdiagnosed
Conjunctivitis-2 Classes					
7		x			Mild classified as normal
10	x				Trace classified as normal
14		x			Mild classified as normal
20				x	Normal classified as severe
23	x				Trace classified as normal
24	x				Trace classified as normal

Table A.27: Classification Error - Conjunctivitis three classes

Patient No.	Error Level 1	Error Level 2	Error Level 3	Error Level 4	Misdiagnosed
Conjunctivitis-3 Classes					
7		x			Mild classified as normal
10	x				Trace classified as normal
14		x			Mild classified as normal
20				x	Normal classified as severe
23	x				Trace classified as normal
24	x				Trace classified as normal
25	x				Severe classified as mild

Table A.28: Classification Error - Conjunctivitis five classes

Patient No.	Error Level 1	Error Level 2	Error Level 3	Error Level 4	Misdiagnosed
Conjunctivitis-5 Classes					
7		x			Mild classified as normal
11	x				Trace classified as mild
10	x				Trace classified as normal
14		x			Mild classified as normal
19	x				Trace classified as mild
20				x	Normal classified as severe
23	x				Trace classified as normal
24	x				Trace classified as normal
25	x				Severe classified as mild

Table A.29: Classification Error - Cellulitis two, three and five classes

Patient No.	Error Level 1	Error Level 2	Error Level 3	Error Level 4	Misdiagnosed
Cellulitis					
8		x			Trace classified as moderate
34	x				Moderate classified as severe