CT Colonography:
Defining Performance and Interventions to Improve Interpretation

Dr David Burling,
MBChB MRCP FRCR

Intestinal Imaging Centre
St. Mark’s Hospital
London

Submitted for the degree of
Doctor of Medicine
Leicester University

2006
Dedicated to my wife and children;

Fiona, Katie, Rosie and Jemima
CT Colonography:
Defining Performance and Interventions to Improve Interpretation

David Burling, MBChB MRCP FRCR

ABSTRACT

CT colonography is now widely regarded as the optimal radiological technique for colonic examination. However, performance characteristics, particularly relating to interpretation accuracy, interpretation times and polyp measurement have been derived primarily from academic centres in the USA. For successful implementation of CT colonography, such performance characteristics must be generalisable to non-academic centres, different patient populations in different geographical locations, and different clinical environments. This thesis aims to investigate current UK implementation and to determine the interpretative performance of observers across the UK and Europe, focussing on those interventions influencing reader accuracy and polyp measurement.

The first chapter reviews the technique, diagnostic performance and clinical role of CT colonography. It is followed by two surveys revealing CT colonography is widely available across the UK NHS, and that hitherto unsuspected complications do occur, although CT colonography appears relatively safe in routine clinical practice when compared to alternatives.
The effect of directed training on reader performance is investigated in a multi-centre European study and shows that experienced radiologists are significantly more accurate and time-efficient when reporting CT colonography compared to specifically trained but less experienced radiologists and radiographers. A subsequent study also shows they are more accurate than radiologists offering CT colonography in UK clinical practice routinely. Trained radiographers can perform as well as their radiologist counterparts.

The accuracy of polyp measurement is investigated and the effect of different visualisation displays determined. Results suggest that an optimised 2D display utilising a 'colon CT window' should be recommended. Finally, we show that an automatic measurement software tool improves inter and intra-observer agreement for polyp measurement 'in vitro' although the benefit for in-vivo measurement is less clear.
# LIST OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ABSTRACT</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Preface</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>Acknowledgements</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>List of tables</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>List of figures</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>Ethical approval statement</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td>Glossary of abbreviations</td>
<td>22</td>
</tr>
<tr>
<td>1.</td>
<td>INTRODUCTION</td>
<td></td>
</tr>
<tr>
<td>1.1</td>
<td>Background</td>
<td>23</td>
</tr>
<tr>
<td>1.2</td>
<td>Hypothesis and Aims</td>
<td>26</td>
</tr>
<tr>
<td>1.3</td>
<td>Thesis strategy</td>
<td>29</td>
</tr>
<tr>
<td>2.</td>
<td>CT COLONOGRAPHY: A review of technique, diagnostic</td>
<td></td>
</tr>
<tr>
<td></td>
<td>performance and clinical role</td>
<td></td>
</tr>
<tr>
<td>2.1</td>
<td>Introduction</td>
<td>31</td>
</tr>
<tr>
<td>2.2</td>
<td>Technique</td>
<td>37</td>
</tr>
<tr>
<td>2.3</td>
<td>Current clinical role</td>
<td>48</td>
</tr>
</tbody>
</table>
3. **CT COLONOGRAPHY PRACTICE IN THE UNITED KINGDOM**

3.1 CT Colonography Practice in the UK: A National Survey

3.2 Potentially serious adverse events associated with CT colonography performed in symptomatic patients: A national survey of the UK

4. **READER PERFORMANCE**

4.1 Effect of directed training on reader performance at CT colonography.

4.2 CT colonography: Interpretative performance in a non-academic environment.

5. **INTERPRETATION TIMES**

5.1 CT colonography interpretation times: Effect of reader experience, fatigue, and scan findings in a multi-centre setting
6. POLYP MEASUREMENT

6.1 Polyp measurement by CT colonography: Agreement with colonoscopy, and effect of viewing conditions on inter- and intra-observer agreement.

6.2 Polyp measurement and size categorisation by CT colonography: Effect of observer experience in a multi-centre setting.

6.3 CT colonography: Automated diameter and volume measurement of colonic polyps compared with a manual technique – in vitro study

6.4 CT colonography: Automatic measurement of polyp diameter compared with manual assessment – an in-vivo study

7. CONCLUSIONS

7.1 Introduction

7.2 Discussion of results

7.3 Conclusion
APPENDIX

1 Survey to determine CT Colonography practice in the UK 256

2 Telephone questionnaire survey of complications associated with CT colonography. 258

3 LIST OF PUBLICATIONS FROM THESIS 259

BIBLIOGRAPHY 268
This thesis represents original work by the author and has not been submitted in any form to any other University. Where use has been made of the work of others it has been duly acknowledged in the text.

The research described in this thesis was carried out in the Intestinal Imaging centre of St. Mark's Hospital with additional data collection from multiple European centres: University College London, London UK; Hope Hospital, Salford, UK; Bradford Royal Infirmary, UK; Academic Medical Center, Amsterdam, the Netherlands; Stedelijk Ziekenhuis, Roeselare, Belgium; Mater Misericordiae University Hospital, Dublin, Ireland; University of Pisa, Pisa, Italy; La Sapienza, Rome, Italy; Candiolo Oncologic Hospital, Turin, Italy; Queen Alexandra Hospital, Portsmouth, UK; Royal Cornwall Hospital, Truro, UK; Nottingham City Hospital, UK; Royal Lancaster Hospital, UK; Royal united Hospital, Bath, UK; Royal Oldham Hospital, UK; Queen Elizabeth hospital, Birmingham, UK.

The research described in this thesis was carried out under the supervision of Professor Steve Halligan, now of University College London.
ACKNOWLEDGMENTS

The author would like to express his sincere thanks to the many individuals, without whom this thesis would not have been possible.

Firstly to Professor Steve Halligan, my supervisor, who has provided unfaltering support throughout the planning and implementation of the studies included in this thesis. His passion for research is inspirational and without his involvement, little of this work would have been possible.

I would also like to thank my predecessor at St. Mark's, Dr Stuart Taylor. His considerable intellectual support and experience were greatly appreciated.

St. Mark's Hospital is a unique institution, providing a rich academic environment for research, which has benefitted this thesis. My particular thanks go to Professor Wendy Atkin of Cancer Research UK, for her insight into colorectal cancer epidemiology; Dr Brian Saunders for his colonoscopy expertise and advice; Professor Clive Bartram, Dr Michele Marshall, Dr Arun Gupta and Dr Andrew Slater, fellow radiologists, for their advice and support. I would also like to thank the receptionist, Josie, secretaries Bharti and Fowzan, and radiographers at St. Mark's Hospital for their friendship and support throughout my research fellowship.
I am indebted to Douglas Altman and his colleagues at the Cancer Research UK/NHS Centre for Statistics in Medicine in Oxford for their statistical advice and contributions. In addition, I would like to acknowledge the support and contribution of Paul Bassett, a statistical consultant who made valuable contributions to several of the studies in this thesis.

I would like to thank the many colleagues both in the UK and Europe, who in many instances have become firm friends, for their contributions to the studies in this thesis namely; Roger Frost (Salisbury NHS Trust, Salisbury, UK), Lesley Honeyfield, Melinda De Villiers (Medicsight plc); David Nicholson, Velauthan Rudralingham, Lisa Renaut (Hope Hospital, Salford, UK); Clive Kay, Andy Lowe, Jane Williams-Butt (Bradford Royal Infirmary, UK); Jaap Stoker, Jasper Florie, Martin Poulus, (Academic Medical Center, Amsterdam, the Netherlands); Victor Van der Hulst (Onze Lieve Vrouwe Gasthuis, Amsterdam, The Netherlands); Philippe Lefere, Jesse Marrannes, Guido Dessey (Stedelijk Ziekenhuis, Roeselare, Belgium); Helen Fenlon, Alan O'Hare, Shane Foley (Mater Misericordiae University Hospital, Dublin, Ireland); Emmanuele Neri, Paola Vagli, Benedetta Politi (University of Pisa, Pisa, Italy); Andrea Iagli, Riccardo Iannaccone, Filipo Mangiapane, Sante Ori (La Sapienza, Rome, Italy); Teresa Gallo, Giulia Nieddu, Saverio Signoretta, Daniele Regge (Candiolo Oncologic Hospital, Turin, Italy); Julian Atchley, Rajpal Dhingsar, Peter Guest, Stephen Hayward, Anthony Higginson, Craig Jobling, Clive Kay, Richard Lilford, Giles Maskell, Ian McCafferty, John McGregor, Dion Morton, Mahesh Kumar Neelala, Michael Noakes, Andrea Philips, Peter Riley, Alasdair Taylor.
Finally, I am very grateful to the following commercial supporters of this research for their help with the workstations and interpretation software used for these studies, both at participating sites and at the trial office: Medcisght plc (London, UK); Barco (Edinburgh, UK); E-Z-EM, inc (Westbury, NY, USA); Vital Images, Inc. (Minnetonka, MN, USA) and General Electric (Milwaukee, Wisconsin, USA).

Funding:
This thesis was funded in part by grants from the European Association of Radiology (EAR), administered by the European Society of Gastrointestinal and Abdominal Radiology (ESGAR); a Kodak Scholarship administered via the Royal College of Radiologists; the NHS Health Technology Assessment programme; and Medicsight plc.
LIST OF TABLES

Table 1. Polyp classification and suggested management strategy

Table 2. Classification of Extra-Colonic Findings

Table 3. Details of nine patients who suffered luminal perforation related to CT colonography

Table 4. Relation between patient category and observer assessment for all observer groups combined.

Table 5. Summary of lesion detection rates according to observer group.

Table 6. Observer accuracy and sensitivity according to experience for all lesions and when the six difficult-to-detect polyps were excluded. Accuracy refers to the percentage of correct classifications for patients with and without lesions whereas sensitivity refers to detection of cancers and polyps only.

Table 7. Ability to detect a second lesion (i.e. a lesion other than the index lesion) in those patients with more than one lesion, split by observer group. There were seven such patients (ED-3). Data in table refers to numbers of observations.

Table 8. Prior CT colonography interpretation experience of non academic subspecialist radiologists, split by whether examinations had been colonoscopically validated or not.

Table 9. Number and percentage of 15 CT examinations correctly classified by the 13 non academic subspecialist observers.

Table 10. Individual accuracy of non academic sub specialist radiologists
Table 11. Mean interpretation times in minutes (SD) and the number of cases interpreted (n) according to reference case category, observer group and for all observers.

Table 12. Effect of fatigue on interpretation times (Odds ratio).

Table 13. Effect of interpretation time on accuracy for each observer group. The regression coefficients indicate the amount that accuracy changes for a one-unit increase in report time (on the logarithmic scale).

Table 14. Agreement between observer CT measurements and colonoscopic reference.

Table 15. Comparison of absolute errors between visualization displays for four observers.

Table 16. Variability in measurement data: Components of variance (units = mm)

Table 17. Differences in polyp measurements between observers (units = mm).

Table 18. Intra-observer agreement.

Table 19. Agreement between measurement of polyp diameter by CT colonography and the colonoscopic reference diameter for each observer group.

Table 20. Frequency of individual polyp detection and correct categorisation by the three observer groups.

Table 21 Mean difference, standard deviation of the differences, and Bland Altman 95% limits of agreement for intra-observer comparisons of polyp diameter and volume measurements for two phantom types.

Table 22 Mean difference, standard deviation of the differences, and Bland Altman 95% limits of agreement for inter-observer comparisons of polyp diameter and volume measurements for two phantom types.
Table 23 Median difference, inter-quartile range, and significance for estimated values of polyp diameter and volume relative to the true value.

Table 24 Ratio of the difference between the estimated and known polyp volume, with 95% limits of agreement on a log-transformed scale (measurements taken from QRM phantom only).

Table 25. Inter-observer agreement.

Table 26. Agreement with colonoscopy.
LIST OF FIGURES

Figure 1. 3D 'endoluminal views' with white arrows depicting abnormalities
a. 10mm polyp in the ascending colon of a 62 years man
b. Same polyp as in 1a, close up.
c. Same patient with a synchronous sigmoid colon cancer, seen as an intra-luminal mass

Figure 2. a,b. 6mm sigmoid colon polyp in a 52 years woman seen on both prone (1a) and supine (1b) images.

Figure 3. a-d. 12mm caecal polyp seen as a filling defect in otherwise tagged fluid on
a. Prone axial display (Colon CT windows Width 1500, Level -150)
b. Magnified prone sagittal view
c. 3D endoluminal view (difficult to see as polyp only just protrudes above fluid level)
d. Endoscopic view of 12mm polyp subsequently confirmed as adenomatous polyp.

Figure 4. A; axial and, B; coronal. CT colonography images (supine acquisition) in a female patient with retroperitoneal perforation (table 2, case 3) – Black arrows outline the extraluminal retroperitoneal air surrounding the descending colon in Fig 4A and B. The patient was entirely asymptomatic and required no treatment. No obvious cause for the perforation was attributed.

Figure 5. A; sagittal and, B; coronal. CT colonography images (prone acquisition) in a female patient with intra- and retroperitoneal perforation (table 2, case 8) – White arrows outline retroperitoneal air seen surrounding the rectum and extending around both the ascending and descending colon in Fig. 5A and around the rectum only in
Fig 5B. This was attributed to a sigmoid carcinoma (not shown) and the patient underwent laparotomy, at which time the primary tumor was removed.

**Figure 6.** 62-year old woman with a 12mm sigmoid flat adenoma. Axial view. Images acquired using a four detector row scanner (100mA, 120 kV). This polyp was missed by all observers.

A; the adenoma (arrows) is barely visible using a standard CT colonography window (Hounsfield units level -150, range 1500, white arrows).

B; the adenoma is more clearly visible using standard abdominal CT windows (Hounsfield units level 40, range 400, white arrows).

**Figure 7.** 66-year old woman with a 8mm descending colon adenoma. Axial view. Images acquired using a four detector row scanner (100mA, 120 kV). This polyp was visible on the prone dataset only and was missed by all observers, excepting one experienced observer.

A; The polyp (arrow) is visible on the axial view using CT colonography windows; Hounsfield units level -150, range 1500.

B; 3D volume rendered endoluminal view.

**Figure 8.** Graph of observer accuracy overall, grouped by experience.

+= mean value: 74.2% for experienced observers, 66.6% for radiologists, 63.2% for technologists.

**Figure 9.** Graph of observer accuracy grouped by experience excluding 6 “difficult-to-detect” lesions.

+= mean value: 83.7% for experienced observers, 76.9% for radiologists, 72.0% for technologists.
Figure 10a,b Circumferential cancer located at the caecal pole on a. supine 2D axial view (arrows) and b. 3D endoluminal 'volume rendered' display (arrows): Correctly identified by 8 of 13 (62%) observers.

Figure 11 a, b. Large polyp (12mm maximal diameter), located in the recto-sigmoid on a. prone 2D axial view (arrows) and b. 3D endoluminal 'volume rendered' display (arrows): Polyp was only visible on the prone acquisition, partially submerged in untagged fluid and was correctly identified by 3 of 13 (23%) observers.

Figure 12 Medium sized polyp (8mm maximal diameter), located in the ascending colon on prone 2D axial view (ring); only visible on the prone acquisition (submerged in untagged fluid on supine) and correctly identified by 6 of 13 (46%) observers.

Figure 13. Interpretation times (in minutes) for each case, categorised according to observer group

Figure 14. Effect of interpretation time on accuracy for each observer group displaying the amount that accuracy changes for a one-unit increase in report time.

Figure 15. 65-year old male. CT colonography using a 4-detector row machine, collimation of 1.25mm, pitch 1.5, and mA 100. 6mm sigmoid polyp viewed on 2D 'colonography window' (A), 2D 'abdominal window' (B) and 3D 'surface rendered' view (C).

Figure 16 A; Effect of visualization display on polyp diameter measurement. Bland Altman dot plot comparing CT measurements with colonoscopy reference for each observer (1 to 4) according to visualization display. B; Interobserver agreement. Bland Altman dot plot comparing CT measurements with colonoscopy reference for each observer (1 to 4) according to visualization display.

Figure 17. A; Effect of visualization display on intraobserver agreement. Bland Altman dot plot comparing CT measurements with colonoscopy reference for each
observer (1 to 4) according to visualization display. B; Intraobserver agreement.

Bland Altman dot plot comparing CT measurements with colonoscopy reference for each observer (1 to 4) according to visualization display.

**Figure 18a-c.** Bland-Altman agreement plots for each of the three observer groups.

a. Experienced radiologists b. Less experienced radiologists c. Radiographic technicians. The middle solid horizontal line represents the mean difference between the pairs of measurements (difference between observer and colonoscopy estimate); essentially representing the mean 'measurement error'. The upper and lower horizontal lines represent the 95% limits of agreement i.e. the range of 'measurement errors' that would occur on 95% of occasions.

**Figure 19.** 7mm polyp located in the rectosigmoid colon and displayed on the supine axial image (Fig 2a. 'colon' window: W1500, L -150) and 3D volume rendered 'endoluminal' display (Fig 2b). Polyp overestimated as large by 5 observers: the line across the polyp (Fig 2a) represents software calipers placed by the author corresponding to a CT diameter of 7mm.

**Figure 20** 10mm polyp located in the rectum and displayed on the prone axial image (Fig 3a. 'colon' window: W1500, L -150) and 3D volume rendered 'endoluminal' display (Fig 3b). Polyp underestimated as medium by 9 observers: the line across the polyp (Fig 3a) represents software calipers placed by the author corresponding to a CT diameter of 10mm.
Figure 21 Representative image from the QRM lung phantom with cork granules removed from the phantom cavity. Only the nodules related to the 'chest wall' were measured as these best mimicked colonic polyps: There are both sessile and pedunculated types.

Figure 22 Representative image from the Whiting phantom showing mural polyps of both polypoid and sessile morphology.

Figure 23 Workstation screenshot showing a window for automated volume and diameter measurement of a 13mm pedunculated polyp.

Figs 24 a-e 10mm maximal diameter tubular adenoma located in the sigmoid colon of a 62 years old man; displaying measurements found using colon CT windows (Width 1500, Level −200, fig 1a); lung (Width 700, Level −600, fig 1b); abdominal (Width 400, Level 10, fig 1c); bone (Width 3500, Level 400, fig 1d) and using a 3D endoluminal view (fig 1e).

Fig 25 Automatic segmentation and extraction of the 10mm polyp (colonoscopic diameter) shown in Fig 24, which provided an assessment of diameter without the need for manual boundary identification.
ETHICAL APPROVAL STATEMENT

All of the patients contributing data to this thesis gave full consent (either verbal or written depending on the local stipulations of the European centre in question) for the primary study with which they were involved. Ethical permission for data sharing for the purposes of this thesis was obtained by the collaborators from contributing research centers.
GLOSSARY

CT: computed tomography
CTC: computed tomography colonography
SD: standard deviation
CI: Confidence interval
P: Probability value
CHAPTER 1

BACKGROUND, HYPOTHESIS, AIMS, AND STRATEGY

1.1 Background

Reader performance and experience

CT colonography is a technique whereby computerised tomographic (CT) images of the cleansed and distended colon are acquired primarily for detection of colonic neoplasia. Additional diagnostic information from extra-colonic organs may also be obtained. Introduced more than a decade ago [1], CT colonography is now increasingly regarded as the preferred radiological technique for examining the colon, likely improving on the performance characteristics and patient experience associated with double contrast barium enema [2-4].

However, the accuracy of CT colonography for detecting adenomatous colonic polyps, the precursor of colorectal cancer, is highly variable. Meta-analyses have shown sensitivity for large polyps (defined as 10mm maximal diameter or larger) and cancer as approximately 90% and 96% respectively but the synthesised data is heterogenous [5]. Indeed large polyp sensitivity for the three largest prospective, multi centre studies of CT colonography performance (all from the USA) varied between 52% and 94% [2;6;7].
The reasons for such variability are multifactorial and not fully understood. Critical analysis cites variability in reader experience and in the software/hardware specification utilised, researcher bias, and the methods used for interpretation [8-11]. Of these, the principal investigator of this thesis and co-workers hypothesised that reader inexperience is the most contributory variable [11]. Indeed it is unsurprising that the utility of CT colonography may decline when interpreted by observers with insufficient experience, particularly when compared to conventional colonoscopy performed by experienced practitioners [2;6;12].

There is a learning curve for CT colonography with interpretative accuracy generally improving with experience [13-15]. However, the curve is not always predictable and the performance of some individuals may even decline following directed training [14]. As a result, we are left with the challenging question of what experience of CT colonography is required to achieve competence? To date, the only responses have been based on anecdotal estimates by expert consensus, for example recommending review of 40 to 50 endoscopically validated cases [16].

Polyp measurement

Accuracy for detecting polyps and cancer is the first major challenge for CT colonography but unlike conventional colonoscopy, whereby detection is followed by polyp excision or biopsy for subsequent histological assessment, the biological significance and management strategy for polyps detected at CT colonography (i.e. whether subsequent polypectomy should be scheduled) is determined almost exclusively by their maximal diameter. As a result, appropriate polyp management strategies that are based upon diameter alone are in turn contingent upon accurate
polyp measurement. However, there are several potential confounders that conspire to thwart accurate measurement, such as choice of 2D versus 3D displays, the 2D anatomical plane chosen for measurement, and the CT window settings utilised. Prior to the research undertaken towards this thesis, no study had investigated the effect of such variables on estimates of polyp size and subsequent categorisation of patient risk.

**Geographical and population differences**

In the UK and in many European centres, the clinical role for CT colonography is likely to be as a relatively non-invasive test for investigating patients with symptoms possibly indicating colorectal cancer. This contrasts clearly with the focus in the USA, where CT colonography has been promoted as a primary screening technique for colorectal cancer and its precursor, the adenomatous polyp. These different clinical environments underline the importance of examining how CT colonography performs in different geographical locations and clinical environments where, for example, access to high technology imaging platforms or radiology training may differ. This thesis has focussed on CT colonography practice in the UK and Europe.

The aim of this thesis is to investigate how CT colonography has been implemented in the UK and any complications encountered, and to investigate the effect of directed training on reader performance across Europe for several different observer groups: Experienced radiologists and inexperienced but trained observers; radiologists and radiographers. Performance of sub-specialist radiologists, who have not been trained formally but who offer a CT colonography service in day-to-day UK clinical practice is also compared to these observer groups to examine generalisability. Finally, accuracy
of polyp measurement and the effect of different viewing conditions on assessment of polyp diameter is investigated, coupled with assessment of automatic measurement computer software in both in-vitro and in-vivo studies.

1.2 Hypothesis and Aims

1. There is considerable interest in CT colonography amongst UK radiologists but there is a perceived lack of CT scanner capacity and reader experience that might limit implementation. Consequently, we are unaware how widely CT colonography has disseminated into routine clinical practice.

Aim: To determine the availability of CT colonography in the UK NHS, the technical parameters used, the number of examinations performed, by whom, and their experience.

2. It is claimed widely that CT colonography is free from serious adverse-events but isolated case-reports of potentially serious complications are starting to emerge and some abstracted work on this topic is marred by spectrum bias. As a result, the incidence of serious adverse events associated with CT colonography performed in routine clinical practice is currently unknown.

Aim: To determine retrospectively the incidence of potentially serious adverse events associated with CT colonography performed in symptomatic patients.
3. Training and experience likely improves reader performance at CT colonography but to what extent? In particular, it is unclear to what extent currently recommended training schedules affect observer performance, and whether they are adequate for their intended purpose.

_Aim:_ To define the interpretative performance of radiologists experienced in CT colonography and to compare this with novice observers who have undergone a period of directed training.

4. Reader performance for CT colonography has only been investigated in academic centres, whereas the majority of CT colonography examinations are performed in a non-academic environment.

_Aim:_ To investigate observer accuracy and report times for radiologists performing CT colonography in routine clinical practice, outside an academic environment.

5. Successful implementation of CT colonography demands time efficient interpretation but report times and accuracy are potentially influenced by a number of variables.

_Aim:_ To assess the effect of reader experience, fatigue, and scan findings on interpretation time for CT colonography in a multi-centre setting.

6. The biological significance of colonic polyps detected by CT colonography is determined by their maximal diameter but it is unclear how accurate the measurements made using CT colonography are in reality.
**Aim:** To determine inter- and intra-observer agreement for estimates of polyp diameter made from CT colonography, including the effects of different visualisation displays and prior expertise.

7. Polyp management strategies are contingent upon accurate measurement of maximal polyp diameter and subsequent correct categorisation of the polyp. **Aim:** To determine to what extent measurement error on CT colonography influences polyp categorisation according to established management guidelines.

8. Manual assessment of maximal polyp diameter is subject to interobserver measurement error. It is possible that this could be reduced by automated measurement. **Aim:** To investigate inter- and intra-observer agreement of automated measurement of polyp diameter in vitro.

9. Any additional benefit of automated measurement software found in vitro may not apply in vivo because the circumstances of the measurement are more unpredictable. **Aim:** To compare automated polyp measurement with conventional manual assessment in vivo.
1.3 Thesis strategy

Chapter 2 reviews the technique, diagnostic performance and clinical role of CT colonography predominantly via an extensive literature review. The fundamental technical parameters and interpretation methods utilised are discussed, followed by a review of published performance characteristics and generally accepted current clinical role.

Chapter 3 investigates the current provision of CT colonography in the UK and also the incidence of potentially serious adverse events, via two national surveys of UK radiologists. The wide availability of CT colonography is revealed, and barriers to more extensive implementation are discussed. The complication rate for CT colonography is determined and strategies for avoiding colonic perforations are provided.

Chapter 4 defines the range of interpretative performance of radiologists experienced in CT colonography when asked to interpret a validated dataset of cases, and compares their performance with novices (both radiologists and radiographers) who have been trained using a schedule typical of generally accepted current guidelines. The relative performance of sub-specialist gastrointestinal UK radiologists, with no formal training in CT colonography but who still offer a service in day-to-day practice, is also investigated.
In chapter 5, interpretation times for CT colonography are investigated and the influence of experience and fatigue on report times is assessed.

Chapter 6 deals with accuracy of polyp measurement and investigates the effect of different CT viewing conditions on inter- and intra-observer agreement. The effect of measurement error on subsequent polyp categorisation and patient management is also investigated in a multi-centre setting. Finally computer software used to automate polyp measurement is compared to manual assessment in both in-vitro and in-vivo scenarios.

Chapter 7 concludes this thesis with a discussion of the main findings of the thesis and the implications for CT colonography interpretation in routine clinical practice.
CHAPTER 2

CT COLONOGRAPHY:
A review of technique, diagnostic performance
and clinical role

2.1 INTRODUCTION

2.1.1 Historical background

Advances in computed tomography (CT) technology have paved the way for increasingly sophisticated diagnostic techniques. The introduction of helical and, more recently, multi-detector row CT technology, has greatly reduced scan acquisition time along with improvements in z-axis resolution. Acquisition of these data is inherently 3-dimensional and consequently, reconstruction of CT data into clinically useful multi-planar and 3-dimensional displays became feasible.

In 1994, David Vining and colleagues harnessed these new CT data rendering techniques to create the first virtual ‘fly through’ of the colorectum, simulating the passage of an optical colonoscope [1]. The underlying technique used to provide this data utilises thin collimation abdominopelvic CT scan acquisitions of the cleansed colon, which has been distended via a rectal catheter. The complete technique including the use of 3D perspective rendering for interpretation was termed “virtual colonoscopy” or alternatively “CT colonography” [17].
The introduction of this new, relatively non-invasive radiological technique for examining the colon captured the attention of radiologists, other health professionals and the general public alike, not least because of the attractions of the novel endoluminal display. However, a decade ago, the time required to reconstruct data into 3D endoluminal displays, meant that CT colonography was excessively time consuming and required very high specification computers with sufficient memory to permit sophisticated data processing. As a result, CT colonography was impractical for implementation into routine clinical practice at that time. However over the intervening years, rapidly advancing hardware and software technology solutions have provided increasingly sophisticated, time-efficient methods of interpretation, increasing the potential for CT colonography to be utilised in a day-to-day clinical environment. Currently, data can be rendered in a matter of minutes and interpretation can be performed in 10 to 15 minutes, with the rate-limiter being the observer rather than computational power.

2.1.2 Early landmark clinical studies

Early single centre studies compared CT colonography to its conventional counterpart, optical colonoscopy [17-20] using it as a reference-standard. The first of these by Hara and co-workers (including David Vining) in 1996[20] based their scanning protocol on their experience of scanning an artificial colon phantom. They subsequently scanned ten patients who then underwent same day conventional colonoscopy. The two radiologist observers in this study detected 100% (5 of 5) large polyps (10mm or larger), 71% (5 of 7) medium polyps (5 to 9mm) and 11% -28% (2-5 of 18) small polyps (<5mm). Hara concluded that CT colonography was feasible for polyp detection>or = 5mm in diameter. This and other studies showed promising
results but patient numbers were small. However two landmark studies were published in 1999 and 2001, with larger patient numbers, demonstrating that CT colonography had excellent sensitivity for detecting clinically important polyps and cancer [21,22]. These are described in greater detail below:

Between 1997 and 1999, Fenlon et al [21] conducted a prospective study of 100 patients (60 men, 40 women) at 'high risk' for colorectal neoplasia, comparing findings at CT colonography (acquired using a single detector CT scanner) with immediate subsequent optical colonoscopy. This study examined the performance of CT colonography via a per-polyp analysis, whereby each polyp found at CT was only deemed a true positive finding if it was matched to a corresponding polyp found at optical colonoscopy. Results were also analysed on a 'per-patient' basis whereby the CT was considered to be truly positive when at least one polyp identified on CT was matched to a lesion seen at optical colonoscopy. All other finding at CT were considered to be false positives.

Results of this study showed that, of 100 patients undergoing both investigations, 51 had normal findings by CT; and of the remaining 49, a total of 115 polyps and 3 cancers were found by CT. CT colonography correctly identified all 3 cancers and 82 of 115 polyps (71% per-polyp sensitivity for all polyp sizes). As one might expect, sensitivity increased with increasing polyp diameter such that CT identified 20 of 22 polyps that were 10 mm or more in diameter (91 percent), 33 of 40 that were 6 to 9 mm (82 percent), and 29 of 53 that were 5 mm or smaller (55 percent). On a per-patient basis, the results were improved (since one patient may harbour multiple polyps) with an overall sensitivity of 82% and specificity of 84% for all polyp sizes;
and for polyps measuring 10 mm or larger, the sensitivity, specificity, and positive and negative predictive values for CT colonography were each 96%.

As a result, the authors concluded that in patients with high risk for colorectal neoplasia, CT colonography and optical colonoscopy have similar efficacies for the detection of polyps 6 mm or more in diameter. The authors highlighted the low sensitivity of CT for diminutive polyps (5mm or less in maximal diameter) and predictably but sensibly challenged the need for universal polypectomy for such polyps, when the risk of these polyps harbouring malignancy is very low. Whether removal of diminutive polyps is outweighed by polypectomy-related adverse events remains a subject of heated debate between gastroenterologists, surgeons, and radiologists, and is complicated and confounded by the realisation that polyp size assessment by either CT or optical colonoscopy is itself prone to error.

This paper by Fenlon et al was published in the New England Journal of Medicine[21] and was a milestone in the development of CT colonography but notably, CT interpretation was undertaken in consensus by experienced gastrointestinal radiologists, and the time taken for interpretation was not recorded. Also conspicuous was the comment that CT colonography was a potentially useful tool for screening, mentioned in the first few paragraphs of Fenlon’s paper, even though all the patients recruited were high risk and not representative of an asymptomatic screening population. Indeed, it is noteworthy that there have been several examples in the literature where performance characteristics based on studies of symptomatic patients have been used to fuel the debate for utility of CT colonography as a screening test and also vice versa (these studies are detailed below). It is important to
recognise this difference as prevalence of colorectal abnormality and extracolonic
disease is significantly different between the two populations and technical
considerations, for example CT scan dose and use of intravenous contrast, will also
deriffer. Most obviously, symptomatic patients have larger lesions which will be easier
to detect (by both CT and colonoscopy).

In 2001, Yee et al published a larger series of 300 patients (291 men, 9 women:
conducted in a war veteran's hospital setting) who also underwent CT colonography
(also using single detector CT technology) followed by optical colonoscopy 2-3 hours
later[22]. In this study, the patients were either asymptomatic for screening (96
patients) or symptomatic (204 patients). Overall 524 polyps and 8 cancers were
detected in 182 patients. Once again, readers interpreting the CT colonography
examinations, identified all the cancers (n=8) and the overall sensitivity for polyps was
similar to the Fenlon study at 70%. Moreover polyp sensitivities according to size
category closely matched the results from the Fenlon study with sensitivity for larger
polyps (10mm or larger) of 90%; for medium polyps (5-9mm) of 83% and for small
polyps (less than 5mm) 67%. On a per-patient basis, the overall sensitivity and
specificity were 90 and 72% and there was no significant difference in performance
characteristics for CT colonography between asymptomatic and symptomatic patient
groups.

Again, similar to the Fenlon study, there were two readers, who interpreted the cases
independently and then together in consensus. There were no attempts to investigate
agreement between these observers but median interpretation time for each was
recorded as 31 minutes (range 15-45 minutes) and 27 minutes (15-40 mins) respectively.

These studies raised the profile of CT colonography and showed it was feasible for use in both symptomatic and asymptomatic patients with performance characteristics approaching expert optical colonoscopy. Moreover, CT Colonography appeared adept at preferentially detecting adenomas rather than less significant hyperplastic polyps [21;22], clearly because adenomas are larger on average. Support for CT colonography was fuelled further by studies advocating its superior patient experience and safety profile when compared to optical colonoscopy [3;4;23;24].

However, since these early studies, CT colonography has been subject to intense scrutiny. There have been several multi-centre, multi-observer studies showing wide variability in performance. As a result, critics have raised concerns about the generalisability of this technique whereas advocates have cited the use of older technology; outdated methods of interpretation, reader bias and reader inexperience as reasons for poorer performance. The studies contributing to this debate are summarised later in this chapter, and the questions raised, particularly related to reader performance are in part addressed by this thesis.
2.2 TECHNIQUE

Currently there are several fundamental prerequisites for CT colonography, which have been endorsed by expert consensus [25]. At the present time these include full bowel catharsis; gaseous colonic distension (using air or carbon dioxide); dual patient positioning for CT scan acquisition (generally supine and prone); and dedicated CT colonography software for interpretation, incorporating a facility for both 2D and 3D luminal review.

2.2.1 Bowel preparation

Full catharsis prior to CT colonography remains standard although regimens with reduced or even no purgatives may become the norm in the near future. Optimal bowel preparation regimes differ between optical and virtual colonoscopy because significant quantities of intraluminal fluid, easily aspirated with optical colonoscopy, only serve to obscure the luminal surface in CT colonography. Purgation should leave a dry, clean colon, with minimal residual fluid and faeces, because these potentially hide or mimic colorectal neoplasia. Standard laxatives include sodium picosulphate (UK) and sodium phosphate (USA).

2.2.2 Colonic distension & patient positioning

Colonic distention is achieved by gaseous insufflation of air or carbon dioxide via a rectal catheter with the patient lying on the scanner table. Optimising colonic distension is perhaps the single most important technical parameter for ensuring confident, rapid and accurate interpretation [26-28]. In poorly distended colonic segments, the wall and haustral folds appear thickened thus mimicking neoplasia.
while luminal collapse can easily hide significant pathology [29-31]. There is a reasonable evidence base supporting the various techniques employed to optimise distension. Multi detector row CT scanners permit faster scan times than do older single detector technology [32], with the latest generation machines enabling the entire abdomen and pelvis to be examined in very short (a few seconds) 'breath-hold' acquisitions. This virtually eliminates respiratory related artifact and reduces suboptimal distention due to absorption or escape of colonic gas. Dual patient positioning i.e. supine and prone scan acquisitions or lateral decubitus if the patient has difficulty lying prone [33] provides an opportunity to review each side of the colonic wall in a non-dependant position. The result of repositioning the patient in this way is to redistribute the intraluminal gas, faeces and fluid, thus revealing segments that may otherwise be partly obscured. Although adherent faeces remain a potential cause of false positive lesions, dual positioning improves diagnostic performance and is generally universally recommended [25-28]. Intravenous hyoscine butylbromide, a spasmolytic (unlicensed for use in the USA), has been shown to significantly improve colonic distention [34-36], particularly in patients with diverticular disease [35]. As a result, it is widely used across Europe. Glucagon, an alternative available in the United States, has not been shown to be beneficial [37;38] and is therefore not recommended.

Narrow caliber rectal catheters [34] are preferred, being more comfortable for patients and functioning as well as their wider bore counterparts for distending the colon. In patients with severe anal incontinence, judicious use of a balloon catheter is advocated. However, this manoeuvre carries an increased risk of rectal perforation
[39] and should only be performed carefully by an experienced clinician following digital rectal examination to exclude a rectal tumor.

Choice of insufflating gas (air or carbon dioxide) and method of gas delivery (manual or automated) depends on local experience and availability. The majority of practitioners insufflate the colon with room air delivered manually by squeezing a plastic bulb; undoubtedly a cheap, easy, and effective option. However, evidence from the barium enema and colonoscopy literature suggests that carbon dioxide causes less patient discomfort than air due to its rapid mucosal absorption [40;41], and also facilitates easier subsequent colonoscopy if a patient is referred immediately after CT.

While carbon dioxide delivery is more complex, it can be administered using an automated insufflation device specifically designed for CT colonography (ProtoCol insufflation system, EZEM inc, Westbury, NY). This device has been shown to significantly improve colonic distention when compared to room air or manual carbon dioxide insufflation [42;43]. Furthermore, the device permits a relatively ‘hands free’ approach and the facility to monitor intrarectal pressure, whilst providing a continually updated display of the total volume of gas administered. Additionally, it has a safety shut-down mechanism, preventing further gas delivery if the rectal pressure exceeds the limit (up to 25mmHg), preset by the operator.

2.2.3 Intravenous contrast administration

The use of intravenously administered contrast for CT colonography was first described in 1997 [44] and since then has been advocated for improving detection
and characterization of colonic and extracolonic pathology [45,46]. Morrin and colleagues demonstrated potential benefits of increased reader confidence, improved bowel wall conspicuity and increased reader accuracy for medium sized (6-9mm) polyps by using intravenously administered contrast [45]. Such benefits were particularly important in poorly prepared examinations either due to excessive faecal or fluid residue, where the enhancing polyps can be more readily depicted. More recently, Neri and others have shown significantly increased soft tissue attenuation in both benign polyps and cancers found at CT colonography with the use of intravenous contrast, postulating that increased conspicuity will improve reader accuracy [46,47].

The administration of intravenous contrast increases detection and characterisation of extracolonic pathology overall but the benefit of this is unclear as the majority of such findings are clinically unimportant. Indeed the incidence of important extracolonic pathology (significant aortic aneurysm and cancer) is relatively low (3.6%) [48] and may be similar in both asymptomatic screening and symptomatic populations [49]. Furthermore, the majority of important findings are large and therefore relatively conspicuous even using non-enhanced scans.

While there are potential benefits of routinely using intravenously administered contrast for CT colonography, there are also detrimental effects to clinical workflow and patient throughput. Moreover examination costs are increased and there is a small but significant risk of a contrast reaction. With these factors in mind, experts have generally advocated the routine use of intravenous contrast for patients with symptoms that may be attributable to colorectal cancer [50], but not for asymptomatic
screening populations[25]. However further outcome data is required to help refine these strategies further.

2.2.4 Data analysis

Following scan acquisition, CT data is transferred to a computer workstation loaded with dedicated virtual colonoscopy software, for subsequent interpretation. At the time of writing there are two main paradigms for image interpretation; a primary 2D-axial review, reserving 3D 'endoluminal' displays for problem solving [17]; and a primary 3D 'endoluminal flythrough' review, using 2D displays for problem solving [51]. The 3D 'flythrough' method renders the interface between the colonic gas and the adjacent mucosal surface and thus simulates the optical colonoscopists' view – hence the other frequently used term for CT colonography, 'virtual colonoscopy' (Figure 1a-c). Polyps are thus displayed as protruberances from the endoluminal surface, just as in optical colonoscopy. In contrast, the 2-D approach requires the radiologist to look at the axial CT images only (Figure 2a,b).

Because the 3D method requires considerably more computer power and time (both for rendering and interpretation), it was initially thought too cumbersome by most investigators (see chapter section 2.1.1 above). Initial studies suggested a primary 2D approach was as accurate and considerably more time efficient than using a full endoluminal flythrough, largely due to inherent limitations in computer software capabilities [52]. However recent technological advances have meant that an increasing number of workstations are capable of producing essentially instantaneous automated flythrough projections. This, coupled with the fact that the authors of the largest and most successful CT colonography study to date (see below) have
specifically attributed most of their success to the use of a primary 3D method of review [7], has refocused attention on the analysis method used. This is currently a topic of considerable debate amongst radiologists and has stimulated further research. Three recent (albeit small) studies [53-55] comparing the diagnostic performance of primary 2D versus primary 3D interpretation have shown no significant difference between the methods (although interpretation time was significantly longer for the primary 3D review). However, one of these studies [53] used observers who were mostly familiar with primary 2D interpretation, so the findings may have been subject to spectrum bias. Moreover, a recent meta-analysis [56](see below) has shown greater polyp sensitivity when 3D methods of analysis are used.

Ultimately it is now highly likely that interrogation will be performed using a combination of 2D and 3D review, the balance of which will be dependent on the results of further studies, radiologist preference and the particular patient case in question. For example a poorly prepared patient with excessive faecal residue produces numerous false positive polyp candidates on the 3D-endoluminal display, and would lend itself more to 2D analysis where the typical characteristics of faeces (heterogenous appearance containing tiny pockets of air) is more easily appreciated.
Figure 1.

3D ‘endoluminal views’ with white arrows depicting abnormalities

a. a. 10mm polyp in the ascending colon of a 62 years man
b. Same polyp as in 1a, close up
c. Same patient with a synchronous sigmoid colon cancer, seen as an intra-luminal mass.
Figure 2. a, b.
6mm sigmoid colon polyp in a 52 years woman seen on both prone (2a) and supine (2b) images.

Fig 2a.
2.3 CURRENT CLINICAL ROLE

Since its inception, CT colonography has been advocated for use in both symptomatic and asymptomatic patient populations. From the above discussion it is clear that at the time of writing its clinical role is evolving and ultimately will likely reflect local resources and radiological/endoscopic practice.

2.3.1 Symptomatic population

Most CT colonography studies to date have focused on polyp detection, emphasising a potential role for CT colonography for colorectal cancer screening and prevention, by identifying patients for polypectomy. However, its role in the investigation of symptomatic patients, in whom the prime target is cancer, has been somewhat underemphasised [57]. This is especially surprising since the vast majority of studies have actually examined symptomatic patients. Only one meta-analysis has specifically extracted data relating to cancer detection [5], showing that CT colonography is highly sensitive (96%) in this context.

Amongst the generally accepted indications for CT colonography is its use in symptomatic patients whose colonoscopy has failed [58;59], frequently because of an obstructing tumor, diverticular stricture or redundant colon, and there is now very good evidence to support this role [58-60]. However, use in frail and elderly patients, where the target pathology is cancer, is becoming increasingly common despite a lack of hard evidence-base in this population. Standard abdominal or minimal preparation CT (without the use of any bowel purgation) remains an alternative in this
patient group, although sensitivity for this technique is likely in the region of 85% with an associated relatively poor positive predictive value of just 0.5 [61]. Although comparative trial data is lacking, CT colonography would appear the pragmatic alternative for patients in whom optical colonoscopy with intravenous sedation is inappropriate, but are fit enough for bowel preparation.

Despite the increasing utility of CT colonography, barium enema remains the standard radiological colonic investigation, particularly outside the USA; 300,000 of these examinations are performed in the UK alone each year. As discussed above, CT colonography has several potential advantages over barium enema; it is possibly more accurate [2;62]; visualises extracolonic organs (which may or may not be useful), is strongly preferred by patients [3;4;23]; and the consequences of anal incontinence and perforation are less dramatic. Radiation dose is equivalent or even less for CT colonography when ‘low dose’ and ‘dose modulation’ techniques are employed [63;64]. Also, perhaps of greater concern, interpretative skills in barium enema are diminishing, even amongst experts [65], as fewer radiologists actually perform the technique [66] coupled with an undoubted preference for newer cross sectional techniques amongst younger radiologists [67]. It is unclear at the time of writing whether CT colonography is more or less difficult to interpret than barium enema.

Overall it is therefore likely that a significant proportion of barium enema work will transfer to CT colonography over the next few years, although at present the required infrastructure (CT scanners, trained radiologist) is often lacking. It should be stressed that at the time of writing, CT colonography is not recommended for investigating
patients with inflammatory bowel disease, because identifying true colonic neoplasia from regenerative epithelium and benign strictures is fraught with uncertainty. Moreover patients with severe disease are at greater risk of perforation [68].

2.3.2 Screening population

At the time of writing, CT colonography has yet to be advocated as a primary screening tool by any national or expert organisation and generally remains non-reimbursable by the major health care providers in the USA. When demonstrably successful, CT colonography is potentially reimbursable as an effective first line screening investigation when complemented by optical colonoscopy for those in whom pathology is detected [69].

In anticipation and in support of reimbursement, there is a working group of radiologists in the USA who have formulated polyp management strategies based on findings at screening colonoscopy, shown in table 1 [70].
Table 1. Polyp classification and suggested management strategy

<table>
<thead>
<tr>
<th>Classification of Exam</th>
<th>Meaning</th>
<th>Management strategy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inadequate Study</td>
<td>Inadequate preparation or distension</td>
<td>Await prior colon studies for comparison</td>
</tr>
<tr>
<td>Normal colon</td>
<td>No colonic abnormality</td>
<td>Routine screening (7 yearly)</td>
</tr>
<tr>
<td>Benign Lesion</td>
<td>No polyp &gt;5mm Lipoma Non-neoplastic abnormality e.g. diverticular disease</td>
<td>Routine screening (7 yearly)</td>
</tr>
<tr>
<td>Indeterminate Lesion</td>
<td>&lt;3 Polyps 6 to 9mm</td>
<td>Shorter term surveillance (within 3 years*)</td>
</tr>
<tr>
<td>Polyp; possibly advanced adenoma</td>
<td>&gt;2 Polyps 6 to 9mm Polyp &gt;9mm</td>
<td>Refer for colonoscopic removal</td>
</tr>
<tr>
<td>Colonic mass, Likely malignant</td>
<td>Causing luminal narrowing</td>
<td>Urgent surgical referral**</td>
</tr>
</tbody>
</table>

* subject to local practice and patient preference (consider endoscopy).

** Communicate with referring physician: endoscopic biopsy may be indicated.
However further widespread implementation of screening CT colonography still
remains hampered by concerns over variable trial performance, generalisability of
results, uncertainties over target lesions and subsequent polyp management, and the
impact of detected extra-colonic pathology. This is clearly a rapidly changing field,
and in the light of current research, new technology, and the aim of the American
College of Radiology to introduce quality assurance standards for virtual colonoscopy,
the American Cancer Society are planning to review the role of virtual colonoscopy
[71].

2.4 PERFORMANCE

2.4.1 Diagnostic accuracy

As mentioned above, the data from studies investigating the utility of CT
colonography should be considered according to the patient group examined. The
following summary of diagnostic performance studies will attempt to do so.

2.4.1.1 Single centre studies

Although the number of larger multi-center studies is increasing, the majority of CT
colonographic performance data is derived from single centre studies that have
recruited patients scheduled for colonoscopy (which has also been used as the
reference test). The larger studies by Fenlon et al and Yee et al are described in
detail above but there are two subsequent studies which should be highlighted.

The first from Rome, published in 2002 by Laghi et al [72] used both single and multi-
detector row CT scanners to investigate 185 consecutive patients (79 men, 86
women) with symptoms of colorectal cancer, followed by same day optical colonoscopy. A total of 30 colorectal cancers and 37 polyps (12 were 10mm or larger) were found in 52 patients using optical colonoscopy, which was incomplete in nine cases (6%). CT colonography identified; all thirty cancers (100%) and successfully examined the proximal colon for those nine with incomplete colonoscopy; 29 of 37 (78%) polyps overall; 11 of 12 (92%) large polyps (10mm or larger); 14 of 17 (82%) medium polyps (6-9mm) and 4 of 8 (50%) small polyps (5mm or smaller). Overall per patient sensitivity and specificity were 92% and 97% respectively. These performance characteristics are similar to the studies by Fenlon and Yee and emphasise the utility of CT colonography for cancer diagnosis and staging, as these patients also had contrast administered intravenously.

The second was a prospective multi-observer study from the Mayo clinic [12] which investigated the performance of CT colonography compared to same day optical colonoscopy in 703 asymptomatic patients (442 males, 261 females; mean age 64 years, range 50-84 years) at higher than average risk for colorectal cancer. Either a single or 4-row multi-detector CT scanner was utilised and interpretation of each patient's scan was performed by two of three radiologists, experienced in abdominal radiology and with greater than 150 CT colonography case interpretations each. Results were reported for each radiologist and also for double reading, combining the reports of two individual readers. Readers were told to ignore polyps measuring less than 5mm in maximal diameter.

Overall 153 polyps measuring 5mm or larger in diameter were found by optical colonoscopy. The per-polyp sensitivities for detecting large polyps (10mm or larger)
for the three readers and for double reading was 34%, 32%, 73% and 63% respectively. Corresponding figures for medium polyps (5-9mm) were 35%, 29%, 57% and 54%. Per patient sensitivity for identifying patients with at least one polyp 5mm or larger ranged from 41 to 69%, with a double reading sensitivity of 65%.

These data showed high inter-observer variability in performance despite similar experience and training amongst the three readers. Also polyp detection was considerably lower than for optical colonoscopy. The reasons underpinning such poor results were probably multi-factorial but the authors cited the low prevalence of abnormality as the major factor. While this study was not a truly screening population, the prevalence of abnormality was relatively low compared to prior studies (5% of patients harboured polyps measuring 5mm or larger in maximal diameter, compared to for example 19% in Fenlon’s study and 16% in Yee’s study). Other likely factors included older CT hardware and software technology including relatively thick scan collimation (5mm) and possibly the use of non proprietary customised software.

This was one of the first studies to examine individual performance rather than interpretation by consensus between two or more readers. The authors found the reasons for the considerable inter-observer variability were difficult to explain although previous studies [73;74] had also found variability. They suggested that experience and training were unlikely causative factors of inter-observer variability in their study but they did suggest that double reading, more efficient 3D displays and the assistance of computer aided diagnosis (CAD), may help to improve agreement. These were clearly topics for further investigation and several, including reader agreement and the effect of directed training on reader performance, have been
addressed in this thesis. This thesis has also investigated the possible benefit of CAD for polyp measurement.

2.4.1.2 Multi-centre screening studies

The largest screening study to date, comparing CT colonography to same day colonoscopy in 1233 asymptomatic individuals (728 men, 505 women; mean age 57.8 years) was published in the New England Journal of Medicine by Pickhardt et al [7]. This study involved three centres and six board radiologist readers, each trained using at least 25 CT colonography examinations (two had interpreted more than 100 studies each). This study also employed the technique of segmental unblinding to enhance the reference standard, optical colonoscopy.

Segmental unblinding involves a study co-ordinator revealing the CT colonography report, segment by segment, during endoscopic withdrawal from the caecum thus providing an opportunity for the colonoscopist to re-examine a segment if there was a discrepancy between the CT report and the endoscopic findings. Utilising this technique, excellent performance characteristics were obtained for CT colonography; sensitivity of 94% and specificity of 96% for polyps measuring 10mm or larger, 94% and 92% for polyps measuring 8mm or larger; and 89% and 80% for polyps measuring 6mm or larger. These results compared favourably to expert optical colonoscopy (99.4% caecal intubation rate) which had sensitivities of 88%, 92% and 92% for the three polyp size thresholds. Moreover, two polyps were malignant and one of these, an 11mm diameter polyp was missed on optical colonoscopy prior to unblinding of the CT findings. Indeed, segmental unblinding affords a fairer
comparison between CT and optical colonoscopy so that optical false negatives are not wrongly presented as CT false positives, where this is the case.

According to the authors, the impressive results reported from this study were heavily dependent upon the primary 3D method of interpretation, which they deemed to be the most important difference in methodology from underperforming studies (see below). Subsequent studies have either supported this assertion or shown no difference [53;56]. However, most researchers agree that 3D tends to be more time consuming [53]; the mean interpretation time across all three centres was 20 minutes, although this was reduced to 17 minutes in the second half of the study. The relative time for interpretation using a primary 2D method of interpretation, and the effects of fatigue across readers with differing experience has been investigated in this thesis.

For many radiologists, this study provided the necessary evidence required to support widespread implementation of CT colonography as a primary screening test. However, detractors point to the lack of generalisable data and data published by gastroenterologist researchers (see below) questioned the generalisability of CT colonography, resulting in the demand for further evidence.

2.4.1.3 Multi-centre symptomatic studies

Two gastroenterologist led, prospective multi centre studies were published soon after Pickhardt’s paper, showing considerably poorer results[2;6].

The first of these was by Cotton et al [6], published in the Journal of the American Medical Association. This study investigated 615 patients (45% men, 55% women;
mean age 61 years), with symptoms of colorectal cancer, recruited between 2000 and 2001 from 9 hospital centres. Recruitment was slower than expected and highly variable between centres ranging from 10 to 188 participants per centre. CT colonography was performed using older CT technology and slice collimation of up to 5mm. Reader experience was also limited with the study requiring readers to have an experience of only 10 previous CT colonography examinations. In contrast, endoscopists were highly experienced as evidenced by caecal intubation rates during this study of 99%.

Per patient sensitivity for detecting those individuals with a polyp 10mm or larger with CT colonography was 55%, and for those with a polyp 6mm or larger was 39% (specificity 90%) compared to optical colonoscopy which achieved corresponding sensitivities of 100% and 99% (specificity 100%) respectively. Notably, the only centre with significant experience prior to study commencement contributed the largest number of patients by far (182) and also had the best per patient sensitivity of 82% for detection of patients with a 6mm or larger polyp. Indeed, excluding this centre, the combined sensitivity for all the other 8 centres was 24%.

Perhaps unsurprisingly, this study was heavily criticised by a number of different groups active in CT colonography research [8-11]. Older technology, reader inexperience and researcher bias were cited as the main criticisms. Indeed, comparing novice CT colonography interpretation with expert optical colonoscopy is not comparing 'like with like' and the wide variability in performance between centres of differing experience supports this criticism.
A further study [2], again led by a gastroenterologist, was published in the Lancet in 2005 and compared CT colonography with optical colonoscopy and double contrast barium enema in 614 patients (70% men, 30% women; mean age 57 years) with symptoms of colorectal cancer. While multi-detector CT technology was used, many readers were still relatively inexperienced compared to the gastroenterologists performing the optical colonoscopies, with approximately half having interpreted less than 50 CT colonography cases (although they also received an additional training module) compared to highly experienced gastroenterologists with a mean of 14 years endoscopy experience.

Per-patient sensitivity for all polyps 10mm or larger were 59% for CT colonography, 48% for barium enema and 98% for optical colonoscopy; for polyps measuring 6-9mm, corresponding sensitivities were 51%, 35% and 100%. Once again these data suggest CT colonography underperforms compared to optical colonoscopy in symptomatic patients.

The accompanying editorial by Halligan [75], emphasised the potential for CT colonography to be used as a non-invasive test for cancer, rather than simply a screening test for polyps. As such, its main comparator should be barium enema, rather than optical colonoscopy, and the results of Rockey’s study reinforced CT colonography’s performance superiority over barium enema. This editorial also highlights the potential of lead-researcher bias with the assertion that future studies should be supervised independently by people with no vested interest in the outcome [75].
Collectively these two studies questioned the generalisability of CT colonography, and dampened the enthusiasm for widespread implementation, despite the glaring differences in methodology from Pickhardt's study.

2.4.1.4 Meta-analyses

Since 2003, in an attempt to obtain more reliable data, three meta-analyses have been performed [5;56;76]. One of these considered all relevant publications up to and including February 2005 [56]. Overall, 33 studies involving 6393 patients were included; 63.6% of participants were male, mean age of 61.9 years and 74% were at high risk for colorectal cancer. Analysis focused on per-patient data, deemed most applicable to a screening test i.e. correctly identifying patients who should be referred subsequently for optical colonoscopy.

Average sensitivity for detecting large polyps (i.e. 10mm or larger) was 85% (average specificity 97%) falling to 70% (93%) respectively for polyps 6 to 9mm, and 48% (92%) for polyps less than 6mm diameter. However, sensitivity for large polyps for individual studies varied between 48% and 100%, with subgroup analysis suggesting that narrower CT scanner slice collimation (i.e. slice thickness), use of multidetector-row scanners, and perhaps a greater proportion of analysis using 3D displays all contributed to improved performance. It should be noted that the meta-analysis [56] was performed on sensitivity and specificity separately, with no account of their correlation (threshold effect), which is particularly important in meta-analysis of diagnostic tests (as opposed to treatment effects from randomized trials). A Spearman's test was used to investigate a threshold effect but this approach will likely underestimate overall estimates of sensitivity and specificity since the sROC is a
curve (i.e. the summation of individual ROC curves from each study) rather than a straight line. Threshold effects are particularly important in studies of diagnostic accuracy because different observers reading the test will likely have different individual thresholds for a positive result, as will different studies; some studies restricted their attention to polyps 1cm or more for example.

The investigators also found marked heterogeneity between individual studies [56], broadly suggesting that their methodology is different. This is typified by the results from three individual prospective multi-center studies described above [2;6;7]. In contrast to variable sensitivity, CT colonography is consistently highly specific in all three meta-analyses (>95% for large polyps). Of course, the low prevalence of abnormality in a screening population means that specificity will not suffer as much as sensitivity in the context of a relatively insensitive diagnostic test. However, if true, the high negative predictive value of CT colonography may ultimately favour its use as a future screening test.

2.4.1.5 Use of optical colonoscopy as the reference standard

Most comparisons have used colonoscopy rather than barium enema (in many ways the natural comparison) and employed expert and highly motivated colonoscopists. For example all three multi-centre trials described above achieved cecal intubation rates of 99%. These figures are not comparable with those obtained from national audits of routine clinical practice, where completion rates averaging 57% are encountered instead [77]. Unlike the endoscopists, CT colonographers were relatively inexperienced in many trials since it is a new and emerging technology, so that any performance comparison is likely biased in favour of the reference. Indeed, by
definition in such comparisons the new test can never be superior to the reference standard unless the latter is modified by segmental unblinding. The onus is on the radiological community to provide thorough and effective training. It should be stressed that CT colonography and optical colonoscopy should not necessarily be viewed as competing investigations, rather complimentary, each with its own role in patient management (for example virtual colonoscopy used for polyp detection and optical colonoscopy for their subsequent, preferably same day, removal).

2.4.2 Extra-colonic findings

Unlike any other colonic investigations, CT colonography visualises extra-colonic organs, thereby permitting detection of potentially significant pathologies beyond the colon. Moreover, if a colonic cancer is found, it may be staged accurately [78]. A recent systematic review [48] of 17 studies (3 of which were performed in asymptomatic populations) and 3488 patients, reported a high incidence of incidental extra-colonic lesions (58%) discovered during virtual colonoscopy and, of these, a moderately high proportion required further investigation (14%). Ultimately, the incidence of clinically important pathology was relatively low, with just 4% having either extra-colonic cancer or an aortic aneurysm.

Clearly there is a significant danger that ultimately benign yet initially indeterminate extra-colonic findings may trigger unnecessary, expensive and potentially harmful diagnostic workup, averaging out at approximately 28$ to 34$ per patient screened at 2000 and 2003 rates respectively [79;80]. It is likely that CT colonography in asymptomatic patients will be performed using low dose techniques without intravenous contrast, increasing the difficulty when assessing the solid organs outside
the colon. Indeed, whether extra-colonic organs should be scrutinised at all in the context of a screening programme for colorectal cancer is the subject of intense debate. In any event, it is critical that unimportant findings are declared as such.

Recently a CT colonography working group has provided guidance for reporting extra-colonic findings [70] grading the importance of extra-colonic abnormality in order to guide the referring clinician about the need for follow up (Table 2). It is important to emphasise that all but three studies contributing to the above systematic review [48] investigated symptomatic populations (where extra-colonic pathology is the cause of symptoms believed to originate from the colon in a proportion of referrals) and the incidence in screening populations will inevitably be lower. Importantly, nearly half the incidental cancers found (most commonly renal and ovarian tumors) were discovered at an early stage (N0M0), when they are likely more amenable to treatment, indicating a further potential benefit of virtual colonoscopy.
Table 2. Classification of Extra-Colonic Findings

<table>
<thead>
<tr>
<th>Classification</th>
<th>Meaning</th>
<th>Some examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Limited Exam</td>
<td>Exam compromised by</td>
<td>Ultra low dose scans</td>
</tr>
<tr>
<td></td>
<td>artifact</td>
<td>Respiratory artifact</td>
</tr>
<tr>
<td>Normal Exam</td>
<td>No abnormalities found</td>
<td></td>
</tr>
<tr>
<td>Clinically insignificant</td>
<td>No work up indicated</td>
<td>Simple renal or hepatic cysts</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Anatomic variants</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Bone hemangioma</td>
</tr>
<tr>
<td>Likely insignificant</td>
<td>Work up may be indicated</td>
<td>High density renal cyst</td>
</tr>
<tr>
<td>Clinically significant</td>
<td>Communicate to referring clinician</td>
<td>Complex renal cysts</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Aortic aneurysm</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Possible cancer</td>
</tr>
</tbody>
</table>

The ultimate role, if any, for CT colonography in this scenario will require extensive modeling, with the models populated utilising data obtained from clinical trials and framed to summarise the health effects and costs of alternative approaches to detection of significant colonic neoplasia in patients of different ages, prior risks and preferences.
2.5 SUMMARY

Since its inception in 1994, CT colonography has developed into an increasingly promising and attractive alternative for the investigation and detection of colorectal neoplasia. Nevertheless, CT colonography has been subject to considerable scrutiny, not least due to concerns about variability in reader performance and the generalisability of CT colonography away from academic centres and into routine clinical practice.

There are several barriers to widespread acceptance of CT colonography, including the need for evidence based information regarding the level of training and experience required for acceptable performance levels (currently based on anecdotal expert opinion); and also to demonstrate that CT colonography is an accurate and reliable method for assessment of polyp diameter, given that this variable is frequently the sole arbiter for subsequent polyp management (Table 1). This thesis aims to tackle these issues, by defining the performance of CT colonography interpreted by readers with differing levels of experience and also investigating interventions that might help to improve performance and agreement between readers.
CHAPTER 3

CT COLONOGRAPHY PRACTICE IN THE UNITED KINGDOM

3.1 CT Colonography Practice in the UK: A National Survey

3.1.1 INTRODUCTION

Local implementation of CT colonography across the UK and Europe has generally been uncontrolled with few precedents to help guide successful service provision. Advice is now emerging from those who have successfully implemented the technique [69;81] together with guidance from professional bodies such as the CT colonography working group, established for the 2003 International Virtual Colonoscopy Symposium [70]. Ultimately an individual centres' priorities, resources, patient population and infrastructure will guide the local implementation strategy.

In the United Kingdom, there are several potential barriers to implementation. Critics cite the lack of randomised evidence supporting its use either for screening or symptomatic patients. CT scanner capacity is limited in the United Kingdom and, furthermore, it is generally accepted that CT colonography is both difficult to interpret and time-consuming. The precise technique employed also varies widely. However, it has attracted a great deal of professional and public attention with the result that there is considerable pressure on departments to provide it on a day-to-day basis.
While this thesis primarily aims to investigate reader performance, it is important to first determine the availability of CT colonography in the UK NHS, the technical parameters used, the number of examinations performed, by whom, and their experience.

3.1.2 MATERIALS AND METHODS

The Royal College of Radiologists database of UK departments of radiology was used to identify 236 departments of which 216 offered an adult gastrointestinal radiological service. A short questionnaire was addressed to the clinical director (or lead gastrointestinal radiologist if known) of these 216 departments and mailed by the principal investigator of this thesis in February 2003 along with a stamped, addressed envelope for reply. The questionnaire was designed so that it could be completed rapidly and without free text. It was organised into three broad sections: scanner hardware and software and general provision of a CT colonography service; barriers to implementation; experience and technique used (Appendix 1).

138 replies were received; a response rate of 64%, which was judged to be adequate using definitions of acceptable response rates derived from the social-sciences literature [82]. The responses were collated by the principal investigator of this thesis and frequencies determined in order to determine the availability of CT colonography across the UK, barriers to implementation, experience with the technique, and the technical parameters used.
3.1.3 RESULTS

Of the 138 departments who replied, 137 had a CT scanner, of which 74 (54%) were multi-detector row, and 81 (59%) had software specifically designed for CT colonography interpretation. Fifty (36%) offered CT colonography in day-to-day clinical practice. Of the 87 (64%) who did not, scanner capacity limited by competing clinical demands was the most frequent reason cited for non-implementation (68 of 87; 78%). Lack of radiological training or expertise in the technique was cited by 50 (57%) and financial restrictions by 38 (44%).

Of the 50 departments offering CT colonography as part of their day-to-day clinical practice, 19 (38%) performed approximately one study per month, with 25 (50%) performing one per week, 2 (4%) performing one per day, and 4 (8%) performing more than one study per day. Over the last five years, 14 (28%) departments had performed up to 20 studies, 14 (28%) up to 50, 9 (18%) up to 100, 7 (14%) up to 300, and 6 (12%) had performed more than 300. Indications for CT colonography in those departments offering a service included incomplete colonoscopy in 39 (78%), failed barium enema in 36 (72%), and as an alternative to barium enema in frail or immobile patients in 37 (74%).

Full bowel preparation was routinely administered prior to examination by 46 of the 50 (92%) departments offering CT colonography in day-to-day practice. Dual positioning was a routine manoeuvre in 45 (90%) and the colon was insufflated with air in 42
(84%) and carbon dioxide in 8 (16%). Intravenous contrast was routinely administered by 30 (60%).

CT colonography was performed by radiologists with a declared subspecialty interest in gastrointestinal radiology in 32 (64%) centres; by radiologists with a subspecialty interest in CT scanning or a general radiologist whose remit included CT scanning in 13 (26%); and by both these groups of radiologists in 5 (10%). In 18 (36%) centres radiographers and/or nurses performed the study, with subsequent radiologist review.

3.1.4 DISCUSSION

There are several potential barriers to the widespread adoption of CT colonography in the United Kingdom. In the United States, the technique is extensively promoted as an effective screening tool for colorectal polyps whereas in the United Kingdom health care is primarily a national government responsibility and their agencies require randomised evidence of efficacy before implementation. In any event, limited CT scanner capacity effectively prevents CT screening in the UK at the time of writing.

However, symptomatic colorectal cancer is a UK Government priority and CT colonography could have a prominent role for rapid and effective diagnosis. Both colonoscopy and barium enema are widely used for this purpose but at the time of writing there are no randomised comparisons to suggest that CT colonography is more sensitive or cost-effective than either. Instead, most studies of CT colonography have been comparatively small within-subject comparisons that have emerged from relatively few centres. The result is a lack of level one evidence [83] supporting
general implementation of CT colonography. Despite this, the technique has attracted considerable medical attention and this survey confirms that CT colonography is widely available in day-to-day practice across the UK; over one-third of departments offered a service. It was also very clear that further implementation would follow if scanner capacity were improved.

As described in the preceding chapter, the best available evidence supports CT colonography for failed whole-colon examinations [58-60] and 78% of active departments used it for this indication. There is a widely held belief that CT colonography is more acceptable and effective than established alternatives when examining frail, elderly patients. However, within-subject comparisons specifically aimed at this group have only evaluated standard CT techniques [61;84;85] and it could easily be argued that the bowel preparation required for colonography is a positive disadvantage in these patients. Despite this lack of direct evidence, 74% of departments used a colonographic technique in frail or immobile patients as an alternative to more established whole-colon techniques.

There is a broad consensus from expert centres regarding the optimal technical parameters necessary for high quality studies and these seem to have been generally adopted in the UK. For example, nearly all departments employed dual positioning. Furthermore, Government initiatives have sited multi-detector row CT scanners in many UK departments. The majority of UK departments utilised intravenous contrast, a procedure that is relatively unusual in the literature and which probably reflects the symptomatic nature of UK practice. For example, intravenous contrast both better images extra-colonic organs (which may actually be the origin of symptoms in this
group), and stages any primary colorectal cancer [45]. Fifty-nine percent of departments had the software necessary to render 3-D endo-luminal views, which is a considerable increase on the 25% suggested by an earlier questionnaire study from 2001 [86] and which might be explained, at least in part, by the UK Government New Opportunity Fund.

While there is broad consensus regarding the technical parameters used, much less is known about the experience required for optimal reader performance. While it is generally accepted that CT colonography is a difficult examination to interpret, and has a steep learning curve, the slope of this curve is unknown. This has not dissuaded UK radiologists from adopting the technique. Performance is related to experience and the present study reveals significant variation in the number of studies performed by those departments offering a service; more than one-third of these performed less than one study a month. Only six departments performed more than 300 studies a year, a figure that is far smaller than the average number of barium enemas performed.

It is increasingly well recognised that sub-specialist radiologists perform better than their generalist counterparts [87]. Although it could be argued that radiologists whose interest is CT scanning are most suited to CT colonography, it seems likely that radiologists with a subspecialty interest in luminal gastrointestinal radiology are best placed to exploit their specific knowledge of colonic anatomy and disease when interpreting the examination. Supporting this, we found gastrointestinal radiologists interpreted most studies.
The questionnaire nature of our study is a potential weakness. However, our response rate was above the 60% required for a 'good' response [82]; indeed 50% is deemed 'adequate' [82].

In summary, this survey demonstrates that CT colonography is widely available in the UK, with approximately one-third of departments offering a service. Experience and throughput varies considerably. Limited CT scanner capacity is the major barrier to further dissemination.
3.2 Potentially serious adverse events associated with CT colonography performed in symptomatic patients: A national survey of the UK

3.2.1 INTRODUCTION

CT colonography has long been regarded as a relatively non-invasive and safe investigation with no reported deaths and few significant adverse events. In comparison to colonoscopy, cardiovascular effects are significantly less and largely related to use of spasmolytic [24]. This is because no sedation is administered during CT colonography. Indeed, many CT colonography researchers have claimed that the procedure is entirely free from serious adverse events.

With a superior safety profile compared to optical colonoscopy, it has been claimed that CT colonography would be a more suitable test for implementation in a screening program [88]. For example, in preparation for a national screening program in the United Kingdom, it has been estimated that 12 patients per annum would die as a consequence of colonoscopy-related adverse events [89]. Moreover there is a growing awareness that screening programmes should not necessarily be restricted to younger participants (fifty percent of colorectal cancer is diagnosed in patients older than 75 years) and it is the elderly who are most at risk from colonoscopy related adverse-events (notably due to sedation). With this in mind, the imminent UK scheme will have an 'opt-in' clause for older patients, no matter how old.
That said, the radiation dose associated with CT colonography needs to be considered and will vary according to the technical parameters used, and the patient population studied; younger patients are more at risk from radiation-induced tumors. For symptomatic patients, higher radiation doses are justified on a risk/benefit basis and are necessary if intravenous contrast is administered. For a screening population the high contrast between colonic mucosa and intra-luminal gas means that radiation dose can be significantly lower than either a conventional CT or barium enema, without significant detriment to image quality [90-92]. Undoubtedly, technological improvements such as dose modulation techniques will serve to lower radiation dose further [92].

Despite this favourable experience to date, there have been few previous attempts to establish the frequency of adverse events associated with CT colonography. For example, a single center reported minor complications in 12 of 343 (3.5%) asymptomatic patients [88]. Although initially believed free from serious adverse events because of its relatively non-invasive nature, isolated case reports have emerged that suggest CT colonography is also associated with potentially serious adverse events, in particular luminal perforation [68;93].

Sosna and co-workers obtained data from three centers, and suggested that CT colonography was associated with a perforation rate of 1 in 2393 (0.04%) examinations [94]. These authors extended their work to include 21 centers across 7 countries, identified by international visibility, and found 9 perforations in 24,365 examinations (perforation rate 1 in 2,707, 0.04%): five of these required laparotomy
However, surveys directed towards specialist units are inevitably subject to considerable selection and spectrum bias and while they reflect the complication rate in these centers, they do not necessarily reflect the true incidence of potentially serious adverse events. As a case in point, a survey of 3196 colonoscopies performed as part of a research study to investigate procedural success and complications achieved total intubation in 97.2% with no perforations or death [96]. In contrast, a UK audit of day-to-day clinical practice in 68 predominantly non-specialist units [77] found total intubation rates of approximately 50% and 12 perforations in 9223 examinations (1 in 769 examinations, 0.13%). In fact, colonoscopy was implicated as a possible factor in six deaths (0.07%) occurring within 30 days of the procedure [77].

A fundamental theme of this thesis is to examine how CT colonography performs away from specialist units, in routine day-to-day clinical practice, and including assessment of patient risk. Also, this thesis has established that in the UK, CT colonography is generally used for examining symptomatic patients. With these factors in mind, this survey aims to retrospectively determine the incidence of potentially serious adverse events associated with CT colonography performed in patients with symptoms of colorectal cancer.

3.2.2 MATERIALS AND METHODS

The fifty UK centres responding to the survey detailed earlier in this chapter (Chapter section 3.1) who offered a CT colonography service in day-to-day clinical practice were invited to participate in this follow-up survey. The principal investigator of this
thesis and a second colleague (both had extensive experience in performing (200 and 100 cases respectively) and interpreting CT colonography examinations (600 and 150 endoscopically validated cases respectively) attempted to contact by telephone the lead gastrointestinal radiologist from each of these 50 centres.

The lead gastrointestinal radiologist from each center was questioned in February 2005 by one of the two radiologists (mentioned above – the principal investigator contacted 30 and his colleague, 20 centres). We asked a series of six questions, read from a study sheet (Appendix 2). In particular, we asked how many serious adverse events (including number of deaths) they had experienced in their department, if any, and the total number of examinations performed at the time of the present survey so that we had both the numerator and denominator in order to determine the potentially serious complication rate associated with CT colonography. We also asked additional questions related to those aspects of technique that could possibly influence perforation rates, namely whether a rectal balloon was inflated in situ and whether an automated insufflation device was used (Appendix 2). We were particularly interested in luminal perforation (rectal and colonic) but asked for details of any potentially serious adverse event. We defined a potentially serious adverse event as either luminal perforation where gas was observed beyond the bowel wall or any other complication that required a period of observation before the patient was thought to be able to leave the department safely.

If a serious adverse event had been identified, the lead gastrointestinal radiologist from the respective centre was contacted again by the principal investigator of this thesis and further questions related to the adverse event were asked in detail. In
particular, questions related to diagnosis of the adverse event in question, its clinical severity, treatment, and ultimate clinical outcome were asked. The details of the anatomical distribution of gas in those patients with luminal perforation (specifically whether this was retroperitoneal, intraperitoneal, or both) were also sought, so that we could determine the theoretical risk of fecal peritonitis.

3.2.2.1 Statistical analysis

Responses were collated, tabulated, and raw frequencies determined. Fisher's Exact test was used to determine differences in the proportion of perforations encountered between centers using an inflated rectal balloon catheter and those who did not, and also between research (3) and non research (47) centers. Research centres were defined as those which had published peer-reviewed indexed articles relating to CT colonography and were identified by performing a Medline search of CT colonography articles published over the last 10 years. All other centres were categorised as non-research centres. Statistical significance was assigned to a probability value of 5% or less and analysis performed using Stats Direct Stats Direct version 2.4.4. (Stats Direct Ltd., Sale, Cheshire, UK).
3.2.3 RESULTS

3.2.3.1 Contact, examinations performed

The lead gastrointestinal radiologist at all 50 centres (who responded to the initial survey and offered CT colonography in day to day practice - see above) was contacted successfully, a response rate of 100% for the present survey. Of these, 47 centres still offered a CT colonography service in routine clinical practice; the remaining three no longer offered a service but could not recall any serious adverse events related to CT colonography.

In total, 17,067 CT colonography examinations had been performed by the 50 centres (mean number per center 359, range 10 to 3000); 36 (72%) had performed a total of 100 examinations or more. At the time of our telephone survey, on average, 5 (10%) centres performed more than one examination per day, 21 (42%) centres performed one examination per day, 14 (28%) performed one per week, 7 (14%) performed one per month and 3 (6%) no longer performed CT colonography.

3.2.3.2 Potentially serious adverse events

No deaths were reported. Thirteen (1 in 1313 patients, 0.08%) patients had suffered potentially serious adverse events believed to be related to CT colonography. Of these, there were three self-limiting vasovagal episodes, and one attack of cardiac angina successfully treated with sublingual glyceryl trinitrate spray. All patients were discharged home after a period of observation, apparently well. There were nine luminal perforations giving a perforation rate of 1 in 1896 patients, 0.05%. The nine
perforations came from six centres. All six had performed a total of 100 or more examinations and collectively contributed 6500 (38%) to the total number of CT colonography examinations performed. The individual circumstances for the nine perforation cases are summarized in Table 3. Five (56%) of the nine perforations had an attributable cause as follows:

- A radiologist resident, believing he was examining the distal limb of a loop colostomy (surgical details were not communicated on the request form), actually inflated a rectal stump – perforation occurred at the suture line at its apex (Table 3, case 1).
- A rectal catheter was traumatically inserted through apparently normal rectal wall by a radiographic technician (Table 3, case 6).
- A transverse colon perforation occurred in a patient with previously undiagnosed ulcerative colitis (Table 3, case 7).
- An obstructing sigmoid carcinoma was believed to be the cause underlying a sigmoid colon perforation (Table 3, case 8).
- Perforation was discovered on plain radiography in a 72-year old woman (Table 3, case 9) with multiple co-morbidities including known diverticular disease and rheumatoid arthritis treated by non-steroidal anti-inflammatory drugs (NSAIDs). She attended feeling unwell and abdominal radiography performed before CT colonography revealed intraperitoneal gas. Perforation was deemed to have been precipitated by her bowel preparation. This patient did not therefore undergo CT colonography.
<table>
<thead>
<tr>
<th>Case Center</th>
<th>Gas Used</th>
<th>Gas Used Insufflation</th>
<th>Device Used</th>
<th>Device Used Insufflation</th>
<th>Associated Symptoms</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Abdominal pain</td>
</tr>
<tr>
<td>2</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Abdominal pain</td>
</tr>
<tr>
<td>3</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Abdominal pain</td>
</tr>
<tr>
<td>4</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Abdominal pain</td>
</tr>
<tr>
<td>5</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Abdominal pain</td>
</tr>
<tr>
<td>6</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Abdominal pain</td>
</tr>
<tr>
<td>7</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Abdominal pain</td>
</tr>
<tr>
<td>8</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Abdominal pain</td>
</tr>
<tr>
<td>9</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Abdominal pain</td>
</tr>
</tbody>
</table>

Table 3: Details of nine patients who suffered luminal perforation related to CT colonography.
Eight perforations were discovered during or after the CT procedure (the exception being Table 3, case 9). Four patients (44%) with perforation were entirely asymptomatic (Table 3 cases 2,3,4,5). The finding of extra-luminal gas in these four patients was discovered incidentally by the reporting radiologist between 6 hours and four days following the procedure. All of these four patients had returned home, feeling well and were subsequently contacted and managed at home conservatively. Thus the symptomatic perforation rate was 1 in 3413 patients, 0.03%. Because of the lack of an attributable cause in these four, it was difficult to establish the exact site of perforation. Extraluminal gas was located intraperitoneally surrounding the cecum in two cases, the ascending/transverse colon in one case and retroperitoneally in the descending colon in one case (Fig 4).

A single patient whose sigmoid carcinoma had been diagnosed on CT colonography (Table 2, case 8) underwent laparotomy because of peritonitis, fears of tumour perforation, and the knowledge that surgery would be required to treat the underlying tumour in any event (Fig 5). Eight (89%) patients with perforation were thus treated conservatively either as in- or out-patients (Table 2). To the respondents knowledge, all patients with perforation were alive and well at the time of our survey.
Figure 4.

A; axial and, B; coronal. CT colonography images (supine acquisition) in a female patient with retroperitoneal perforation (table 3, case 3) – Black arrows outline the extraluminal retroperitoneal air surrounding the descending colon in Fig 4 A and B. The patient was entirely asymptomatic and required no treatment. No obvious cause for the perforation was attributed.

Figure 4A
Figure 4B
Figure 5.

A; sagittal and, B; coronal. CT colonography images (prone acquisition) in a female patient with intra- and retroperitoneal perforation (table 3, case 8) – White arrows outline retroperitoneal air seen surrounding the rectum and extending around both the ascending and descending colon in Fig. 5A and around the rectum only in Fig 5B. This was attributed to a sigmoid carcinoma (not shown) and the patient underwent laparotomy, at which time the primary tumour was removed.

Figure 5A
Figure 5B
3.2.3.3 Use of rectal balloon catheter

Twenty-nine (58%) centres never used an inflated balloon catheter, 7 (14%) centres occasionally used one (for on average 14% of their examinations when anal incontinence was encountered; range 1 to 50%), and 14 (28%) centres always used one. Of these 14 centres, the balloon catheter used was a rectal retention catheter in 9 and an inflated Foley catheter in 5. Overall, 9,378 CT colonography examinations were performed using an inflated balloon in the rectum (in which there were 6 perforations) and 7,689 without (in which there were 2 perforations). There was no significant difference in the proportion of perforations associated with and without rectal balloon inflation (P=0.3). Six centres overall (12%) utilized an automated insufflation device. Two perforations were associated with CT colonography performed using automated insufflation (table 3, cases 4 and 5).

3.2.3.4 Research, non-research centres

Three of the 50 centres (6%) had published peer-reviewed indexed articles relating to CT colonography and one of these had experienced two perforations (Table 3, center 3). These three centres contributed 4,350 patients to the total of 17,067. There was no significant difference in the proportion of perforations originating from research and non-research centres (p = 0.82).
3.2.4 DISCUSSION

This study found that the incidence of potentially serious adverse events in patients with symptoms of colorectal cancer were uncommon, occurring in 0.08% of patients examined. Of these, the majority were luminal perforations, occurring in 0.05% of studies performed, of which five (0.03%) were symptomatic. This compares favorably to routine day-to-day colonoscopy practice performed in the same or similar hospitals, where the symptomatic perforation rate appears to be over four times more frequent (i.e. 0.13%) [77]. A UK survey of day-to-day double contrast barium enema practice found 30 perforations in 738,216 studies (1 per 24,607, 0.004%) [39] but the consequences of barium peritonitis are likely more devastating than perforations occurring during CT colonography; 10% of such patients died [39]. The symptomatic perforation rate associated with CT colonography was 0.03% - four patients were entirely asymptomatic. It should be borne in mind that CT colonography is exquisitely sensitive for extra-luminal gas. Colonoscopy cannot detect extra-luminal gas and so cannot detect asymptomatic perforations, and barium enema will inevitably also be less sensitive than CT. As a result, colonoscopy, and to a lesser extent barium enema, will underestimate the true perforation rate as asymptomatic perforations will generally be missed. Moreover half of the colonoscopy related perforations in the audit discussed above were not recognized at the time of colonoscopy, perhaps as symptoms were masked by intravenously administered sedation. For this reason, the most relevant comparison between perforation rates for CT versus colonoscopy may actually be 0.03% versus 0.13%; i.e. symptomatic perforation seems 4.3 times less
frequent for CT. It should be noted that our survey and the colonoscopic comparator [77] was performed in the same and similar hospitals.

With respect to perforation during double contrast barium enema, it has been estimated that rectal perforation is increased by a factor of 2.5 when retention balloon catheters are used [97], possibly due to their diameter (20mm), radial force when expanded, and relatively stiff construction: In our survey one such catheter was forcibly inserted through the rectal wall and a balloon retention catheter was employed in 56% of the perforations we encountered. It has been suggested that if they are to be used, then these wide bore, stiff plastic catheters should be inserted carefully by an experienced radiologist using fluoroscopic guidance and only then after rectal examination [97]. Given that they do not benefit CT colonography significantly [34], their routine use should probably be discouraged. Although we were unable to demonstrate a significant association between balloon inflation and perforation, it should be noted that the event rate (i.e. perforation) was low and we likely lacked statistical power to confidently exclude an association.

Some workers recommend automated insufflation devices to improve colonic distension and reduce post-procedural abdominal pain since these machines utilise carbon dioxide rather than room air [42;43]. These devices also allow rectal pressure to be monitored and cease insufflation if rectal pressure rises above 25mmHg. We encountered two perforations using these devices, both of which were cecal, the colonic segment most prone to perforation. Further research quantifying the intraluminal pressures generated by automated insufflation in different colonic segments should be undertaken.
A number of the perforations that we encountered could potentially have been avoided. For example, one perforation occurred because the surgical anatomy was not fully appreciated, and this stresses the importance of having relevant information available. A relatively inexperienced technician forcibly perforated the rectum, emphasizing that rectal insertion is potentially dangerous, and practitioners need to be aware of this and appropriately supervised where necessary. Indeed, intuitively one might have anticipated an increased number of perforations from centers with limited experience of CT colonography. However, all the perforation cases occurred in six centers which collectively had performed 38% of the total number of examinations. One possible explanation is that centers performing large numbers of CT colonography are more likely to delegate the task of colonic distension to less experienced members of the team who may be less careful when inserting the rectal catheter, inflating the balloon or distending the colon. Indeed, of the five perforations discovered during or after CT colonography, two were performed by radiographic technicians and one by a radiology resident. Perforation was due to underlying pathology in two patients, ulcerative colitis and a sigmoid carcinoma respectively. It would be difficult to avoid these complications in advance because the underlying pathology was not known but it does emphasise that the initial acquisition (whether prone or supine) should be scrutinised for pathology and care taken when discovered. There have been previous case-reports of perforation associated with inflammatory bowel disease [68] and this, coupled with the fact that many such patients are young, and the considerable difficulties distinguishing dysplasia from regenerative epithelium, suggest that CT colonography is best avoided in this clinical scenario. Overall, of the
four cases of symptomatic perforation discovered during CT colonography (Table 2),
three (75%) were potentially avoidable.

Although we have attempted to circumvent selection bias, there are limitations to our
study. Most notably, such retrospective surveys, valuable as they are, are based on
self-reported practice and are therefore subject to recall bias. For this reason we
chose to focus on serious complications associated with CT colonography, not only
because it is most important that these are reported in the peer-reviewed literature
but because these are also more likely to be recalled accurately than minor
complications. For example, we found that the radiologists reporting potentially
serious adverse events were very clear about the details of each individual case. It is
also important to emphasise that this survey relates to a symptomatic population,
where the proportion of elderly and frail patients is over-represented – cardiac angina,
as encountered by us, is unlikely in younger patients for example. It is very possible
that potentially serious adverse events would be reduced significantly in a younger,
asymptomatic screening population, where the prevalence of abnormality is less. A
prospective design would be able to control for recall bias and additionally for
potential confounders such as the size of catheter and balloon used, the experience
of the individual performing the insufflation, and the use of spasmolytics.

In summary, we found potentially serious adverse events associated with CT
colonography in 0.08% of patients when used to investigate patients with symptoms
of colorectal cancer. However, symptomatic luminal perforation occurred over four
times times less frequently than equivalent rates published for colonoscopy. Some
complications of CT colonography are potentially avoidable. When they did occur,
potentially serious adverse events were managed conservatively with success in the majority of cases.
CHAPTER 4
READER PERFORMANCE

4.1 Effect of Directed Training on Reader Performance for CT Colonography: Multi-Center Study

4.1.1 INTRODUCTION

In Chapter section 3.1, this thesis discussed the lack of guidance regarding local implementation of CT colonography, but also found that in spite of this, the technique was widely available across the UK. There is a growing body of evidence to support the use of CT colonography in a few specialist centres, where the technique and interpretation methods have been optimized, and experienced readers interpret the examinations. In contrast, however, there is a lack of evidence examining the generalisability of CT colonography when interpreted by readers with differing experience in routine clinical practice.

There have been attempts to investigate interobserver variability for interpretative performance at CT colonography and the most notable of these was a retrospective multicentre study by the Mayo group, published in the same journal as the study described in Section 2.4.1.1 [74]. This retrospective study compared the interpretative performance of eighteen radiologists with differing prior CT colonography experience,
and asked each to review a dataset of approximately 60 cases, which most read within three days.

Diagnostic accuracy was assessed by using the non-parametric area under the receiver operating curve (AUC) and showed wide variability in individual reader performance. The average AUC for identification of patients with lesions of 10mm or larger was 0.8 (with a wide range from 0.58 to 0.99). This study showed better performance amongst more experienced readers; the average of the AUCs for the most experienced readers (prior interpretation experience up to 50 CT colonography cases) was 0.82 (CI, 0.75-0.88) compared to 0.77 (CI, 0.7-0.85) for the least experienced readers (prior interpretation experience <10 cases).

Indeed this study also suggested a learning curve, with improved performance related to increasing experience, a finding indicated by previous studies. In 2001, Spinzi et al showed that sensitivity of CT colonography for polyp detection increased from 32% to 92% when comparing the first and second halves of the study [13]. In 2004, Taylor et al [14] also showed a learning curve in a study assessing the interpretative performance of three inexperienced radiologists who each read 100 colonoscopically validated cases and received limited individual training in the form of feedback after the first fifty. While one of the readers significantly improved, one reader got worse and one reader was unchanged, indicating the learning curve can be unpredictable even with structured training.
These findings have important consequences for training requirements. While it is clear that training will increase interpretation experience and this in turn can improve performance, there is a lack of evidence regarding what constitutes adequate training for independent reporting. Furthermore, there are no agreed standards for acceptable reader performance.

A consensus statement from the working group of experts in the USA suggested that ideally CT colonography training should be carried out in a 'hands-on' workshop environment, reviewing up to fifty endoscopically validated examinations over two days [16]. However this figure is arbitrary and available evidence (as described above) suggests that fifty cases may not be sufficient for a significant proportion of attendees. There is also some evidence to suggest that it is the quality of training that is most important rather than simply the number of cases reviewed [2].

The European Society of Gastrointestinal and Abdominal Radiology (ESGAR) has acknowledged these concerns and is interested in developing evidence-based guidelines for training and possible accreditation in CT colonography interpretation. With this in mind, a major aim of this thesis is to define the interpretative performance of radiologists experienced in CT colonography and to compare this with novice observers who had undergone directed training, using colonoscopy as the reference standard.
4.1.2 MATERIALS AND METHODS

4.1.2.1 Dataset composition and accrual

We collated a dataset of normal and abnormal CT colonography studies from seven participating centres, the prevalence and morphology of neoplasia of which was modeled to simulate that expected in patients testing positive to fecal occult blood (FOBT) (10-12), i.e: cancer (10%), large polyps (30%), medium polyps (10%), normal colorectum (50%). This procedure aimed to create a mix of abnormal/normal examinations in order to investigate sensitivity for different classes of neoplasia and specificity between observer groups. It also ensured the dataset had clinical relevance. Each centre was asked to submit 10 examinations, aged 50 to 69 years, mimicking accrual in FOBT trials [98-100]; one patient with cancer, three with large polyps, one with a medium polyp, and five normal patients, i.e. there were four possible diagnostic categories – cancer, large polyp, medium polyp, normal. In line with FOBT trials, a large polyp was defined as one measuring 10 mm or more and a medium polyp as one measuring less than this (6 to 9mm for the purposes of this study).

In order to reflect normal variation in data quality, patients from each centre were accrued in a strictly chronologically consecutive fashion. That is to say, consecutive patients were allocated to an appropriate diagnostic category until all four categories were full. Thus, the first patient with cancer encountered completed recruitment to that category whereas three consecutive patients with large polyps were necessary to
complete accrual to that category. Patients with multiple lesions were classified by the largest – the 'index' lesion. So that studies reflected accurately the natural and inevitable technical variation found in day-to-day practice, centres were obliged to submit all eligible examinations unless it was deemed non-diagnostic (defined as any study where the local principal investigator would normally recommend a repeat colonography or another test because of insurmountable technical problems, e.g. segmental collapse or retained fluid). In all patients CT findings were defined by subsequent same-day colonoscopy performed by experienced practitioners. Polyp size was defined by the colonoscopic measurement, estimated using adjacent biopsy forceps.

Participating centres were chosen because they had active CT colonography research programs at the time the study protocol was written, and so could contribute patients. We also stated that as long as examinations were chronologically consecutive, centres could submit retrospective examinations as long as technical stipulations for CT colonography were satisfied. Five centres used retrospective data alone and two centres used prospective data alone. Ethical permission for data sharing was covered by the local stipulations at each centre. Four centres using retrospective data did not require specific approval for this study because data had been collected as part of a local study with ethical approval for additional analyses and data-sharing. The fifth centre using retrospective data was asked to gain verbal consent via telephone from the patients selected. The two centres using prospective data had ethical permission for additional analyses and data-sharing for on-going studies as long as patient identifying information was removed before data sharing. We applied the latter stipulation to all data collected for the study.
Imaging was performed using full bowel purgation; prone and supine acquisition; multi-detector row machines; collimation no greater than 2.5mm. The gas insufflated and spasmolytic use was left to local discretion. Low-dose protocols were permissible but intravenous contrast was not (since this is unlikely in a screening program). Faecal tagging was not permitted since this was not common practice at the time of data accrual (May to November 2003). Studies were archived to compact disk and transferred to the trial office with technical and diagnostic category information for each examination.

Of the seven centres submitting data, three provided data from symptomatic patients, one from asymptomatic patients, and three from both. Four centres submitted a full dataset of 10 patients; one of these examinations could not be opened at the trial office. The three remaining centres submitted five, four and three examinations respectively due to difficulties satisfying protocol requirements, notably those related to age. Thus there were 51 patients of whom 27 (53%) were normal and 24 (47%) had an index lesion as follows: 8 cancer, 12 large polyp, 4 medium polyp. Seven (29%) of these patients had a second lesion; two cancer patients each had one additional large polyp; one cancer and four large polyp patients each had one additional medium polyp.

4.1.2.2 Observers
The dataset was interpreted by three groups of observers; experienced radiologists, trained radiologists, and trained radiographic technologists. Nine centres (including all
An experienced radiologist was defined as someone who had considerable practical and/or research experience of CT colonography prior to the study. Individual experience ranged from 325 examinations to 1200 examinations (median 750), with the number validated by colonoscopy ranging from 120 to 600 (median 200).

Each experienced radiologist identified a local radiologist and radiographic technologist who had interpreted 10 or fewer examinations prior to this study. We stipulated these radiologists were familiar with interpretation of standard abdomino-pelvic CT and the technologists were familiar with acquisition of abdomino-pevic CT data. These radiologists and technologists were then trained in CT colonography interpretation by the local experienced radiologist using normal and abnormal examinations that had been acquired locally and verified by subsequent colonoscopy; there was no attempt to administer the same training dataset to all participating centers since we wished to emulate existing training programs for conventional CT (where, in general, trainees learn using examinations accumulated locally). However, we did stipulate that 50 individual examinations should be interpreted, unaided initially and then followed by face-to-face discussion with the local trainer on a patient-by-patient basis, so as to closely mimic standard day-to-day training practice. Trainers and trainees used whatever was their preferred local reading platform, in line with their everyday practice. We also stipulated that training should occur over several separate sessions and over several weeks, again to reflect standard teaching practice.
4.1.2.3 Reading conditions and outcome measures

After training, an individualised test dataset of 40 examinations was prepared by the principal investigator of this thesis for each participating centre. This was sampled from the dataset of 51, excluded any examination submitted by the center in question, and was balanced in terms of the prevalence of abnormality. The order of the cases was randomized in order to mix abnormal and normal cases and all readers read the cases in the same order, excluding any cases from their own cente. All identifiers were removed. The experienced radiologist (nine in total), trained radiologist (nine) and trained technologist (ten) from each centre then interpreted this dataset over two days. The principal investigator of this thesis visited each center to supervise reading, which was conducted using individual laptops with 17inch screens and software that allowed a primary 2D analysis with 3D for problem solving (Voxar ColonScreen version 2.2, Barco, Edinburgh, UK). Observers were familiarised with the software when necessary and the supervisor was available at all times. Reading was performed in a quiet environment with ambient light. Observers were asked to read at their own pace with no requirement to finish within a pre-specified time. Observers had read the study protocol and knew that their own examinations, if any, had been excluded, but had no specific information about the composition of their individualised dataset.

Observers categorised each patient as either 'normal' or 'abnormal', using a datasheet. 'Abnormal' was further categorised as follows: 'cancer', 'large polyp' or 'medium polyp'. Large polyps were defined by a maximal 2D transverse diameter of
1cm or larger and medium polyps by 6mm to 9mm, using software calipers. Observers noted any polyp measuring 5mm or less but categorised such patients as 'normal', a procedure that allowed false-negatives due to measurement error to be distinguished from false-negatives due to perceptual error. Observers were unaware of other's responses. Prone and supine image co-ordinates and segmental location were recorded for each abnormality perceived (so that false-positive responses could be distinguished from true-positives in the same patient). Multiple responses were possible. There were six segments: rectum, sigmoid, descending, transverse, ascending, cecum, and observers were provided with an annotated diagram of segmental definitions. Observers were free to classify an examination as technically inadequate, although the protocol had taken steps to avoid non-diagnostic studies.

Datasheets were collated and responses compared with the known diagnostic category. The principal investigator of this thesis (who had experience of over 300 endoscopically verified examinations at the time) independently evaluated each examination to confirm the CT findings declared by the submitting centre and to confirm the CT co-ordinates of the abnormality, which was then used to determine whether observers' responses were true-positive or false positive. All but one of the endoscopically validated lesions could be identified but there was difficulty locating four flat adenomas, measuring 40mm, 30mm, 15mm and 12mm, one of which was only visible using standard abdominal CT windows (Hounsfield units level 40, range 400) (Figure 6). One flat adenoma (40mm) could not be identified despite good bowel preparation and distension, and the endoscopic data.
Figure 6.

62-year old woman with a 12mm sigmoid flat adenoma. Axial view. Images acquired using a four detector row scanner (100mA, 120 kV). This polyp was missed by all observers.

A; the adenoma (arrows) is barely visible using a standard CT colonography window (Hounsfield units level -150, range 1500, white arrows).
B; the adenoma is more clearly visible using standard abdominal CT windows (Hounsfield units level 40, range 400, white arrows).

4.1.2.4 Statistical analysis

Observer responses were compared with the known patient category and lesion co-
ordinates and true-positive, true-negative, false-positive and false-negative counts
determined. Individual and group performance was determined by calculating the
number and percentage of examinations in which the index lesion (and second lesion
in 7 patients) was correctly identified, and normal examinations that were correctly
categorised. False-positive classifications in normal patients and also in patients with
a known index lesion elsewhere were determined. Two measures were derived for
each reader, namely sensitivity for lesions and accuracy (defined as the overall
percentage of correct categorisations). For both measures, cases classified as
'inadequate' were excluded. Because observers largely read the same examinations (and observations were thus correlated to some extent), a bootstrap analysis was used to investigate differences between observer groups. 1999 samples were redrawn randomly from the original sample with replacement and analysis of each resultant dataset. The results of interest were calculated for each bootstrap sample and the distribution of values used to obtain a bootstrap confidence interval. A probability value was also calculated by considering how many of the values were further from zero than the actual value observed with the data. Statistical significance was attributed at a probability level of 5%. Analysis was performed using Stata 8.0 (StataCorp, College Station, Texas, USA).

4.1.3 RESULTS

The 28 observers read a total of 1084 individual examinations; 22 of 28 (79%) readers (including all nine experienced observers) read all 40 examinations assigned to them; 2 read 39, 1 read 37; 1 read 35 and 2 read 27 examinations each because of time constraints.

Overall, 68% classifications were correct (Table 4).
Table 4.

Relation between patient category and observer assessment for all observer groups combined.

<table>
<thead>
<tr>
<th>Patient category</th>
<th>Seen</th>
<th>Missed</th>
<th>False</th>
<th>Inadequate</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cancer</td>
<td>121 (79)</td>
<td>33 (21)</td>
<td>-</td>
<td>0</td>
<td>154</td>
</tr>
<tr>
<td>Large polyp</td>
<td>134 (47)</td>
<td>147 (53)</td>
<td>-</td>
<td>2</td>
<td>283</td>
</tr>
<tr>
<td>Medium polyp</td>
<td>33 (36)</td>
<td>59 (64)</td>
<td>-</td>
<td>0</td>
<td>92</td>
</tr>
<tr>
<td>Normal</td>
<td>448 (81)</td>
<td>-</td>
<td>73</td>
<td>34</td>
<td>555</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>736</td>
<td>239</td>
<td>73</td>
<td>36</td>
<td>1084</td>
</tr>
</tbody>
</table>

n = number
Correct patient classification declined in parallel with decreasing size of the index lesion: 79% cancer patients were detected, 47% large polyp patients, and 36% medium polyp patients (Table 4). Of the remaining 348 categorisations; 239 were patients in whom the index lesion had been missed; 73 were false positive; and 36 were judged technically inadequate. Of the 36 inadequate assessments, 23 (64%) related to one patient whose diagnostic category was normal. Of the 13 other technically inadequate assessments, 11 related to patients whose diagnostic category was normal. Overall, the false positive rate was 12% (21/176) for experienced observers, 12% (20/169) for radiologists and 14% (26/188) for technologists. One experienced observer and two technologists made no false positive categorisation. Six readers had false positive rates of 20% or more: one experienced observer, two radiologists and three technologists.

4.1.3.1 Observer performance (Table 5)

Experienced observers detected more lesions overall than did the other two groups; 66% overall versus 51% for radiologists and 47% for technologists. This observation was also the case when all subgroups of lesions were considered individually.
Table 5.

Summary of lesion detection rates according to observer group.

<table>
<thead>
<tr>
<th>Index lesion</th>
<th>All lesions</th>
<th>Cancer</th>
<th>Large polyps</th>
<th>Medium polyps</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Seen</td>
<td>Missed</td>
<td>Seen</td>
<td>Missed</td>
</tr>
<tr>
<td>Observer group</td>
<td>n (%)</td>
<td>n (%)</td>
<td>n (%)</td>
<td>n (%)</td>
</tr>
<tr>
<td>Experienced</td>
<td>116 (66)</td>
<td>60 (34)</td>
<td>47 (92)</td>
<td>4 (8)</td>
</tr>
<tr>
<td>Radiologist</td>
<td>85 (51)</td>
<td>82 (49)</td>
<td>34 (71)</td>
<td>14 (29)</td>
</tr>
<tr>
<td>Technologist</td>
<td>87 (47)</td>
<td>97 (53)</td>
<td>40 (73)</td>
<td>15 (27)</td>
</tr>
</tbody>
</table>
Subset analysis revealed that some polyps were clearly more difficult to detect than others, a phenomenon that applied across all observer groups. For example, of the 12 patients whose index was a large polyp, two of these polyps were missed by all 24 observers who read these two examinations. The large polyp in a further two of these patients was identified and categorised correctly by only one observer, an experienced reader. All of these four difficult-to-detect polyps were morphologically flat. Of the four patients whose index lesion was a medium polyp, only one observer (4%) saw the index polyp in one examination out of the 23 who were assigned it (Figure 7), and only two observers (10%) identified the index polyp in a second from the 21 who were assigned it. Thus there were two difficult-to-detect medium polyps. Overall, the six difficult-to-detect polyps (four large and two medium) had considerable influence on our results, decreasing accuracy for both observer groups and individuals.
Figure 7.

66-year old woman with a 8mm descending colon adenoma. Axial view. Images acquired using a four detector row scanner (100mA, 120 kV). This polyp was visible on the prone dataset only and was missed by all observers, excepting one experienced observer.

A; The polyp (arrow) is visible on the axial view using CT colonography windows; Hounsfield units level -150, range 1500.
B; 3D volume rendered endoluminal view.
4.1.3.2 Accuracy and sensitivity Table 6

Observer accuracy and sensitivity according to experience for all lesions and when the six difficult-to-detect polyps were excluded. Accuracy refers to the percentage of correct classifications for patients with and without lesions whereas sensitivity refers to detection of cancers and polyps only. (CI = confidence interval p = probability value)

<table>
<thead>
<tr>
<th>Observer group</th>
<th>Overall accuracy (%)</th>
<th>Excluding difficult cases (%)</th>
<th>Sensitivity (%)</th>
<th>Excluding difficult cases (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experienced</td>
<td>74.2</td>
<td>83.7</td>
<td>65.5</td>
<td>85.3</td>
</tr>
<tr>
<td>Radiologist</td>
<td>66.6</td>
<td>76.9</td>
<td>50.7</td>
<td>69.7</td>
</tr>
<tr>
<td>Technologist</td>
<td>63.2</td>
<td>72</td>
<td>47.3</td>
<td>63.5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Group comparisons</th>
<th>Difference: % (95% CI)</th>
<th>p</th>
<th>Difference: % (95% CI)</th>
<th>p</th>
<th>Difference: % (95% CI)</th>
<th>p</th>
<th>Difference: % (95% CI)</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experienced vs radiologist</td>
<td>7.6 (1.2 to 14.3)</td>
<td>0.017</td>
<td>6.8 (0.5 to 13.1)</td>
<td>0.035</td>
<td>14.9 (4.3 to 25.2)</td>
<td>0.007</td>
<td>15.6 (5.7 to 25.8)</td>
<td>0.004</td>
</tr>
<tr>
<td>Experienced vs technologist</td>
<td>11.0 (0.5 to 17.7)</td>
<td>0.003</td>
<td>11.7 (5.2 to 17.9)</td>
<td>0.001</td>
<td>18.2 (8.3 to 28.3)</td>
<td>0.002</td>
<td>21.9 (6.2 to 16.9)</td>
<td>0.001</td>
</tr>
<tr>
<td>Radiologist vs technologist</td>
<td>3.4 (-3.4 to 10.1)</td>
<td>0.33</td>
<td>4.9 (-1.8 to 11.7)</td>
<td>0.16</td>
<td>3.4 (-7.2 to 13.7)</td>
<td>0.52</td>
<td>6.1 (-0.8 to 10.7)</td>
<td>0.31</td>
</tr>
</tbody>
</table>
Regarding the bootstrap analysis, experienced observers were significantly more accurate and sensitive overall than the other two observer groups. This was the case for all analyses, irrespective of whether the six difficult to detect polyps were included or not. However, there was no significant difference in measures of accuracy or sensitivity when the radiologists were compared to the technologists in any of the comparisons made. Although these measures show that experienced readers performed best on average, there was considerable overlap between the groups when individual performance was considered (Figures 8 and 9); for example, one technologist and two radiologists exceeded the mean accuracy for the experienced group.
Figure 8.

Graph of observer accuracy overall, grouped by experience.

+= mean value: 74.2% for experienced observers, 66.6% for radiologists, 63.2% for technologists.
Figure 9.
Graph of observer accuracy grouped by experience excluding 6 "difficult-to-detect" lesions.
+ = mean value: 83.7% for experienced observers, 76.9% for radiologists, 72.0% for technologists.
4.1.3.3 Second lesions

The ability of observers to detect the seven second lesions is shown in Table 7. The second lesion was a large polyp in 14 of the examinations read by experienced readers, 14 of the examinations read by radiologists, and 15 of the examinations read by technologists, and was detected by all of the experienced readers, 93% of the radiologists, and 80% of the technologists (Table 7). The second lesion was a medium polyp in 35 of the examinations read by experienced readers, 33 of the examinations read by radiologists, and 38 of the examinations read by technologists, with detection rates of 37%, 21%, and 16% respectively.
Table 7.

Ability to detect a second lesion (i.e. a lesion other than the index lesion) in those patients with more than one lesion, split by observer group. There were seven such patients (ED-3). Data in table refers to numbers of observations.

<table>
<thead>
<tr>
<th>Observer group</th>
<th>Overall</th>
<th>Large polyps</th>
<th>Medium (R1-3)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Seen</td>
<td>Missed</td>
<td>Seen</td>
</tr>
<tr>
<td></td>
<td>N (%)</td>
<td>n (%)</td>
<td>n (%)</td>
</tr>
<tr>
<td>Experienced</td>
<td>27 (55)</td>
<td>22 (45)</td>
<td>14 (100)</td>
</tr>
<tr>
<td>Radiologist</td>
<td>20 (43)</td>
<td>27 (57)</td>
<td>13 (93)</td>
</tr>
<tr>
<td>Technologist</td>
<td>18 (34)</td>
<td>35 (66)</td>
<td>12 (80)</td>
</tr>
</tbody>
</table>

n = number
4.1.3.4 Prior experience levels

We performed a subset analysis comparing experienced readers whose \textit{a priori} experience exceeded 1000 examinations (4 individuals) with those whose experience was less than 1000 examinations (6 individuals), and found no significant difference: Detection rates for cancer, large polyp, medium polyps 100\%, 56\%, 43\% for the most experienced versus 86\%, 58\%, 53\% respectively. False positive rates were also similar; 80\% for the most experienced versus 84\%. 
4.1.4 DISCUSSION

As noted earlier in this chapter, highly experienced individuals agree that specific and supervised training is a prerequisite for acceptable performance [16] and that such training should involve interpretation of forty to fifty endoscopically validated examinations. Our data show that the overall sensitivity of novices trained using this scheme is significantly inferior to experienced observers. For example, both trained groups detected approximately 70% cancers, significantly less than the 92% detected by experienced observers. This discrepancy seemingly suggests that the training we administered (fifty cases) was inadequate if the aspiration is to reach the performance standard set by experienced observers.

However, it could be argued that our proposition, i.e. that trainees must achieve the competence of experienced observers, is flawed. A distinction can be made between 'best-achievable' and 'acceptable' performance. Radiologists with a subspecialty interest perform better than their generalist peers on average, due to their ability to make decisions based on prior experience [87]. Whether all radiologists interpreting CT colonography need to be as capable as those with extensive experience is a debate for the wider radiological community. The answer will depend on where the examination is positioned with respect to generalist and subspecialist camps. A number of observations will inform this. Firstly, subspecialisation is increasing and has affected radiology since the 1920s [101], not least because it is thought to benefit patients [102]. The ultimate position of CT colonography as a specialist examination is thus more likely today than previously. A parallel may also be drawn with barium
enema, which is widely considered a general examination despite compelling evidence that interpretation is best handled by those with extensive experience [65]. Furthermore from its inception, the diagnostic performance of CT colonography has been compared continually to colonoscopy, a better test than barium enema [103]. Comparisons between skilled colonoscopists and less capable colonographers damage the reputation of CT colonography [6].

It may be possible to stratify acceptable performance contingent on the clinical setting. For example, it has been argued for mammography that the highest aptitude is necessary for screening, where patients are asymptomatic and lesions are often difficult to detect [104]. The same might apply to CT colonography. Like mammography, colonography may also be used to investigate symptomatic patients (who actually constitute the largest group by far to have undergone the procedure in research studies), where lesions tend to be larger and easier to detect, potentially requiring less interpretative skill. This hypothesis is supported by our findings, which showed that detection rates increased in line with lesion size for all observer groups. However, while the trained novices found cancers the easiest index lesion to detect, whether a potential patient (or health-policy maker) would be happy with a 70% average detection rate is a subject for wider debate.

The aim of our study was not to investigate the performance characteristics of CT colonography. Rather, we aimed to determine the performance of novices relative to experienced observers, following training using a schedule in line with proposed guidelines [16]. This approach takes no account of individual aptitude. It is inevitable that some individuals will outperform others despite similar professional backgrounds.
and training. Our data revealed considerable overlap in individual performance between all groups. Notably, there were individuals from both trained groups whose accuracy exceeded the mean value obtained for experienced observers. Conversely, one experienced observer lay below the mean value achieved by both trained groups, even after excluding difficult-to-detect lesions. On the face of it, our data suggest that competence might be achieved after administering a training schedule based on 50 validated studies, but only by certain 'gifted' individuals. Merely undertaking such a schedule is insufficient to guarantee competency and it is self-evident that competent individuals will need to be identified in some other way, possibly by examination. Again, this is a subject for a wider debate. It should be noted that because our sample dataset was relatively small, the observed variability between observers would likely exceed the real variability because of sampling error. A larger study would likely show performance that regressed towards the mean for each group. Because of this it would be unwise to overemphasize the performance of individuals studied in the present study.

Considering aptitude further, we found no difference between the trained radiologists and radiographic technologists on average, despite the relative wealth of interpretative experience with CT in the former group. Also, the spread of individual abilities was similar between these two groups. This suggests that the paradigm for CT colonography interpretation differs from that encountered in routine CT scanning, with the result that radiologists may not have an intrinsic advantage (perhaps unless we are also considering extra-colonic lesions, something that we chose not to address). This may be explained by the fact that a single organ is being interrogated for a single disease (i.e. neoplasia) so an extensive medical knowledge base confers
no substantial advantage. Further, the skills required for colonic navigation are different than those employed when interpreting conventional CT studies and interpretation takes longer, with greater potential for fatigue and error [12]. Our data possibly support the concept that radiographic technologists may prove a valuable resource, especially when radiologists are in short-supply. This is already the case for the barium enema, and is cost-effective [105;106]. Other workers have investigated the value of using radiographic technologists and medical students to interpret CT colonography in the context of a second reader, finding that they can perform on par with radiologists [107].

Although our primary aim was to assess the relative performance of experienced observers and trained novices, we should explore the reasons behind the overall detection rate of only 57% for large polyps, which lags behind some studies [7] and meta-analyses [5;56]. This was undoubtedly heavily influenced by the disproportionately high percentage of flat adenomas and may not translate to series more representative of the general population: a third of large polyps were flat and one was invisible on CT, even in retrospect. Large series using dye-spray colonoscopy suggest that 13% to 15% of large adenomas are flat [108;109]. Ironically, the higher proportion in our study was a result of our attempts to make the dataset reflect everyday practice. We prevented investigators from submitting only their 'best' examinations by stipulating that accrual was chronologically consecutive. Some contributing centers had ongoing research relating to hereditary cancer, the effect of which was to increase our prevalence of flat lesions. The consequence of this was two-fold. Most obviously detection rates were reduced, but flat adenomas also diminished our power to discriminate between groups because they are a
challenge for all observers [110]. However, they can be detected if interpretation is careful [110]; one experienced observer identified two. The proportion of flat adenomas should be reported in future studies of CT colonography.

Our study does have limitations. We originally intended that all participating centres would contribute 10 patients but not all did so; three did not contribute at all because they could not satisfy protocol stipulations. The dataset, although designed to reflect what might be expected in an FOBT screening program, was by necessity a simulation. For example, cancers detected by screening are earlier stage than those presenting with symptoms [98-100] and, conversely, adenomas detected by FOBT are larger than those in asymptomatic groups. We have discussed the difficulties that the proportion of flat adenomas posed. Reading conditions were by necessity artificial. Interpretation is fatiguing [74] and practitioners are currently unlikely to read 20 examinations per day. However, this was a pragmatic necessity for this study and this paradigm has been adopted successfully in other high-profile studies that have investigated large numbers of observers from several centers [74]. Our original intention was also for observers to use their preferred platform but DICOM incompatibility issues prevented this. Instead we assembled the dataset onto laptops that could be transported easily to each center. These laptops had high-resolution 17-inch screens and the software used a “2D with 3D for problem solving” approach, which was the preferred method of analysis for the majority of experienced readers at the time. All normal software functions were preserved on the laptop. Because some readers had been trained using another platform locally, we took care to ensure that the software used for testing observers was easy to learn, and the study supervisor was available at all times to help if necessary. While there is some evidence that the
type of platform used does not influence accuracy [74], it is also possible that
accuracy may have improved had a primary 3D read been possible [7]. However, it
should be stressed that we aimed to investigate the relative performance of observers
and not the confounding effect of the visualisation platform. Whether the reading
platform used has a differential effect on experienced versus less experienced
readers clearly merits further research. Laptops also meant that study loading times
were prolonged compared to a workstation, which may have frustrated some readers.
Investigators have examined the effect of administering an identical training schedule
to novices using a teaching file and test set [111] but we decided to leave the patient
selection and schedule largely to the discretion of the local trainer (beyond
stipulations relating to the number of examinations and length of training) because we
considered this better reflected current teaching practice. As a result, differences in
performance could potentially be explained by variations in the quality of local
training, just as occurs in residency programs in general. For example, some trainers
may have emphasised the need for a careful soft-tissue read to look for flat lesions
whereas others may have not. Whether an identical training scheme and materials
administered via a training course is superior or inferior to more prolonged but less
standardised local training is a subject that needs further investigation. We have
already stated that because our dataset was relatively small, observed variability
between readers may be increased as a consequence. Also, not all observers read
the same examinations, to avoid recall bias because of interpreting patients from their
own center, but we did balance the prevalence of abnormality across datasets so that
they remained comparable.
In conclusion, experienced observers asked to interpret CT colonography perform significantly better on average than novices who have been trained using 50 endoscopically validated examinations. However, individual performance is variable and some trainees may outperform some experienced radiologists. On average, we found no performance difference between trained radiologists and radiographic technologists, suggesting that prior interpretation of conventional abdominal CT may not confer benefit for CT colonography.
4.2 CT colonography: Interpretative performance in a non-academic environment

4.2.1 INTRODUCTION

Most patients undergoing CT colonography in the UK are examined in district general hospitals, outside of an academic environment (Chapter 3.1). Because the technique has disseminated rapidly, most practitioners have had little or no specific training in either interpretation or technical performance. In the face of this, there is evidence that CT colonography examinations are difficult to interpret, even when individualised training has been administered as found in the preceding study, Chapter section 4.1.

Although we have established the relative performance characteristics of readers with differing experience in an academic setting (Chapter section 4.1), there is a paucity of equivalent data for CT colonography performed in day-to-day clinical practice outside of a research study. While research studies offer the opportunity for within-subject verification of CT findings, in day-to-day clinical practice this opportunity for training is limited by economic and research constraints, and patient inconvenience. Like many health technologies that have disseminated rapidly, initial learning is often acquired "on the hoof", because established and validated training programmes are just not widely available.
The survey described earlier in this thesis in Chapter section 3.1, found that of the fifty departments performing CT colonography, only three (6%) had initiated the service via clinical research. This survey also found that CT colonography was most often interpreted by radiologists with a subspecialty interest in gastrointestinal radiology. The purpose of the present study was to investigate interpretative accuracy and reporting times for a representative group of sub-specialist radiologists who were interpreting CT colonography in routine clinical practice.

4.2.2 MATERIALS AND METHODS

4.2.2.1 CT colonography data
We selected a dataset of fifteen CT colonography studies from cases accrued for the previous study described above (Chapter Section 4.1). The 15 cases selected for the present study were chosen on the basis that they had been categorised correctly by the majority of experienced observers in this prior study; i.e. they were “easy” for experienced readers.

These 15 CT colonography examinations were categorised according to the largest colonic lesion (the 'index lesion') found at same-day colonoscopy performed by experienced practitioners as follows: five patients had a cancer, 5 patients had a polyp 1cm or larger, one patient had an 8mm polyp, and four patients were normal. Polyp size was defined by the colonoscopic measurement, estimated using adjacent biopsy forceps.
4.2.2.2 Observers

Thirteen radiologists from seven UK centres took part in the present study: one centre contributed three observers, four centres contributed two observers and two centres contributed one observer. All observers were fully-accredited radiologists of Consultant grade who held non-academic positions but had a declared sub-specialty interest in gastrointestinal radiology ('non-academic sub-specialists'). All centres had implemented CT colonography into their routine clinical practice for investigation of patients with symptoms of colorectal cancer as a response to favourable reports in the radiologic literature. At the time of the present study, their individual experience varied considerably (Table 8). None of these observers had undergone formal training in CT colonography interpretation, via a workshop for example.

All observers were unaware of the prevalence of abnormality in the dataset and were unaware of how observers in the prior study had fared when interpreting the same cases.
Table 8. Prior CT colonography interpretation experience of non academic subspecialist radiologists, split by whether examinations had been colonoscopically validated or not.

<table>
<thead>
<tr>
<th>Reader</th>
<th>No. of validated cases</th>
<th>No. of non validated cases</th>
<th>Total no. of cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>54</td>
<td>60</td>
</tr>
<tr>
<td>3</td>
<td>10</td>
<td>190</td>
<td>200</td>
</tr>
<tr>
<td>4</td>
<td>25</td>
<td>25</td>
<td>50</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>30</td>
<td>33</td>
</tr>
<tr>
<td>6</td>
<td>30</td>
<td>170</td>
<td>200</td>
</tr>
<tr>
<td>7</td>
<td>8</td>
<td>72</td>
<td>80</td>
</tr>
<tr>
<td>8</td>
<td>10</td>
<td>20</td>
<td>30</td>
</tr>
<tr>
<td>9</td>
<td>50</td>
<td>300</td>
<td>350</td>
</tr>
<tr>
<td>10</td>
<td>25</td>
<td>75</td>
<td>100</td>
</tr>
<tr>
<td>11</td>
<td>10</td>
<td>90</td>
<td>100</td>
</tr>
<tr>
<td>12</td>
<td>15</td>
<td>105</td>
<td>120</td>
</tr>
<tr>
<td>13</td>
<td>10</td>
<td>90</td>
<td>100</td>
</tr>
</tbody>
</table>

4.2.2.3 Interpretation

The 13 observers read all 15 examinations using large screen laptop computers and commercially-available software as in the preceding study (Chapter section 4.1). Once again, software functionality was demonstrated to observers during an initial training session administered by the principal investigator of this thesis.

Reading took place over one day in the observers' own hospital in a quiet room, uninterrupted by day-to-day clinical activity. Observers were asked to complete a dedicated study proforma, similar to that used in the study described in chapter 4.1.
i.e. classifying each of the 15 CT colonography examinations as either normal (if no colonic polyp exceeding 5mm in maximal diameter was found) or abnormal, and then recording the colonic findings as follows: Any lesion found was categorised as either cancer, large polyp or medium polyp, its size recorded (in mm), and the interpretation start/finish time noted (interpretation times are revealed and discussed in the following chapter). The location of any colonic lesion was also indicated by colonic segment, via reference to a study diagram available throughout the reading. Scan coordinates were noted, for prone and supine acquisition if a lesion was visible on both. The principal investigator of this thesis was present throughout to provide technical support but no training beyond software functionality was given. Case order was randomised and all observers then read the cases in the same order. No conferring was permitted.

4.2.2.4 Analysis

Observer responses were collated by the study coordinator and compared with the known diagnostic category for each individual case, in order to determine accuracy (defined as the number of cases correctly classified, including normal cases). The performance of these sub-specialist radiologists was then compared to each of the three groups of observers who had participated in the prior study of diagnostic accuracy, described in Chapter section 4.1, so that their relative performance could be judged. Analysis of variance was used to compare diagnostic accuracy and interpretation time across groups. The relationship between accuracy and the number of CT colonography examinations reported prior to the study was determined using Spearman's Rank correlation. Statistical significance was attributed at a probability
level of 5%. Analysis was performed using Stata 7.0 (StataCorp, College Station, Texas, USA).

4.2.3 RESULTS

The number and percentage of examinations in which the index lesion was correctly identified, and normal examinations were correctly categorized by the non-academic sub specialist observers are shown in Table 9.
Table 9. Number and percentage of 15 CT examinations correctly classified by the 13 non academic subspecialist observers.

<table>
<thead>
<tr>
<th>CT category &amp; examination no.</th>
<th>Number of ‘correct’ examination categorisations n=13</th>
<th>% correct categorisations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cancer 1</td>
<td>10</td>
<td>77</td>
</tr>
<tr>
<td>Cancer 2</td>
<td>13</td>
<td>100</td>
</tr>
<tr>
<td>Cancer 3</td>
<td>8</td>
<td>62</td>
</tr>
<tr>
<td>Cancer 4</td>
<td>12</td>
<td>92</td>
</tr>
<tr>
<td>Cancer 5</td>
<td>13</td>
<td>100</td>
</tr>
<tr>
<td>Large polyp 6</td>
<td>9</td>
<td>69</td>
</tr>
<tr>
<td>Large polyp 7</td>
<td>3</td>
<td>23</td>
</tr>
<tr>
<td>Large polyp 8</td>
<td>10</td>
<td>77</td>
</tr>
<tr>
<td>Large polyp 9</td>
<td>11</td>
<td>85</td>
</tr>
<tr>
<td>Large polyp 10</td>
<td>9</td>
<td>69</td>
</tr>
<tr>
<td>Medium polyp 11</td>
<td>6</td>
<td>46</td>
</tr>
<tr>
<td>Normal 12</td>
<td>7</td>
<td>54</td>
</tr>
<tr>
<td>Normal 13</td>
<td>13</td>
<td>100</td>
</tr>
<tr>
<td>Normal 14</td>
<td>13</td>
<td>100</td>
</tr>
<tr>
<td>Normal 15</td>
<td>11</td>
<td>85</td>
</tr>
</tbody>
</table>
Overall, 56 of 65 (86%) cancers were correctly identified; 42 of 65 (65%) of large polyps and 6 of 13 (46%) small polyps were correctly identified; and 44 of 52 (85%) normal cases were correctly categorised. Three examinations caused particular difficulty, one cancer located at the caecal pole (Figure 10a,b) which only 8 of 13 (62%) observers correctly identified, one large polyp which was only visible on the prone acquisition (and was partially submerged in untagged fluid) [Figure 11 a,b] which 3 of 13 (23%) observers correctly identified and the single medium sized polyp, also only visible on the prone acquisition (submerged in untagged fluid on supine) which only 6 of 13 (46%) observers correctly identified [Figure 12].
Figure 10a, b

Circumferential cancer located at the caecal pole on a. supine 2D axial view (arrows) and b. 3D endoluminal 'volume rendered' display (arrows): Correctly identified by 8 of 13 (62%) observers.
Figure 11 a, b.

Large polyp (12mm maximal diameter), located in the recto-sigmoid on a. prone 2D axial view (arrow) and b. 3D endoluminal 'volume rendered' display (arrows): Polyp was only visible on the prone acquisition, partially submerged in untagged fluid and was correctly identified by 3 of 13 (23%) observers.
Figure 12.

Medium sized polyp (8mm maximal diameter), located in the ascending colon on prone 2D axial view (ring); only visible on the prone acquisition (submerged in untagged fluid on supine) and correctly identified by 6 of 13 (46%) observers.
4.2.3.1 Individual non-academic sub-specialists performance

The number of correctly categorised cases and the accuracy (i.e. percentage of correctly categorised cases with 95% confidence intervals) are shown for each individual observer in Table 10. There was wide inter-observer variability in performance with accuracy ranging between 53 and 93%. All bar one observer correctly categorised at least 10 of the 15 examinations. Of the 13 observers, three missed two cancers, three missed one cancer and the remaining seven correctly identified all the cancers.
Table 10. Individual accuracy of non academic sub specialist radiologists

<table>
<thead>
<tr>
<th>Reader</th>
<th>Number of 'correct' cases n=15</th>
<th>% Accuracy (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8</td>
<td>53 (27, 79)</td>
</tr>
<tr>
<td>2</td>
<td>10</td>
<td>67 (38, 88)</td>
</tr>
<tr>
<td>3</td>
<td>10</td>
<td>67 (38, 88)</td>
</tr>
<tr>
<td>4</td>
<td>11</td>
<td>73 (45, 92)</td>
</tr>
<tr>
<td>5</td>
<td>11</td>
<td>73 (45, 92)</td>
</tr>
<tr>
<td>6</td>
<td>11</td>
<td>73 (45, 92)</td>
</tr>
<tr>
<td>7</td>
<td>11</td>
<td>73 (45, 92)</td>
</tr>
<tr>
<td>8</td>
<td>11</td>
<td>73 (45, 92)</td>
</tr>
<tr>
<td>9</td>
<td>12</td>
<td>80 (52, 96)</td>
</tr>
<tr>
<td>10</td>
<td>12</td>
<td>80 (52, 96)</td>
</tr>
<tr>
<td>11</td>
<td>13</td>
<td>87 (60, 98)</td>
</tr>
<tr>
<td>12</td>
<td>13</td>
<td>87 (60, 98)</td>
</tr>
<tr>
<td>13</td>
<td>14</td>
<td>93 (68, 100)</td>
</tr>
</tbody>
</table>
4.2.3.2 Observer group performance

Overall, the accuracy of non-academic sub-specialists was 75% (standard deviation of 10). They were significantly more accurate than the radiographic technicians (56%, SD 12; p=0.003) and significantly less accurate than the experienced radiologists (88%, SD 11; p=0.04) studies in Chapter section 4.1. There was however no significant difference between their accuracy and that of the less experienced, trained radiologists from academic centres (71%, SD 20; p=0.48).

4.2.3.3 Effect of prior interpretative experience

There was a significant positive correlation between prior experience i.e the total number of examinations read in the past and observer accuracy (correlation coefficient of 0.52 (p=<0.001), irrespective of whether this experience was using colonscopically validated or non validated CT colonography examinations; correlation coefficient of 0.39 (p=0.01) and 0.53 (p=<0.001) respectively.

4.2.4 DISCUSSION

We found that, on average, sub-specialist radiologists in the UK interpreted CT colonography examinations as accurately as radiologists who had been trained using 50 colonoscopically validated cases. This number of training cases has been suggested as sufficient to attain competence [16;70], albeit without a large evidence base. These results may appear encouraging, bearing in mind that no observer had received formalised training. However, our dataset was skewed in that all
abnormalities had been detected previously by a majority of experienced radiologists investigating CT colonography in an academic setting as described in Chapter section 4.1. There was also a deliberate attempt to increase the prevalence of larger lesions because we were most interested in detection of symptomatic cancer rather than asymptomatic polyps. Detection of symptomatic cancer is the clinical rationale for the vast majority of CT colonography studies performed in the UK.

The radiologists in the present study detected 86% of cancers overall, a figure similar to the proportion detected in the national UK barium enema audit (85%) [112;113]. Whether such performance is acceptable is beyond the scope of this article but there is clearly potential for improvement. These data could be interpreted variously. Some might argue that CT colonography is no better than barium enema while others might argue a detection rate of 86% is encouraging taking into account the fact that participants had no formal training. We were also able to show that, on average, increasing local experience was significantly associated with improved accuracy overall. However, this relationship will be confounded on an individual level by aptitude - the second worst performer (67% accurate) was also the second most experienced, having previously read 200 scans.

As for the preceding study (Chapter section 4.1), we found wide variability for detection of large and medium polyps, with overall performance falling well below that suggested by previously published data. For example, sensitivity for polyps 1cm or larger was 65%, as opposed to between 81-93% suggested by meta-analyses [5;56;76]. Nevertheless, we also identified individual observers within this group who had a clear aptitude for CT colonography; three observers identified five of the six
(83%) polyp cases. While the vast majority of published research studies relating to CT colonography have not used patients' representative of a screening population, a role in screening (i.e. adenoma detection) is continually advocated. Our results suggest strongly that day-to-day clinical practice alone is insufficient experience to recommend that radiologists (or others) report screening examinations unless they have undergone specific training or can demonstrate audited measures of satisfactory performance in this context.

The limitations of this study are shared with the preceding study, and these have been discussed fully in Chapter section 4.1.

In summary, the accuracy of untrained sub-specialist radiologists reading CT colonography examinations in day-to-day clinical practice is comparable to the performance of radiologists trained using 50 validated cases. However there is wide variability in individual performance, which generally falls short of the average performance suggested by meta-analysis of published data. Experience improves accuracy but alone is insufficient to determine competence. Our findings reiterate the need for formal training and incontrovertibly indicate the need for formal accreditation if the role of CT colonography is extrapolated to screening.
CHAPTER 5

INTERPRETATION TIME

5.1 CT colonography interpretation times: Effect of reader experience, fatigue, and scan findings in a multi-centre setting

5.1.1 INTRODUCTION

As discussed earlier in this thesis, interpretation of CT colonography can be both time consuming and fatiguing. The time taken for readers to interpret CT colonography is influenced by a number of intrinsic and extrinsic factors. Extrinsic factors (i.e. those not under the direct control of the radiologist) might include the quality of bowel preparation and distension in an individual patient, colonic configuration, the reading environment, the need for other organ review, and the number of cases that must be interpreted per reading session. Intrinsic factors also play a substantial role. For example, some readers are more inclined to perform a meticulous search for flat or diminutive polyps, which prolongs interpretation. The reading platform preferred by the observer will also play a part; longer interpretation times are generally encountered when using a primary 3D interpretation [53;54].

However, the most obvious intrinsic factor that may influence interpretation time is observer experience, and it is presently unclear whether experienced readers
interpret faster than those who are less experienced, or, conversely, whether they take longer because they are more careful. One study found that readers speeded up as they became more experienced [15], whereas a second suggested that sensitivity increased when observers slowed down [14]. Some authors have specifically investigated rapid reporting (less than 5 minutes), finding no significant effect on sensitivity and specificity [114]. To our knowledge there is no published data examining the relationship between interpretation time and reader experience or how reading time may influence diagnostic accuracy. We aimed to determine the relationship, if any, between reader experience and interpretation times when using a 2D primary read, and to examine how this relates to accuracy.

5.1.2 MATERIALS AND METHODS

This study utilised data collected as described in Chapter section 4.1 and 4.2 and assessed the interpretation time for the same observer groups. As part of these studies, observers loaded each individual case into the software and then recorded the time when they started interpretation, noting again when they had finished. Times were recorded using individualised report proformas. Observers were asked to ignore extracolonic findings and limit their interpretation to the colon. They noted, measured, and categorised any polyp or cancer observed.

Results were collated by the principal investigator of this thesis, tabulated in a worksheet (Microsoft Excel® version 2000, Microsoft Corporation, Redmond, WA, USA), and transferred to Stata 8.0 (StataCorp, College Station, Texas, USA) for additional analyses by a statistical advisor.
5.1.2.1 Statistical analysis

The mean interpretation time and standard deviation were calculated for each observer group according to the case category (cancer, medium polyp, small polyp, normal). Overall, the distribution of individual case interpretation times was skewed towards shorter times and so the data was logarithmically transformed to provide a normal distribution. Subsequently, for data derived from the larger study described in Chapter section 4.1, differences between the interpretation times of the three observer groups; of reading normal versus abnormal cases; and the effects of fatigue were analysed using two level cross classified multi-level linear regression and, due to the logarithmic data transformation, expressed as ratios of interpretation times (with 95% confidence intervals). Fatigue was assessed in two ways; by comparing interpretation times for the last five cases, with an initial five (excluding the first case because many observers had used this case to familiarise themselves with the software); and by comparing interpretation times between cases read in the morning versus those read in the afternoon. Linear regression was used to compare the relationship between individuals' overall accuracy (calculated as the percentage of correct assessments including normal cases and false positive assessments) and their mean interpretation times. These data were then combined for each observer group. Finally, the data for sub-specialist, non-academic radiologists reading a subset of 15 CT colonography examinations (described in chapter 4.2) was compared, using analysis of variance, to these three observer groups, who had read the same 15 cases as part of their individualised datasets.
5.1.3 RESULTS

Mean interpretation times (with corresponding standard deviation) and the number of cases interpreted for each observer group, according to the case category, are shown in Table 11.
Table 11. Mean interpretation times in minutes (SD) and the number of cases interpreted (n) according to reference case category, observer group and for all observers.

<table>
<thead>
<tr>
<th>Reference case category</th>
<th>Experienced radiologists</th>
<th>Less experienced radiologists</th>
<th>Radiographic technicians</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cancer</td>
<td>11.8 (SD 6.7) n=51</td>
<td>17.1 (SD 9.6) n=48</td>
<td>17.3 (SD 8.2) n=55</td>
<td>15.4 (SD 8.6) n=154</td>
</tr>
<tr>
<td>Large polyp</td>
<td>12.4 (SD 5.3) n=95</td>
<td>17.2 (SD 7.9) n=91</td>
<td>16.8 (SD 7.5) n=97</td>
<td>15.4 (SD 7.3) n=283</td>
</tr>
<tr>
<td>Small polyp</td>
<td>10.7 (SD 4.4) n=31</td>
<td>16.2 (SD 8.0) n=29</td>
<td>18.8 (SD 9.4) n=32</td>
<td>15.3 (SD 8.3) n=92</td>
</tr>
<tr>
<td>Normal</td>
<td>9.9 (SD 4.5) n=183</td>
<td>14.5 (SD 7.2) n=176</td>
<td>17.0 (SD 7.4) n=196</td>
<td>13.9 (SD 7.2) n=555</td>
</tr>
<tr>
<td>Total</td>
<td>10.9 (SD 5.2) n=360</td>
<td>15.7 (SD 7.9) n=344</td>
<td>17.1 (SD 7.7) n=380</td>
<td>14.6 (SD 7.5) n=1084</td>
</tr>
</tbody>
</table>
Overall, on average, observers spent 14.6 minutes interpreting each individual case. Experienced radiologists read faster than the less experienced observers (radiologists and radiographic technicians) for all case categories. The overall ratio of interpretation time for these less experienced radiologists and radiographic technicians compared to experienced radiologists was 1.41 (CI 1.12, 1.78) and 1.61 (CI 1.28, 2.02) respectively (P=<0.001). The ratio for radiographic technicians compared to the less experienced radiologists was 1.14 (CI 0.91, 1.44) indicating little difference between these less experienced groups. There was greater individual variation in interpretation time amongst observers from the less experienced groups when compared to experienced radiologists, demonstrated by larger standard deviations. This finding was further illustrated when individual times for each observer, categorised according to their group, were displayed graphically (Fig 13).

Experienced and less experienced radiologists, but not radiographic technicians, interpreted normal cases significantly faster than abnormal cases; the ratio of interpretation time for abnormal cases compared to normal cases was 1.24 (CI 1.09, 1.42) [P=<0.001] for experienced radiologists, 1.17 (CI 1.02, 1.33) [P=0.02] for less experienced radiologists and 1.00 (CI 0.88, 1.14) [P=0.99] for radiographic technicians. Also, the difference in interpretation time between experienced radiologists and radiographic technicians was proportionally larger for normal cases than abnormal cases; the ratio of interpretation times for radiographic technicians compared to experienced radiologists was 1.79 (CI 1.41, 2.27) for normal cases and 1.44 (CI 1.14, 1.83) for abnormal cases. This discrepancy was much less marked.
when comparing less experienced radiologists to experienced radiologists; 1.45 (CI 1.14, 1.86) and 1.36 (CI 1.07, 1.74) for normal and abnormal cases respectively.

Overall, observers spent significantly less time interpreting cases read in the afternoon compared to the morning (Table 12). In addition, their interpretation times decreased by a greater extent over the course of the study, such that on average, observers only took 71% of the time to report cases at the end of the study reading compared to the beginning.
Table 12. Effect of fatigue on interpretation times (Odds ratio).

<table>
<thead>
<tr>
<th>Variable</th>
<th>Case Group</th>
<th>Ratio (95% CI)</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time of day</td>
<td>Morning</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Afternoon</td>
<td>0.90 (0.85, 0.95)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>First/last cases</td>
<td>Initial 5 cases (2-6)</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Final 5 cases</td>
<td>0.71 (0.63, 0.80)</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

Results of individual observer performance have been discussed in Chapter section 4.1; the median accuracy for each observer group was 76.9% for experienced radiologists, 69.2% for less experienced radiologists and 66.7% for radiographic technicians. For this study, an individual's overall accuracy was compared to their interpretation time and the results are summarised below in table 3.
Table 13. Effect of interpretation time on accuracy for each observer group. The regression coefficients indicate the amount that accuracy changes for a one-unit increase in report time (on the logarithmic scale).

<table>
<thead>
<tr>
<th>Group</th>
<th>Regression Coefficient (95% CI)</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expert</td>
<td>-7.1 (-17.8, 3.5)</td>
<td>0.18</td>
</tr>
<tr>
<td>Radiologist</td>
<td>1.2 (-9.3, 11.7)</td>
<td>0.82</td>
</tr>
<tr>
<td>Technician</td>
<td>16.3 (0.6, 32.0)</td>
<td>0.04</td>
</tr>
</tbody>
</table>

Also the effect of interpretation time on accuracy for each observer group is displayed graphically (Fig 14). Accuracy improved significantly with longer interpretation times for radiographic technicians but accuracy was unaffected by the length of interpretation time taken by less experienced radiologists. For experienced radiologists, Interpretative accuracy fell with longer interpretation times although this effect did not reach significance.
Figure 13. Interpretation times (in minutes) for each case, categorised according to observer group.
Figure 14. Effect of interpretation time on accuracy for each observer group displaying the amount that accuracy changes for a one-unit increase in report time.
5.1.3.1 Interpretation time for non academic sub-specialists (chapter 4.2)

Overall, the non-academic sub-specialists took, on average, 12.4 minutes per examination (SD 3.5). For the 15 cases used in their study (chapter 4.2) they interpreted cases significantly faster than the trained radiologists and radiographic technicians (16.2 mins, SD 4.3; p=0.05 and 17.9 mins, SD 6.4; p=0.005 respectively). Interpretation time however was similar to experienced observers from the prior study who reported examinations in 11.7 mins (SD 2.3; p=0.74).
5.1.4 DISCUSSION

For successful implementation both as a symptomatic and screening modality, CT colonography interpretation must be time efficient. This study, which examined multiple observers of differing experience across several European countries, found that, on average, readers took less than fifteen minutes to interpret a case (excluding extra-colonic organ review). These interpretation times are similar to previous multi-observer studies utilising a primary 2D interpretation method [2;74] and shorter than those employing a primary 3D approach [7]. Such interpretation times are comparable to optical colonoscopy [2], but prolonged when compared to barium enema.

From a health economic perspective, faster reporting times are desirable, but possibly not at the expense of sensitivity and specificity. Prior to this study we wondered whether experienced observers took longer to report equivalent studies than less experienced individuals, because they are more careful. When teaching on CT colonography workshops, it had been our anecdotal experience that novice observers tended to read too fast for them to be able to detect polyps reliably. However, our data show that experienced observers read significantly faster than others, taking on average ten minutes for normal studies. This latter finding is particularly relevant for screening where the large majority of examinations will have no significant pathology (i.e. no adenomatous polyp measuring 6mm or larger).

It should be borne in mind that the less experienced observers from the academic settings used for this study are not directly comparable to the novices encountered at many CT colonography workshops. In particular, they had received some training, and this will hopefully have emphasised the need to be thorough and meticulous.
Shorter interpretation times are potentially associated with impaired accuracy. A study [73] of four radiologists who were relatively inexperienced with CT colonography found that the two observers with the longest interpretation times (averaging 21 and 25 minutes respectively) detected more polyps than those who read faster (14 and 15 minutes respectively). We found this to be the case for our radiographic technician group, whose accuracy improved with longer interpretation times. In contrast, accuracy for our less experienced radiologist group seemed unaltered by the time taken for interpretation and, conversely, the experienced radiologists who took longest to read a case performed on average worse than the rest of their group. The reasons underlying this observation are obscure and certainly merit further investigation. Whatever the explanation, individual interpretation times varied considerably within each observer group, with some readers reporting very quickly, irrespective of their experience. This finding is potentially encouraging since the best performing experts also reported fastest, perhaps alleviating concerns relating to fatigue, workflow and CT throughput. It is also reassuring that recently abstracted data shows that cases can be reported rapidly (i.e. less than 5 minutes) without sacrificing sensitivity or specificity [114].

When comparing the non academic subspecialists to these three observer groups from an academic setting, we found the sub-specialist radiologists interpreted as fast as experienced observers in academic centres, which has positive implications for workflow in busy clinical practice.
Like other studies [74], the reading environment was necessarily artificial because of pragmatic considerations (observers read 40 cases over two days), although it should be noted that some screening centres require equivalent work rates of their radiologists. Our findings may not be generalisable to routine clinical practice therefore. For example, observers may have been fatigued and lost concentration over the study period. Conversely, it is also possible that concentration was enhanced overall by a carefully controlled reading environment, away from routine clinical practice. The desire to do well, especially in comparison to their peers, may also have influenced readers' performance.

We were unsure how fatigue would affect interpretation time. For example, interpretation time could increase as a consequence of diminished concentration. Alternatively, it could be argued that interpretation might shorten if the desire to finish the study rapidly was not suppressed and vigilance suffered as a result. We attempted to assess the effect of fatigue on interpretation time and found a marked reduction in the time taken to read the last five cases compared to an initial five. This could suggest that when provided with a large number of cases to read, observers may be less thorough towards the end of the reading period. Alternatively, the fall in interpretation time may also be due to increasing familiarity with both the software and study environment. We attempted to minimise confounders by using highly intuitive and relatively simple software, and undertaking reading in the observer's own hospital. We also ignored the time taken to read the first case, where the learning curve was likely to be greatest.
Computer aided detection (CAD) for CT colonography has recently gained regulatory approval and these systems will undoubtedly influence interpretation times in the future. It is unclear whether they will reduce interpretation times, which is hoped for, or indeed prolong them if the regulatory requirements stipulate that the study be first read in its entirety unaided by the software.

In summary, experienced radiologists interpret CT colonography faster than less experienced radiologists and radiographic technicians who had been trained using 50 cases, and spend less time interpreting normal cases. Interpreting cases more slowly improves accuracy amongst radiographic technicians but seems to reduce accuracy amongst experienced radiologists. All observer groups interpreted cases faster as the study period progressed. Comparison with non academic sub-specialist radiologists showed this group report as fast as experienced academic radiologists. More research relating to the optimal reading environment and work-flow is needed.
CHAPTER 6

POLYP MEASUREMENT

6.1 Polyp measurement by CT colonography: Agreement with colonoscopy, and effect of viewing conditions on inter- and intra-observer agreement

6.1.1 INTRODUCTION

Most colorectal cancer arises from pre-existing adenomatous polyps. While most adenomas are destined never to become cancer, malignant transformation occurs in a minority. Malignant transformation is multifactorial [115;116] and is positively correlated with the degree of dysplasia, villous histology, and the maximum diameter of the polyp [117;118]. Only 1% of adenomas measuring less than 10mm are malignant in contrast to 50% that measure over 20mm [119]. Additionally, patients with an adenoma measuring 10mm or more have a higher incidence of both synchronous and metachronous advanced adenomas [120-122]. As a result, individuals with larger polyps are at higher risk for developing colorectal cancer than the general population and surveillance guidelines advise more frequent examination in this group [123;124].
The biological significance of an adenoma is thus heavily influenced by its maximum diameter. Diameter can be determined during colonoscopy, most often by comparison with adjacent biopsy forceps. This procedure can be followed immediately by polypectomy. CT colonography is increasingly advocated as an alternative screening modality that is less invasive and safer than colonoscopy. However, immediate polypectomy is not possible and when a polyp is detected recommendations for future management must be made on the basis of measurements obtained from the CT dataset. In general, large polyps will require subsequent endoscopy for polypectomy whereas small polyps may be safely left in situ because the risk of malignant transformation is outweighed by the small but significant risk of adverse-events related to colonoscopy and, the cost and inconvenience of a second procedure [70]. It is therefore important to establish that measurements of polyp diameter obtained at CT colonography are both accurate and reproducible. However, while several published studies have focussed on the variability of polyp measurements obtained during colonoscopy [125-127], at the time of writing no study has determined this for CT colonography. A degree of inter- and intra-observer disagreement is inevitable and this could potentially be confounded by the viewing conditions (for example window width and level), the method of image rendering used, and the experience of the observer.

We aimed to determine the level of agreement between CT colonography measurements of maximum polyp diameter and their colonoscopic equivalent and to additionally determine the level of inter- and intra-observer measurement agreement for CT, including the effects of different viewing conditions and prior expertise.
6.1.2 MATERIALS AND METHODS

6.1.2.1 Polyp dataset preparation and CT technique

Forty-eight polyps in 24 patients were identified by the principal investigator of this thesis from a database of CT colonography examinations accumulated as part of a study comparing CT with optical colonoscopy [57;128]. All polyps were protruberant with no morphologically flat lesions. Their diameters ranged between 2mm and 12mm. All sizes were represented in 1mm increments, excepting 11mm; four were 12mm, five 10mm, one 9mm, three 8mm, two 7mm, seven 6mm, fourteen 5mm, six 4mm, three 3mm and three 2mm. The reference size was defined as the maximal diameter estimated during colonoscopy, performed by a single expert endoscopist who had performed over 5000 colonoscopies. Maximal polyp diameter was estimated by comparison with adjacent open biopsy forceps, measuring 7mm, pushed against the polyp. Twelve patients had one polyp, seven had two, one had three, two had four, one had five, and one had six polyps.

In brief, CT colonography had been performed following full bowel preparation and distension using carbon dioxide for insufflation. Prone and supine scanning was performed using a 4-detector row CT (LightSpeedPlus, General Electric, Milwaukee, Wisconsin, USA), collimation of 1.25mm to 2.5mm, pitch 1.5, and mA 50 to 100. This technique has been described in detail previously [57]. Same day optical colonoscopy was performed following CT colonography.

The 24 individual patient studies were anonymised, their order randomised, and each then given a study number. The principal investigator of this thesis determined the
segmental location and CT co-ordinates for each of the known 48 polyps and recorded this on a study sheet. This was achieved by reference to the previous CT colonography and optical colonoscopy reports, and by viewing the prone and supine CT datasets for each study on a commercially available workstation using proprietary software (Advantage Windows 4.1 and Navigator colon package, General Electric, Milwaukee, Wisconsin, USA). The axial slice number for the epicentre of each individual polyp was noted for the prone and supine studies individually, if visible on both. This procedure facilitated polyp identification and location for the study readers (since polyp detection was not an aim of this study) and allowed them to select which one of the paired studies from which to make their measurements.

6.1.2.2 observers

Four observers interrogated each of the 48 polyps. The observers were unaware of the reference measurement for each polyp and also unaware of the distribution of polyp sizes in the dataset. All four observers were familiar with the proprietary software for reporting routine abdomino-pelvic CT and the principal investigator of this thesis also ensured they were familiar with the display and measurement functions required for 2D and 3D blinded assessment (see below). Three of the observers (numbers 1 to 3) had previous experience of CT colonography interpretation with a minimum of 150 examinations each at the time of the study while the fourth (observer 4) had no prior experience and only received instruction on how to use the software, not on how best to measure polyps. Each observer was asked to measure the maximal diameter of each polyp indicated on the study sheet using one of three image display methods in turn: 'colonography window', 'abdominal window', and '3-D
window'. Measurements were made using the same proprietary workstation and software used by the principal investigator of this thesis to select the polyps.

6.1.2.3 Polyp measurement

A four quadrant display was used (4:1) and the first measurement was made from the 2-D axial images or 2-D multiplanar or oblique reformatted images, on either the prone or supine acquisition, according to whichever of these was felt by the observer to better depict the maximal diameter of the polyp. The observer was also able to magnify the image according to their individual preference. The workstation was set to measure from 2-D images and the standard colonography window display used at our institution was used for viewing (window width 1500, level -150, Fig. 15a). The measurement was made by placing software calipers across what was judged to be the maximal diameter of the polyp. In order to reduce bias, the screen annotation function was disabled during this procedure with the result that the observer was unaware of the value of the measurement made. The observer then changed the display to standard abdominal viewing windows (width 400, level 40, Fig. 15b) and made a second measurement of maximal polyp diameter in an identical fashion to the first, again unaware of the value of the measurement made. Finally, the observer switched to a default endoluminal surface-rendered perspective display (Fig. 15c), repositioned the viewing angle to best depict the polyp, and made a final measurement of maximal diameter after changing the software settings to account for diameter measurement using 3-D rendering. Once this measurement had been made the screen annotation function was enabled and the observer recorded all three individual measurements on the study sheet. Readers were unaware of each other's results.
In order to assess intra-observer agreement, each observer repeated the three measurements on a subset of 10 polyps chosen by the study co-ordinator from the study dataset, which broadly represented all polyp sizes. Each observer re-measured the same subset of polyps and the order of these was again randomized before measurement by all observers.

Once all measurements were completed, the study sheets were collated by the study coordinator and individual observer measurements for each polyp transferred to an Excel Worksheet (Microsoft Excel® version 2000, Microsoft Corporation, Redmond, WA, USA). Analysis was performed by a statistical advisor using Stata (version 7, Statacorp LP, Texas, USA) and MLwiN (version 1.10, Institute of Education, University of London).
Figure 15. A 65-year old male. CT colonography using a 4-detector row machine, collimation of 1.25mm, pitch 1.5, and mA 100. 6mm sigmoid polyp viewed on 2D 'colonography window'.
Fig 15 B; Same patient as Fig 1A. 6mm polyp viewed on 2D 'abdominal window'.
Fig 15C. Same patient as Fig 1A. 6mm polyp viewed on 3D 'surface rendered' view.
6.1.2.4 Statistical analysis

The Bland-Altman method [129] was used to determine the level of agreement between observers' estimates of maximal polyp diameter and the reference endoscopic diameter. This analysis assumes that the mean of the two corresponding polyp measurements is the best estimate of the true measurement. The overall mean difference is determined and +/- 1.96 standard deviations calculated around this to provide 95% limits of agreement. These limits define the range within which the true measurement will lie on 95% of occasions. Also, because each observer measured the same polyps a paired t-test was used to determine if there was a significant difference between each set of measurements and the endoscopic reference size (for each observer). Paired t-tests were also used to determine any significant difference between estimates from the novice observer and the experts when using the three different visualization displays. A repeated measures ANOVA was used to examine the difference in measurements between the four observers. We allowed for the fact that there were repeated measurements on the same polyp, adjusting for this using the Huynh and Feldt method. Repeated measures analysis of covariance (ANCOVA) was used to determine if the differences between observers varied for polyps of different sizes. This was done by including the endoscopic reference size in the analysis and examining if there was an interaction between this and the observer.

In order to assess the repeatability of the CT polyp measurements, the variability between readers was measured regardless of the reference value by calculating the components of variance. It was possible to break down the data variability into that which was due to 'between' and 'within' polyp measurements. A large proportion of
variability between polyp measurements would imply that the observer had relatively little impact on the measurements. Conversely, if a large proportion of the variability was found to be within polyps, then this would imply that the observer had more influence on the measurement. Within polyp variation was further broken down into that attributable to different observers and that attributable to repeat measurements. The components of variance were calculated using a cross-classified multilevel model with the multilevel component allowing for the structure of the data and the cross-classification allowing for repeat observations on the same polyp. Any missing observations were excluded from the analysis.

6.1.3 RESULTS

Two of the experienced observers felt unable to measure accurately a total of four polyps when using the abdominal window display only: observer 1 was unable to measure one polyp with a reference size of 4mm and observer 2 was unable to measure three polyps with reference sizes of 2mm, 3mm and 5mm. The observers deemed this due to windowing effects i.e. the reassignment of grey scale values according to the change to abdominal windows meant that voxels with lower hounsfield units appeared black and were therefore indistinguishable from intraluminal gas. Two experienced observers (observers 2 and 3) and the novice also felt unable to measure a 2mm polyp accurately when using the 3D display. All other measurements were completed.
Of the 48 polyps measured using a 2D display, observer 1 used an axial display for 27, an oblique display for 15, a coronal display for 5 and a sagittal display for 1. Observer 3 used solely axial views for all polyps and observers 2 and 4 used a combination of axial, coronal and sagittal views.

6.1.3.1 Observer agreement with colonoscopic reference size

The mean differences between observers' estimates of polyp diameter and the colonoscopic reference size, standard deviation of these differences, and limits of agreement for each observer are summarised in Table 14 for each of the three visualisation displays used. In addition, the difference between individual observer measurements and the reference size are displayed graphically as Bland Altman dot plots (Figure 16a, b). These data show that the CT measurements, irrespective of the display used, on average overestimated the diameter of the polyp studied when compared to the reference value. However, least measurement error was encountered when using the 2D abdominal window display. The 95% limits of agreement were relatively wide for all observers, and certainly had sufficient span to encompass different size categories for individual polyps. The narrowest limits of agreement spanned 7.6mm for observer 1 using the 2D colonography window display whereas the widest spanned 14.6mm for observer 3 (an experienced observer) using the 3D display. When significant differences between the observers' estimates and the reference values were considered, observers 1 and 2 were not significantly different, but only when using the 2D abdominal window display; these two observers were significantly different from the reference standard when using the other two alternative displays and observers 3 and 4 were significantly different when using all
three displays (Table 14). The novice (observer 4) had errors that were significantly larger than expert observer 1 for all three displays (p=0.001, 0.01, and <0.001 for 2D colonography, 2D abdominal, and 3D displays respectively); significantly larger than observer 2 for two of the displays (p= 0.04 and 0.001 for 2D colonography and 3D displays respectively). No significant difference was found between the novice and Observer 3 for any display. For all displays, measurement error was smallest, on average, for observer 1, followed by observer 2, with observers 3 and 4 having greater errors. Supporting this data, when the absolute error was calculated for each observation and observer, and compared to the visualisation platform used, there was no significant difference between the 2D colonography window display and 3D display but a significant difference when the 3D display was compared to the 2D abdominal window display in all but one comparison (Table 15).
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0.02</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>0.05</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>0.03</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>0.01</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>0.00</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>

Table 14 Agreement between observer CT measurements and colposcopic reference.

Bland-Altman method, paired t-test.
Table 15.
Comparison of absolute errors between visualisation displays for four observers.

<table>
<thead>
<tr>
<th>Observer</th>
<th>Mean Difference (95% CI) (3D -2D colon display) mm</th>
<th>Significance (p)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.04 (-0.34, 0.43)</td>
<td>0.83</td>
</tr>
<tr>
<td>2</td>
<td>0.00 (-0.51, 0.51)</td>
<td>1.00</td>
</tr>
<tr>
<td>3</td>
<td>0.68 (-0.12, 1.37)</td>
<td>0.05</td>
</tr>
<tr>
<td>4</td>
<td>0.23 (-0.20, 0.67)</td>
<td>0.29</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Observer</th>
<th>Mean Difference (95% CI) (3D -2D abdominal display) mm</th>
<th>Significance (p)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.49 (0.05, 0.93)</td>
<td>0.03</td>
</tr>
<tr>
<td>2</td>
<td>0.38 (-0.19, 0.97)</td>
<td>0.19</td>
</tr>
<tr>
<td>3</td>
<td>1.40 (0.57, 2.23)</td>
<td>0.001</td>
</tr>
<tr>
<td>4</td>
<td>1.09 (0.52, 1.65)</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>
Figure 16 A; Effect of visualisation display on polyp diameter measurement. Bland Altman dot plot comparing CT measurements with colonoscopy reference for each observer (1 to 4) according to visualisation display.
B; Interobserver agreement. Bland Altman dot plot comparing CT measurements with colonoscopy reference for each observer (1 to 4) according to visualisation display.
6.1.3.2 Observer agreement irrespective of reference size

Between-polyp variance (variability when measuring different polyps) and within-polyp variance (variability when measuring the same polyp, further broken down into that attributable to different observers and that due to random variation - ‘between measurement variance’) are shown in Table 16. These data show that measurements obtained using the three visualisation displays are relatively similar with very little within-polyp variation due to differences between observers; approximately 70% of the total variability in the data was attributable to between-polyp differences. Of the within-polyp variation, the majority was due to random variability of the measurements rather than differences between observers. When the differences between observers’ measurements were examined using ANOVA, the most significant difference between them occurred when using the 3D display (Table 17). There was no evidence that this difference was significantly influenced by polyp size [2D colon window display (p=0.96); 2D abdominal window display (p=0.84); 3D display (p=0.86), ANCOVA].
### Table 16.

Variability in measurement data: Components of variance (units = mm)

<table>
<thead>
<tr>
<th>Display</th>
<th>Between polyp variance</th>
<th>Within polyp variance</th>
<th>% Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Between observer variance</td>
<td>Between measurement variance</td>
<td>Between polyps</td>
</tr>
<tr>
<td>2D'colo'n'</td>
<td>8.13</td>
<td>0.10</td>
<td>3.03</td>
</tr>
<tr>
<td>2D'abd'o'</td>
<td>7.23</td>
<td>0.13</td>
<td>2.80</td>
</tr>
<tr>
<td>3D</td>
<td>10.04</td>
<td>0.36</td>
<td>4.34</td>
</tr>
<tr>
<td>p</td>
<td>Mean (SD) Observer 1</td>
<td>Mean (SD) Observer 2</td>
<td>Mean (SD) Observer 3</td>
</tr>
<tr>
<td>-------</td>
<td>----------------------</td>
<td>----------------------</td>
<td>----------------------</td>
</tr>
<tr>
<td>0.001</td>
<td>9.00 (4.52)</td>
<td>8.40 (3.37)</td>
<td>7.33 (3.53)</td>
</tr>
<tr>
<td>0.02</td>
<td>7.21 (3.69)</td>
<td>8.21 (3.70)</td>
<td>7.92 (3.67)</td>
</tr>
<tr>
<td>0.05</td>
<td>6.15 (3.10)</td>
<td>6.30 (3.15)</td>
<td>5.68 (2.93)</td>
</tr>
</tbody>
</table>

ANOVA: Differences in polyp measurements between observers (units = mm).

Table 17.
Predictably, assessment of intra-observer agreement revealed that this was superior to inter-observer agreement, with, in general, the narrowest limits of agreement for the 2D abdominal display and the widest for the 3D display (Table 18). Once again, differences between the two sets of observer measurements are displayed graphically as Bland Altman dot plots (Figure 17 a,b).
Table 18.

Intra-observer agreement.

<table>
<thead>
<tr>
<th>Visualisation display</th>
<th>Observer</th>
<th>Mean difference (first - second)</th>
<th>SD difference</th>
<th>95% limits of agreement*</th>
</tr>
</thead>
<tbody>
<tr>
<td>2D 'Colon'</td>
<td>1</td>
<td>0.10</td>
<td>1.29</td>
<td>(-2.43, 2.63)</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>-0.30</td>
<td>1.49</td>
<td>(-3.22, 2.62)</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.00</td>
<td>1.33</td>
<td>(-2.61, 2.61)</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.78</td>
<td>1.99</td>
<td>(-3.12, 4.68)</td>
</tr>
<tr>
<td>2D 'Abdo'</td>
<td>1</td>
<td>0.10</td>
<td>1.20</td>
<td>(-2.25, 2.45)</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>-0.56</td>
<td>1.13</td>
<td>(-2.77, 1.65)</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.10</td>
<td>1.66</td>
<td>(-3.15, 3.35)</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.11</td>
<td>1.17</td>
<td>(-2.18, 2.40)</td>
</tr>
<tr>
<td>3D</td>
<td>1</td>
<td>-0.50</td>
<td>1.08</td>
<td>(-2.62, 1.62)</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>-1.30</td>
<td>3.50</td>
<td>(-8.16, 5.56)</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>-1.00</td>
<td>1.83</td>
<td>(-4.59, 2.59)</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>-0.89</td>
<td>1.27</td>
<td>(-3.38, 1.60)</td>
</tr>
</tbody>
</table>

*Bland-Altman limits of agreement
Figure 17. A; Effect of visualisation display on intraobserver agreement. Bland Altman dot plot comparing CT measurements with colonoscopy reference for each observer (1 to 4) according to visualisation display.
Figure 17 B: Intraobserver agreement. Bland Altman dot plot comparing CT measurements with colonoscopy reference for each observer (1 to 4) according to visualisation display.
6.1.4 DISCUSSION

Measurement error during colonoscopy is well established [130-132] and we have shown that error also occurs during CT colonography. This error is dependent upon the observer, their experience, and the viewing conditions used to make the measurement. Furthermore, we used a viewing platform from a single manufacturer. It has been shown that lesion conspicuity may be affected by the type of workstation [51] and so it is conceivable that further differences between measurements will occur using platforms from different manufacturers.

Researchers have tended to group polyps into three size categories contingent upon their maximum diameter (small, 5mm or less; medium, 6 to 9mm; large, 10mm or larger) and expert consensus has stated that polyps in the 'small' category are clinically insignificant whereas polypectomy is recommended for those that are 'large' [70]. Such grouping is convenient but measurement error is inevitable and could potentially result in large polyps being inadvertently classified as small and vice-versa, especially where the error occurs around a category cut-point. Supporting this, we found limits of agreement as wide as 14mm, raising the possibility that a polyp whose true diameter was 10mm + could be assigned to the 'small' category as a result of measurement error and thus be either assigned to interval surveillance or even disregarded when the correct course of action is endoscopic polypectomy. Conversely, by including 'clinically insignificant' polyps (2 to 5mm polyps inclusive) in our dataset, we showed that diminutive polyps could be potentially miscategorised as large and therefore inappropriately referred for polypectomy.
Our analysis of variance found that there was some evidence of a significant difference between mean polyp measurements for the four observers especially when the 3D display was used to make the measurement. This might be expected since there are more subjective opportunities for cursor placement using this type of display. The angle and direction from which the polyp is viewed may be continuously varied by the observer, as may be the distance from the polyp, even though the software is calibrated to account for this. Intuitively, these factors suggest that inaccurate measurements are most likely using a 3D display and our results support this hypothesis. Alternatively, it could be argued that a 3D visualisation offers the best chance of detecting the long axis of the polyp, especially when the polyp is large and its morphology irregular. However, we found no evidence that accuracy was influenced by the size of the polyp being measured.

Predictably, intra-observer agreement was better than inter-observer agreement but again the 3D display fared worst in this respect. This is important because polyps left in situ will require re-examination and re-measurement, perhaps after an interval of several years, which makes it less likely that the subsequent measurement will be performed by the same observer who made the first. The clear message is that while 3D visualisation may be the best method with which to detect polyps [7], it is not the most reliable method with which to measure them.

Whether to use colonography or abdominal window displays for measurement is open to debate from examination of our data. Both displays generally over-estimated the diameter of polyps when compared to the reference measurement, but, overall, this
discrepancy was least for the abdominal window display. However, the narrowest limits of agreement were achieved by observer 1 (an experienced observer) when using the colonography display and the least inter-observer error overall (irrespective of agreement with the reference values) was achieved using the colonography display. Also, two observers reported that very small polyps 'disappeared' when they enabled the abdominal window display. Intuitively, this latter finding suggests abdominal windows should not be used when there is an appreciable reduction in polyp conspicuity by utilising these settings. Moreover, in vitro studies of lung nodule measurement have shown far more accurate measurements are obtained using lung window settings (closer to the colonography display) compared to using an 'abdominal window' display [133]. There was also some evidence that observers differed in their ability to make accurate measurements. Notably, two experienced observers (observers 1 and 2) were generally more accurate than observers 3 and 4 (an experienced observer and a novice); the novice observer had significantly larger errors than two of the experienced observers but not the third. Although this suggests that prior experience in CT colonography interpretation may facilitate more accurate measurements, this study is not absolutely conclusive.

There are several limitations to our study and perhaps it raises more questions than it answers, especially with respect to what is the true diameter of a polyp. We used the colonoscopic estimate of polyp diameter as a reference standard since this is conventional practice and accordingly well established. We also used a single expert endoscopist whose experience exceeded 5000 colonoscopies. Despite this, as described above, it is well recognised that colonoscopic measurements are subject to considerable error, even when made by experienced practitioners. Endoscopists tend
to overestimate polyp diameter during in vivo assessment and underestimate diameter during in vitro assessment [134;135]. For pragmatic reasons we used comparison with open biopsy forceps to determine the reference size since this is the method preferred by our endoscopists. Utilising a calibrated linear measuring tool may have been preferable but even this is imperfect because it may not be possible to align the tool adjacent to the maximal diameter of the polyp, leading to a semi-subjective assessment [136]. Alternatively, it could be argued that we should have used pathological estimates as our reference standard as they have been shown to be more reliable than endoscopic measurements [137]. However, even this approach is imperfect since excised polyps tend to decrease significantly in diameter due to cauterisation, vascular collapse, and may either enlarge or shrink with formalin fixation [138;139]. We did not attempt to analyse our data according to polyp morphology for example, our dataset contained no flat polyps, which are notoriously difficult to detect and, by inference, to measure [140;141]. Also, it is conceivable that elliptical polyps may be more accurately measured on 3D due to better depiction of their shape. Indeed, our study provides some indirect evidence to suggest this is the case. While observers were free to use oblique 2D reformatted images, only observer one actually did so, and they were closest to the reference standard. In contrast, observer three solely used the axial plane and was the least accurate of the experts.

All of these factors conspire to thwart a truly reliable estimate of polyp diameter with which comparisons can be made. The Bland-Altman approach is especially relevant to this type of analysis because it can easily be argued that the 'true' measurement is not known with certainty from either colonography or colonoscopy. In this case, the best estimate of the true diameter is the mean of the measurements obtained by the
colonoscopy and colonography, which is the bedrock of the Bland-Altman analysis. We found that absolute agreement between the CT colonography estimate and the colonoscopic reference measurement was variable, evidenced by limits of agreement that were frequently enough to span not only one but two polyp size categories.

We used a single viewing platform (albeit with a variety of visualization displays). Most obviously, it could be argued that the 3D display we used, which was surface-rendered, may be more inherently inaccurate than a volume-rendered alternative [142]. Also for pragmatic purposes, we did not formally assess the effect of display magnification on polyp measurement accuracy although observers were free to alter magnification according to individual preference. Further research on these topics is required. Our use of a single novice may be criticised since this individual was effectively acting as a proxy for all novices. We were able to demonstrate significant differences between our three experienced observers and no doubt there would also be differences between novices whose inherent ability to make these types of measurement will inevitably vary. One potential solution to minimize inter- and intra-observer variation of assessment of polyp diameter is to fully automate the measurement process using computer-aided boundary detection, which removes the subjective element of cursor placement. Further development in this area is awaited.

In conclusion, measurement of maximum polyp diameter during CT colonography is subject to inter- and intra-observer variation, the degree of which is contingent on the observer, their experience, and the viewing conditions used. It is important to be aware that this variation may result in polyps detected by CT being inadvertently assigned to an incorrect size category. In day-to-day clinical practice, the window
display and anatomical plane used for measurement should be documented to facilitate any subsequent interval comparisons. Finally, our study suggests that 3D visualisation displays, commonly used for polyp detection, should not be used for polyp measurement.
6.2 Polyp measurement and size categorisation by CT colonography: Effect of observer experience in a multi-centre setting.

6.2.1 INTRODUCTION

As described in the preceding study (Chapter section 6.1), the decision as to which polyps should be removed and which can be left safely in situ is determined largely by maximal polyp diameter since this predicts the risk of malignancy. As a consequence, a working group has developed patient management strategies based on the largest polyp found at screening CT colonography [70] but it is self-evident that these strategies turn on obtaining accurate estimates of diameter. However, single centre data from this thesis (chapter 6.1) and others has suggested that significant measurement error exists, determined partly by observer experience and the conditions used to display the data during measurement [143]. Measurement error will have variable impact, contingent upon the clinical circumstances. For example, while error of a few mm either way will not affect categorisation of polyps measuring 15mm or more, a similar error for polyps lying close to a category threshold will ultimately determine patient management.

We aimed to investigate measurement error for observers of different experience from different centres by comparison with the colonoscopic reference size. We also aimed to determine to what extent, if any, measurement error influenced subsequent polyp categorisation according to established management guidelines [70].
6.2.2 MATERIALS AND METHODS

We utilised data acquired as part of the reader performance study described in Chapter section 4.1 in which three observer groups were asked to read an individualised subset of 40 cases from the study dataset of 51 cases (8 cancer cases, 16 polyp cases and 27 normal cases). Of the 16 cases with a polyp, 12 were large and 4 were medium sized. The segmental locations of these was as follows; two rectal; four sigmoid; two descending colon; five transverse colon; two ascending colon; one caecal. All lesions identified had been verified by same day optical colonoscopy, at which time the reference diameter was established by comparison with adjacent open biopsy forceps.

To recap, observers had been asked to review each case and to measure the maximum perceived diameter of any polyp identified, so as to facilitate classification into 'large' and 'small' groups. Measurements were obtained using software calipers applied to the 2-D axial or multi-planar-reformat that the observer considered best represented the maximal diameter of the polyp. The software did not permit measurement when using the 3D volume rendered display. Each observer recorded both the image numbers (prone and supine positions) and segmental location of any polyp found. After collating all responses, the principal investigator of this thesis tabulated Individual measurements along with the colonoscopic size and diagnostic category assigned to each case by the submitting centre. The known scan co-ordinates were compared with observers' responses in order to determine whether their responses were true- or false-positive.
6.2.2.1 Statistical analysis

As in the previous study (Chapter section 6.1) the Bland-Altman method [129] was used to determine the level of agreement between observers' estimates of maximal polyp diameter and the colonoscopic diameter. The absolute difference between the observer estimate and the reference size was calculated (ignoring the direction of the difference) and the Kruskal-Wallis test statistic used to examine differences between observer groups. We also examined the effect of measurement error on categorisation and whether the observer managed to detect the polyp or not. These categorical outcomes were compared using the Chi-squared test statistic. Statistical significance was ascribed to a probability value of 5%.

6.2.3 RESULTS

There were 28 observers in total: 9 experienced radiologists, 9 less experienced radiologists, 10 technicians. Of the 16 cases with polyps, 3 observers read all 16; 4 observers read 15; 9 observers read 14; 2 observers read 13; 8 observers read 12; 1 observer read 11; 1 observer read 10.

6.2.3.1 Agreement with colonoscopy

The mean differences between observers' estimates of maximal polyp diameter and the colonoscopic reference size, standard deviation of the differences, and limits of agreement for each of the three observer groups are summarised in Table 19. In addition, the difference between individual observer measurements and the reference size of each polyp, for each observer group are displayed graphically as Bland
Altman agreement plots (Figure 1a-c). These data show that the CT measurements on average underestimated the diameter of the polyp studied when compared to the colonoscopic reference value, by approximately 2 to 3 mm, irrespective of the observer group. The 95% limits of agreement were relatively wide for all observer groups, and had sufficient span to encompass different size categories for individual polyps. However, the narrowest limits of agreement were achieved by the experienced radiologists (Table 19).
Table 19.
Agreement between measurement of polyp diameter by CT colonography and the colonoscopic reference diameter for each observer group.

<table>
<thead>
<tr>
<th>Observer Group</th>
<th>Mean difference (CT minus colonoscopy) mm</th>
<th>SD difference Mm</th>
<th>95% Bland-Altman limits of agreement mm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experienced</td>
<td>-2.1</td>
<td>3.7</td>
<td>-9.4 to 5.2</td>
</tr>
<tr>
<td>Radiologist</td>
<td>-2.8</td>
<td>5.4</td>
<td>-13.3 to 7.7</td>
</tr>
<tr>
<td>Technician</td>
<td>-2.2</td>
<td>4.5</td>
<td>-11.1 to 6.6</td>
</tr>
</tbody>
</table>

SD = Standard deviation
There was no significant difference between observer groups when the difference between the estimated and reference diameters were considered (p=0.39): all groups had a median difference of 2mm. However, the inter-quartile range (IQR) for experienced radiologists (0 to 4mm) indicated a spread of measurements that was closer to the reference values than for the less experienced radiologists (IQR 1mm to 6mm) or radiographic technicians (IQR 1mm to 5mm).
Figure 18a-c

Bland-Altman agreement plots for each of the three observer groups.

a. Experienced radiologists  
b. Less experienced radiologists  
c. Radiographic technicians. The middle solid horizontal line represents the mean difference between the pairs of measurements (difference between observer and colonoscopy estimate); essentially representing the mean 'measurement error'. The upper and lower horizontal lines represent the 95% limits of agreement i.e. the range of 'measurement errors' that would occur on 95% of occasions.

18a. Experienced
18b. Less experienced radiologists

18c. Technicians
6.2.3.2 Polyp detection and case categorisation

There were 16 cases with polyps; 4 medium and 12 large. The frequency with which these polyps were correctly identified by the three different observer groups is shown in Table 20. On average, larger polyps were detected more frequently than medium sized polyps. Per-polyp detection rates for medium polyps ranged from 13% to 88% for experienced radiologists, 0% to 71% for less experienced radiologists, and 0% to 88% for radiographic technicians (Table 20). Per-polyp detection rates for large polyps ranged from 0% to 100% for all three observer categories (Table 20). When per-polyp detection rates were compared overall, experienced radiologists detected significantly more polyps than either of the other two groups (p=0.01).
Table 20. Frequency of individual polyp detection and correct classification by the three observer groups.
There were 167 polyp observations by observers overall; 134 (80%) of these were of
large polyps according to the reference colonoscopy diameter and 33 (20%) were of
medium polyps. The majority of polyps that were detected by observers were
correctly categorized (135 of 167 (81%) of polyps) with respect to their colonoscopic
reference diameter. Experienced radiologists correctly categorized 49 of 55 (89%)
large polyps observed compared to 30 of 42 (71%) for less experienced radiologists
and 28 of 37 (76%) for radiographic technicians. Experienced radiologists correctly
categorized 11 of 15 (73%) medium polyps observed compared to all 9 of 9 (100%)
for less experienced radiologists and 8 of 9 (89%) for radiographic technicians. There
was no significant difference between observer groups when the frequency of correct
case categorization was compared (p=0.28).

Of the 32 polyp observations that were miscategorised, only 5 (16%) of these were
due overestimation of diameter; 5 observations of a single polyp known to be medium
by the reference standard (7mm) were miscategorized by observers as large (Fig
2a,b). The remaining 27 (84%) miscategorisations referred to large polyps that were
misclassified by observers as medium. Thus, overall 27 of 134 (20%) observations of
large polyps resulted in their misclassification as medium; experienced radiologists
misclassified large polyps on six occasions (due to 3 observer estimates of 9mm and
3 observer estimates of 8mm); less experienced radiologists misclassified large
polyps on twelve occasions (due to 7 observer estimates of 9mm, 4 observer
estimates of 8mm and 1 observer estimate of 7mm); radiographic technicians
misclassified large polyps on nine occasions (due to 6 observer estimates of 9mm, 2 observer estimates of 8mm and 1 observer estimate of 7mm).

Therefore, of the 27 observations that misclassified large polyps as medium, 16 (60%) were within 1mm of the 'large polyp' threshold (i.e. 10mm); 9 (33%) measurements were within 2mm; and 2 (7%) were within 3mm. These 27 observations referred to six large polyps: four of these were misclassified by several observers. (one polyp was misclassified 9 times, Fig 3a,b); and a single polyp of 30mm was misclassified by a single less experienced radiologist observer, who had probably appreciated part of the lesion only.
**Figure 19.** 7mm polyp located in the rectosigmoid colon and displayed on the supine axial image (Fig 2a. ‘colon’ window: W1500, L -150) and 3D volume rendered ‘endoluminal’ display (Fig 2b). Polyp overestimated as large by 5 observers: the line across the polyp (Fig 2a) represents software calipers placed by the author corresponding to a CT diameter of 7mm.

Fig 19a

![Image 19a](image19a.png)

Fig 19b

![Image 19b](image19b.png)
Figure 20 10mm polyp located in the rectum and displayed on the prone axial image (Fig 3a. 'colon' window: W1500, L -150) and 3D volume rendered 'endoluminal' display (Fig 3b). Polyp underestimated as medium by 9 observers: the line across the polyp (Fig 3a) represents software calipers placed by the author corresponding to a CT diameter of 10mm.

Fig 20a

Fig 20b
6.2.4 DISCUSSION

Polyp management strategies are a sensible step towards standardisation of CT colonography reporting and the avoidance of unnecessary colonoscopy in patients whose polyps are diminutive. However, this study has shown wide variability amongst observers of different experience and from different centres when asked to measure the same polyp. Perhaps most importantly, we found that large polyps were misclassified as unimportant lesions in a proportion of cases: 20% of observations placed polyps that were considered large by colonoscopy into a smaller category. This observation has clear implications for patient management strategies that are defined by maximal polyp diameter, raising the possibility that individuals with an advanced adenoma could be returned to surveillance alone. At the very least, it suggests caution for those polyps whose diameter when estimated by CT colonography places them close to the 1cm category boundary, which separates 'indeterminate' from 'possibly advanced' lesions [70].

Our findings highlight the difficulty of using a single discrete threshold to define an advanced adenoma. This study and the study described earlier in this chapter, emphasise the necessity of an 'indeterminate' category lying immediately below this, since it is inevitable that some polyps will be miscategorised. Given this, it has been sensibly suggested that a colonographic diameter of 8mm might be a more reasonable threshold to trigger colonoscopy [7]. In our study, a threshold of 8mm would have meant correct categorisation of 99% of those polyps with a reference diameter of 1cm or more; a threshold of 7mm would have meant correct categorization of all such polyps. The single polyp of 30mm that was misclassified by
one less experienced radiologist (who probably appreciated only part of the lesion),
reinforces the need for full characterisation of any abnormality, in order to accurately
depict its true extent.

We found that colonographic diameters were generally less than the reference
measurement. However, there is sufficient uncertainty around the accuracy of
colonoscopic estimates of diameter for this area to require considerable further
research. For example, as described earlier in this chapter, several studies have
found that colonoscopy tends to overestimate polyp size in vivo, especially when the
measurement is made via comparison with adjacent opened biopsy forceps [144].

There are also particular difficulties with larger polyps [145]. It is therefore possible
that CT colonography provides a more accurate assessment of true size. There are
good theoretical reasons to support this: CT colonography provides multi-planar
views that allow a fuller appreciation of the extent of a lesion, without the perspective
distortion suffered by endoscopy. In reality, the mean measurement error for both
modalities is of the order of a few mm, and it has been suggested by colonoscopists
that arguing about this is tantamount to, 'making a mountain out of a molehill' [146].

However, errors of such magnitude assume greater significance for CT colonography
since peri-procedural polypectomy is not possible.

The 95% limits of agreement for each observer group were relatively wide when
compared to the mean difference between the colonographic estimate and reference
diameter. The result is that, for an individual patient, there is the potential for an
observer to make errors of several mm in either direction. Experienced radiologists
had the narrowest limits of agreement with colonoscopy, suggesting that experience
may facilitate more accurate measurement although we could find no statistically significant difference when the median measurement was considered overall. However, it should be noted that looking for a significant difference between mean or median measurements is a poor indicator of how well two tests agrees with each another [147]. It is also possible that polyps that are more difficult to detect are also more difficult to measure accurately, which might disadvantage the experienced observers who, as might be expected [14], detected more of these polyps than other groups. In contrast, it is interesting to note that there is some evidence to suggest that prior experience of colonoscopy is not associated with more accurate measurement [148]. Interestingly, that study also showed that directed training improved measurement accuracy, and that this had most effect on the least experienced endoscopists [148]. Equivalent training modules for CT colonography may also be possible.

There are several limitations to this study. We have already discussed above the potential problems associated with using a colonoscopic estimate as the reference measurement. It is known that endoscopists' estimations of diameter vary, and several endoscopists provided the reference measurements for the present study. However, while an in-vitro phantom-based approach would have allowed the true diameter to be ascertained with more certainty, we wished to study measurement error in a situation more akin to day-to-day clinical practice. All observers used the same software for measurement in order to eliminate confounding due to differing platforms. This did not allow measurement from the 3D endoluminal view, and so we could not assess error in this context. Also, it could be argued that 3D displays allow the observer to more fully depict a polyp's true boundary, especially when the shape
is irregular. There is conflicting evidence suggesting that 3D measurements are variously more accurate [149] or less accurate (see chapter 6.1), and this topic deserves further investigation.

In contrast to the present study, the findings of the previous study, described earlier in this chapter (Chapter section 6.1), found that CT colonography may generally overestimate polyp diameter, which raises the possibility that different software platforms may provide different measurements of the same polyp, for example due to different calibration methods, displays, or software caliper functionality. Clearly, more research in this area is also needed. Further limitations include a disproportionate number of 'large polyps', which was a result of protocol stipulations that intentionally introduced a spectrum bias towards these. However, while potential miscategorisation of small polyps is important, we considered miscategorisation of large polyps to be most important from a clinical perspective.

In conclusion, for the single software platform utilized in this study, observer estimates of polyp diameter made from CT colonography tend to be less than the colonoscopic reference diameter. Caution should be exercised in those polyps whose colonographic diameter is below but close to the 1cm boundary threshold in order to avoid miscategorisation of an advanced adenoma.
6.3 CT colonography: Automated diameter and volume measurement of colonic polyps compared with a manual technique – in vitro study

6.3.1 INTRODUCTION

As described in the preceding chapters (Chapter sections 6.1 and 6.2), manual polyp measurement is prone to human error, potentially leading to patient mismanagement. To minimise human error, several computer software companies have been investigating the use of automated measurement alongside the development of computer aided diagnosis (CAD) software (the prime aim of which is to detect and then highlight polyps to observers).

There has been considerable interest in automated analysis of CT colonography datasets because interpretation is time consuming and fatiguing, and, as has been shown in this thesis, is prone to error because of this. Such software aims to provide the observer with a list of polyp candidates (displayed as visual prompts), which the observer can then interrogate and either confirm or reject as true or false positive findings. It is hoped that such CAD systems will improve polyp sensitivity while simultaneously reducing interpretation times.
6.3.1.1 *How does Colon CAD work?*

The CAD software generally performs several steps as follows [150]:

- **Extraction**: CAD software extracts the colon from its surroundings by density slicing (thresholding).

- **Analysis**: A mathematical algorithm is applied to the extracted surface that analyses the contour of the colonic wall and classifies components of the wall into either polypoid (e.g. colonic polyps) or non-polypoid (e.g. normal wall and haustral folds) regions.

- **Classification**: Additional mathematical classifiers such as tissue Xray-density and estimations of object sphericity are used to help distinguish true polyps from false-positive detections.

- **Output**: CAD analysis results in an output of 'CAD marks' (i.e. annotations or visual prompts) of polyp candidates which are then systematically reviewed and subsequently characterised by the reader into those representing either a true polyp or a false positive prompt. It should be noted that most prompts will be false-positive, irrespective of the system used.
Colon CAD may be offered as a stand alone product or integrated into an existing reading platform as an additional functionality. There are currently three potential reader paradigms incorporating CAD [151]; CAD as a first reader, whereby the CAD output alone is reviewed by the reader and candidates are accepted or rejected; CAD as a 'second reader' whereby the reader first interprets the CT colonography examination in the usual way. The CAD software is then switched on and the reader examines the CAD output to first identify and subsequently characterise any further polyp candidates. The third paradigm uses CAD as a 'Joint' or 'concurrent reader' whereby the reader interprets the CT colonography examination alongside the highlighted CAD output. At the time of writing, regulatory approval has only been gained for a second-reader paradigm.

6.3.1.2 CAD performance

A study from 2002 showed the potential benefit of using CAD software in a 'second reader' paradigm increasing large polyp (10mm or larger) sensitivity from 67 to 89% [152]. More recently however a second group has shown standalone CAD sensitivity (without further case interrogation) for detecting polyps 6mm or larger, of 81%, exceeding the average sensitivity of 3 experts [153]. Also, in this study CAD detected 19 of the 21 (90%) polyps missed by the individual experts [153]. There is currently considerable research into different software products with recently abstracted work showing CAD software associated with; per polyp sensitivity (polyps 10mm or larger) of up to 95% [154]; per patient sensitivity of 96% including patients with small and diminutive polyps [155]; improved polyp sensitivity from 74 to 85% when used by inexperienced radiologists [156]. However despite this enthusiasm, there is evidence
emerging that CAD should be combined with dedicated reader training to optimise CT colonography performance [157].

Ultimately the goal is for CAD software to identify all significant polyps and cancer with very few false positive detections, at which point population screening with CT colonography becomes a much more realistic prospect.

6.3.1.3 Polyp measurement

As described above, most CAD work to date has concentrated on automated polyp detection and automated measurement of polyp size via segmentation has been relatively ignored [158]. However, supported by the studies earlier in this chapter, detection is not the only goal: polyps that have been identified must be measured accurately in order to estimate their biological risk of progression to cancer and CAD may help make this process more accurate.

There are several components to this seemingly simple task. Not only must the measurement approximate closely to the real diameter of the polyp, it must also be both reproducible and repeatable since, in the context of non-invasive screening, small polyps may be left in situ and undergo serial examination at intervals contingent upon the size of the polyp and age of the patient. Reproducibility can be defined as the value below which two single test results obtained under different conditions (for example, different observers) may be expected to lie with a specified probability [159]. Repeatability refers to the strength of agreement between repeated measurements obtained under similar circumstances [159]. If an automated polyp measurement
technique could be shown to be accurate, repeatable, and reproducible, it would have significant clinical utility.

The principal investigator of this thesis and co-workers have developed novel software in collaboration with Medicsight plc (Berkeley Square, London, UK), the aim of which is to automate the measurement process with the hope of reducing measurement error due to observer variability. This study assessed the performance of this software in vitro, using two colonic phantoms.

6.3.2 MATERIALS AND METHODS

6.3.2.1 Phantom construction

Two colonic phantoms were used. The first phantom (QRM, M"ohrendorf, Germany) was originally designed for analysis of lung nodules and contained 40 synthetic polyurethane spheres, ranging in size from 3mm to 10mm in diameter. These spheres had a density of 35 Hounsfield Units when scanned at 120kV. The phantom cavity was normally filled with cork granules to mimic lung parenchyma but these were removed for the purposes of the present study in order to better simulate the gas distended colonic lumen. Also in order to better simulate mural colonic polyps, only the 16 nodules related to the ‘chest wall’ of the phantom were used. The morphology of these was either ‘sessile’ or ‘pedunculated’. There were eight sessile polyps ranging in diameter from 3mm to 10mm in one mm increments, and eight pedunculated polyps of identical diameters. The volume of the polyps ranged from 7.3 mm$^3$ for the 3mm sessile polyp to 520.4 mm$^3$ for the 10mm pedunculated polyp. The
diameter and volume of each polyp was supplied by the manufacturer, with each accurate to one-hundredth of a mm and one hundredth of a mm$^3$ respectively.

The second phantom was made of polymethyl methacrylate and has been described previously, the 'Whiting' phantom [160]. In brief, this phantom consisted of three acrylic cylinders, each containing acrylic spheres that were designed to simulate sessile and pedunculated polyps. Only the largest cylinder, with an internal diameter of 50.8 mm was used for the present analysis. This contained two sets of six polyps with a maximal diameter ranging from 3 mm to 13 mm, two of which were polypoid (maximal diameters of 6 mm and 13 mm) and four of which were sessile (with maximal diameters of 3 mm, 6 mm and two of 11 mm; the latter two had different heights of 1 mm and 3 mm respectively) [160]. The exact volumes of polyps in this phantom were not known. Two identical rows of polyps were arranged 35 mm apart and separated by 90 degrees so that the two rows were arranged at the four o’clock and seven o’clock orientations [160].

6.3.2.2 Scanning parameters

The QRM phantom was scanned using a 16-detector row CT scanner (Lightspeed, GE Medical Systems, Slough, UK) using the following parameters: 300 mA, 120 kV, 0.5 s rotation, 1.375 pitch, 27.5 cm$^{-1}$ table speed, 1.25 mm slice collimation, reconstruction on standard algorithm (Figure 21). The Whiting phantom was scanned using a 16-detector row CT scanner (Sensation 16, Siemens Medical, Iselin, New Jersey) using the following parameters: 100 mAs (0.5 s rotation time), 120 kV, 0.5 s rotation, 1.50 beam pitch, 1.5 mm collimation, 1.0 mm reconstruction interval, reconstruction on standard algorithm (Figure 22). The QRM phantom was scanned
parallel to the z-axis of the scanner while the Whiting phantom was scanned at 45 degrees to the z-axis.
Figure 21

Representative image from the QRM lung phantom with cork granules removed from the phantom cavity. Only the nodules related to the 'chest wall' were measured as these best mimicked colonic polyps: There are both sessile and pedunculated types.
Figure 22

Representative image from the Whiting phantom showing mural polyps of both polypoid and sessile morphology.
6.3.2.3 Image analysis

Two observers, a radiologist and a radiographic technician, independently analysed both phantom datasets using a personal computer and proprietary software that was under development (MedicColon version 1.1, Medicsight PLC, London, UK). Both were trained in the use of this software prior to their analysis. Neither read CT colonography datasets as part of their day-to-day clinical practice.

Each observer identified individual polyps by scrolling through the 2-dimensional axial views of each phantom, using a window width and level of 1500 and -150 respectively. When a polyp was encountered, the individual image that was felt to best represent its maximal diameter was identified and three measurements obtained. Firstly, the maximal polyp diameter was estimated from this image using software electronic callipers, the placement of which was entirely under the subjective control of the observer (the 'manual' method). The observer then used a manual software drawing tool to outline the perceived polyp boundary on every individual contiguous image on which each polyp could be identified. This boundary was drawn freehand using a mouse, and was again placed subjectively by the observer. An automated software segmentation based on this boundary was initiated and provided estimates of the polyp diameter and volume (the 'semi-automatic' method). The observer then placed two software seed points opposite one-another at their perceived junction between the polyp and the phantom wall on the image that best represented the maximal polyp diameter. They then initiated a further software automatic segmentation of the polyp candidate. This second segmentation (the 'fully automatic' method) provided an automatic assessment of the diameter and volume of each
polyp candidate without the need for manual boundary identification (Figure 23). In order to achieve this, the software automatically repositioned the subjectively placed cursors at the best local candidate for the polyp boundary in the immediate vicinity and followed this with automatic segmentation that was based on a region-growing scheme. This scheme used fuzzy logic to separate the polyp from the surrounding normal structures. The result of this process was then mapped to contiguous slices, where two new seed points were initialised. This procedure was repeated until the entire polyp has been analysed. Any failure to automatically segment the polyp was also recorded. Observers were free to magnify the image as they wished and also to use coronal or sagittal multi-planar reformats. Readings were performed during a single session and each observer analysed each dataset twice to provide information on intra-observer variation.
Figure 23

Workstation screenshot showing a window for automated volume and diameter measurement of a 13mm pedunculated polyp.
The observer copied the results displayed on the monitor onto a study sheet. Observers were unaware of each other's results. Neither observer was aware of the size of the polyps present in the Whiting phantom. The radiologist observer was aware of the size of the polyps in the QRM phantom whereas the radiographic technician was not. The results were collated by the principal investigator of this thesis for subsequent analysis.

6.3.2.4 Statistical Methods

For inter- and intra-observer comparisons of interest there were two measurements of the same polyp. The level of agreement between these was examined by calculating the Bland-Altman limits of agreement that measures, in real terms, the size of the range of differences between the measurements that are likely to occur [129]. The difference between corresponding measurements of the same polyp was determined and 95% limits of agreement calculated as follows: Mean difference +/- 1.96*(standard deviation of differences). Comparisons were made between paired results from each individual observer (to estimate intra-observer variation) and between observers (to estimate inter-observer variation). For the latter comparison, the average of the two values for each polyp from each observer was calculated, and these two pairs of average values were then compared. Finally the results for each observer were compared to the known polyp size. Again, the average value for each observer was used for this comparison. The analyses were performed separately for the two phantom types. The two types of polyp in the QRM phantom (sessile and pedunculated) were analysed together.
In order to investigate the relationship between estimated measurements and the known value, the absolute difference between the estimate and the true value was calculated for each polyp, for each method of measurement. i.e. the size of the difference, regardless of the direction of the difference (the average of the two repeat measurements was used for this analysis). The differences between measurements were then compared between methods using the Wilcoxon matched-pairs test.

6.3.3 RESULTS

The observers were unable to detect the 3mm x 1mm sessile polyp in the Whiting phantom and fully automatic segmentation failed for the smallest sessile polyp in the QRM phantom (3mm), leaving 26 polyps for analysis.

6.3.3.1 Intra-observer agreement

The mean difference, standard deviation of the differences, and Bland Altman 95% limits of agreement for intra-observer comparisons are summarized in Table 21. Concerning repeated measurements of polyp diameter, the limits of agreement spanned less than one mm for the fully automated method. This finding applied to all polyp sizes (the true value of which ranged from 3mm to 10mm in the QRM phantom and 3mm to 13mm in the Whiting phantom) for both observers and both phantoms, indicating very high intra-observer agreement. The limits of agreement for the QRM phantom spanned 0.39mm and 0.48mm for the radiologist and radiographer respectively, and 0.24mm and 0mm for the Whiting phantom (the span is the width of the 95% limits of agreement). The limits of agreement for the semi-automatic estimates of diameter were approximately seven to eight times wider overall than
those for the fully-automatic measurement (span for QRM phantom 3.23mm and 1.34mm, radiologist and radiographer respectively; Whiting phantom 2.74mm and 1.6mm). The limits of agreement for the manual estimates of diameter were wider than either the fully-automatic or semi-automatic measurements, being approximately 10 times wider overall than the fully-automatic measurement (span for QRM phantom 3.57mm and 3.21mm, radiologist and radiographer respectively; Whiting phantom 3.2mm and 2.02mm).
Table 21

Mean difference, standard deviation of the differences, and Bland Altman 95% limits of agreement for intra-observer comparisons of polyp diameter and volume measurements for two phantom types.

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Phantom Type</th>
<th>Observer</th>
<th>Mean difference (mm)</th>
<th>SD of differences</th>
<th>95% Bland-Altman limits of agreement (mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fully automatic diameter</td>
<td>QRM</td>
<td>Radiologist</td>
<td>0.03</td>
<td>0.10</td>
<td>(-0.17, 0.22)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Technician</td>
<td>0.01</td>
<td>0.12</td>
<td>(-0.23, 0.25)</td>
</tr>
<tr>
<td></td>
<td>Whiting</td>
<td>Radiologist</td>
<td>-0.02</td>
<td>0.06</td>
<td>(-0.14, 0.10)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Technician</td>
<td>0.00</td>
<td>0.00</td>
<td>(0.00, 0.00)</td>
</tr>
<tr>
<td>Semi automatic diameter</td>
<td>QRM</td>
<td>Radiologist</td>
<td>0.23</td>
<td>0.82</td>
<td>(-1.39, 1.84)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Technician</td>
<td>0.01</td>
<td>0.34</td>
<td>(-0.66, 0.68)</td>
</tr>
<tr>
<td></td>
<td>Whiting</td>
<td>Radiologist</td>
<td>0.14</td>
<td>0.70</td>
<td>(-1.23, 1.51)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Technician</td>
<td>0.09</td>
<td>0.41</td>
<td>(-0.71, 0.89)</td>
</tr>
<tr>
<td>Manual diameter</td>
<td>QRM</td>
<td>Radiologist</td>
<td>0.09</td>
<td>0.91</td>
<td>(-1.70, 1.87)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Technician</td>
<td>0.26</td>
<td>0.82</td>
<td>(-1.35, 1.86)</td>
</tr>
<tr>
<td></td>
<td>Whiting</td>
<td>Radiologist</td>
<td>-0.56</td>
<td>0.82</td>
<td>(-2.16, 1.04)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Technician</td>
<td>0.39</td>
<td>0.64</td>
<td>(-0.37, 1.65)</td>
</tr>
<tr>
<td>Fully automatic volume</td>
<td>QRM</td>
<td>Radiologist</td>
<td>2.3</td>
<td>6.2</td>
<td>(-9.8, 14.4)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Technician</td>
<td>-0.5</td>
<td>6.2</td>
<td>(-12.5, 11.6)</td>
</tr>
<tr>
<td>Semi automatic volume</td>
<td>QRM</td>
<td>Radiologist</td>
<td>-6.2</td>
<td>25.0</td>
<td>(-55.1, 42.8)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Technician</td>
<td>-7.6</td>
<td>26.3</td>
<td>(-59.1, 43.8)</td>
</tr>
</tbody>
</table>

SD = standard deviation
Concerning repeated measurements of polyp volume (made on the QRM phantom alone, with the true value ranging from 7.3mm$^3$ for the 3mm sessile polyp to 520.4mm$^3$ for the 10mm pedunculated polyp), the limits of agreement were again tightest for both observers when comparing the fully-automated measurement, and were approximately four times smaller than the span obtained using the semi-automatic measurement; span 24.2mm$^3$ and 24.1mm$^3$ for the radiologist and radiographer respectively compared to 97.9mm$^3$ and 102.9mm$^3$ for the semi-automatic measurement.

Overall, there was a general tendency for the radiographer to have narrower limits for intra-observer agreement than the radiologist, but this observation was smallest when using the fully-automatic measurement (Table 21).

6.3.3.2 Inter-observer agreement

The mean difference, standard deviation of the differences, and Bland Altman 95% limits of agreement for inter-observer comparisons are summarized in Table 22. In keeping with the results for intra-observer agreement, the limits of agreement for measurement of diameter were narrowest for the fully-automatic method (span 0.12mm and 0.16mm, QRM and Whiting phantoms respectively) and widest for the manual method (2.87mm, 2.18mm), with the semi-automatic method intermediate (1.11mm, 2.19mm). Overall, the 95% limits of inter-observer agreement for the fully-automatic diameter method were approximately 12 times narrower than the semi-automatic method and 18 times narrower than for the manual method.
Table 22

Mean difference, standard deviation of the differences, and Bland Altman 95% limits of agreement for inter-observer comparisons of polyp diameter and volume measurements for two phantom types.

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Phantom type</th>
<th>Mean difference (mm) (radiologist – technician)</th>
<th>SD of differences</th>
<th>95% Bland-Altman limits of agreement (mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fully automatic diameter</td>
<td>QRM</td>
<td>-0.01</td>
<td>0.03</td>
<td>(-0.07, 0.05)</td>
</tr>
<tr>
<td></td>
<td>Whiting</td>
<td>-0.02</td>
<td>0.04</td>
<td>(-0.09, 0.06)</td>
</tr>
<tr>
<td>Semi automatic diameter</td>
<td>QRM</td>
<td>0.71</td>
<td>0.28</td>
<td>(0.16, 1.27)</td>
</tr>
<tr>
<td></td>
<td>Whiting</td>
<td>0.28</td>
<td>0.56</td>
<td>(-0.82, 1.37)</td>
</tr>
<tr>
<td>Manual diameter</td>
<td>QRM</td>
<td>0.11</td>
<td>0.73</td>
<td>(-1.32, 1.55)</td>
</tr>
<tr>
<td></td>
<td>Whiting</td>
<td>-0.15</td>
<td>0.59</td>
<td>(-1.24, 0.94)</td>
</tr>
<tr>
<td>Fully automatic volume</td>
<td>QRM</td>
<td>-2.3</td>
<td>8.5</td>
<td>(-19.0, 14.3)</td>
</tr>
<tr>
<td>Semi automatic volume</td>
<td>QRM</td>
<td>30.5</td>
<td>24.5</td>
<td>(-17.4, 78.4)</td>
</tr>
</tbody>
</table>

SD = standard deviation
Concerning inter-observer agreement for polyp volume, the 95% limits of agreement were again tightest for the fully-automated measurement, and were approximately three times narrower than the span obtained using the semi-automatic measurement; span 33.3 mm$^3$ for the fully-automatic measurement versus 95.8 mm$^3$ for the semi-automatic measurement (Table 22).

Overall, there was a tendency for the radiologist to make larger estimates of diameter than the radiographer when using both the semi-automatic and manual techniques, but this difference largely disappeared when using the fully-automatic technique. For example, the mean difference between the radiologist and radiographer across all polyps for the QRM phantom was 0.71 mm for the semi-automatic method and 0.11 mm for the manual method but only -0.01 mm for the fully-automatic method (Table 22).

6.3.3.3 Difference between observed measurements and true value

Differences between the estimated and true values for polyp diameter and volume for each method of analysis are summarized in Table 23. There were significant differences between all three measurement methods of polyp diameter for the QRM phantom, but for the radiologist only. There were no significant differences between methods for either observer when measuring polyp diameter using the Whiting phantom. Similarly, significant differences between the estimated polyp volume and the true value were found for the radiologist only, due to a large median difference for
the semi-automatic method (Table 23). As might be expected, the magnitude of the differences between the observed and true values tended to increase as the known polyp volume increased. Because of this, a simple difference in values is not appropriate for all volumes. In order to account for this we examined the differences between measurements using a log-transformed scale, where the limits of agreement now represent the ratio of the differences between the sets of measurements (Table 24). There was a tendency for both fully-automatic and semi-automatic methods to overestimate the true value of the polyp volume, especially when the measurement was made by the radiologist (Table 24).
Table 23

Median difference, inter-quartile range, and significance for estimated values of polyp diameter and volume relative to the true value.

<table>
<thead>
<tr>
<th>Phantom and measurement type</th>
<th>Observer</th>
<th>Measurement method</th>
<th>Median difference (IQ range) (mm)</th>
<th>Significance* (p)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Polyp diameter</td>
<td>Radiologist</td>
<td>Fully-automatic</td>
<td>1.00 (0.48, 1.39)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Manual</td>
<td>0.62 (0.24, 0.86)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Fully-automatic</td>
<td>1.00 (0.48, 1.39)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Semi-automatic</td>
<td>0.23 (0.10, 0.32)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Manual</td>
<td>0.62 (0.24, 0.86)</td>
<td>0.005</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Semi-automatic</td>
<td>0.23 (0.10, 0.32)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Technician</td>
<td>Fully-automatic</td>
<td>1.00 (0.48, 1.39)</td>
<td>0.09</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Manual</td>
<td>0.65 (0.17, 0.93)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Fully-automatic</td>
<td>1.00 (0.48, 1.39)</td>
<td>0.57</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Semi-automatic</td>
<td>0.90 (0.50, 1.00)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Manual</td>
<td>0.65 (0.17, 0.93)</td>
<td>0.52</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Semi-automatic</td>
<td>0.90 (0.50, 1.00)</td>
<td></td>
</tr>
<tr>
<td>Whiting</td>
<td>Radiologist</td>
<td>Fully-automatic</td>
<td>0.60 (0.38, 2.45)</td>
<td>0.33</td>
</tr>
<tr>
<td>Polyp diameter</td>
<td></td>
<td>Manual</td>
<td>0.94 (0.37, 1.61)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Fully-automatic</td>
<td>0.60 (0.38, 2.45)</td>
<td>0.07</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Semi-automatic</td>
<td>0.52 (0.18, 1.18)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Manual</td>
<td>0.94 (0.37, 1.61)</td>
<td>0.20</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Semi-automatic</td>
<td>0.52 (0.18, 1.18)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Technician</td>
<td>Fully-automatic</td>
<td>0.60 (0.37, 2.42)</td>
<td>0.72</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Manual</td>
<td>0.76 (0.26, 1.81)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Fully-automatic</td>
<td>0.60 (0.37, 2.42)</td>
<td>0.39</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Semi-automatic</td>
<td>1.05 (0.21, 1.71)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Manual</td>
<td>0.76 (0.26, 1.81)</td>
<td>0.72</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Semi-automatic</td>
<td>1.05 (0.21, 1.71)</td>
<td></td>
</tr>
<tr>
<td>QRM</td>
<td>Radiologist</td>
<td>Fully-automatic</td>
<td>13.4 (4.4, 57.6)</td>
<td>0.008</td>
</tr>
<tr>
<td>Polyp volume</td>
<td></td>
<td>Semi-automatic</td>
<td>25.0 (14.1, 74.2)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Technician</td>
<td>Fully-automatic</td>
<td>12.7 (3.8, 50.5)</td>
<td>0.23</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Semi-automatic</td>
<td>13.6 (5.4, 29.7)</td>
<td></td>
</tr>
</tbody>
</table>

*Wilcoxon matched-pairs test
Table 24

Ratio of the difference between the estimated and known polyp volume, with 95% limits of agreement on a log-transformed scale (measurements taken from QRM phantom only).

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Observer</th>
<th>Ratio difference (observed/true)</th>
<th>95% Bland-Altman limits of agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fully automatic</td>
<td>Radiologist</td>
<td>1.17</td>
<td>(0.89, 1.52)</td>
</tr>
<tr>
<td>volume</td>
<td>Technician</td>
<td>1.18</td>
<td>(0.95, 1.47)</td>
</tr>
<tr>
<td>Semi automatic</td>
<td>Radiologist</td>
<td>1.37</td>
<td>(0.78, 2.39)</td>
</tr>
<tr>
<td>volume</td>
<td>Technician</td>
<td>1.02</td>
<td>(0.63, 1.65)</td>
</tr>
</tbody>
</table>
6.3.4 DISCUSSION

We hypothesised that fully automated measurements would be less prone to inter- and intra-observer error than those that depended on subjective factors to a larger degree, such as cursor placement and boundary identification. We developed novel software with this goal in mind and were able to demonstrate that the two automated methods for diameter estimation had superior inter- and intra-observer limits of agreement when compared to the conventional manual measurement. The fully automated technique fared best, with inter-observer limits of agreement approximately 18 times narrower than the conventional manual method. The clear implication is that different observers measuring the same polyp are much more likely to agree when using this method than when using conventional callipers that have been placed subjectively by eye. A further benefit of the fully-automatic measurement is that it is fast; approximately 30 seconds per polyp. In contrast, the semi-automatic method of volume assessment is slow, taking approximately 5 minutes per polyp. This is because the observer must draw the polyp boundary freehand on all contiguous images that demonstrate it, and the larger the polyp, the longer this process will take. Furthermore, the exact polyp boundary with normal colonic wall may be difficult to identify by eye and therefore prolong the process, especially when dealing with sessile polyps. Its exact position will also be subjective.

We investigated two different observers, a radiologist and radiographic technician, and found significant differences between them when comparing their estimates of polyp diameter and volume to the true measurement. While the technician's estimates
were not significantly different from the true measurement for any of the methods investigated, the radiologist had a tendency to overestimate the true value when using the QRM phantom. This phenomenon was especially noticeable with the semi-automatic estimate of polyp volume and implies that the polyp boundary drawn by the radiologist was generally significantly larger than that drawn by the technologist. However, there was also a general tendency for the fully-automated software analysis to overestimate both the diameter and volume measurements, a phenomenon that we found was greatest with the larger polyps. There was a considerable difference in volume between the largest and smallest polyps studied, from 7.3 mm$^3$ to 520.4 mm$^3$ and it is likely that the overestimation we observed is due to partial volume effects. We are currently working on a software solution to minimise this. These results should be interpreted with caution since only one observer from each professional category was used and each is thus acting as a surrogate proxy for all radiologists and technicians respectively. However, our results do not bear out the assumption that measurements made by radiologists will be more accurate than those made by technicians since we found the reverse to be true.

Our study does have limitations. Most notably, it is a phantom study performed under ideal conditions in vitro and the findings may not be generalizable to the clinical situation in vivo. Supporting this, our results differed to some extent depending on the phantom used, suggesting that the phantom construction had some effect on the function of the software. By way of explanation, a phantom does not exactly mimic the clinical reality, for example the polyp/wall interface and adjacent structures outside the lumen. Moreover, in reality polyps are rarely truly spherical and the algorithm tested in the present study would need to be shown to be effective in situations where
the morphology of the polyp being measured was more complex. Also, the choice of technical parameters such as the collimation and reconstruction interval may also potentially affect the performance of the algorithm. A follow-up study in vivo is clearly indicated. It could also be argued that measurement of polyp volume has little merit for day-to-day practice because it is not made routinely, probably because it would be difficult to calculate. However, the volumetric nature of helical CT data offers the possibility of easy volume measurement and this may eventually prove to be a more sensitive measurement of serial growth than simple estimates of diameter. We also used observers who had no prior experience of CT colonography, which we felt was acceptable because our aim was to determine differences in simple measurements rather than differences in detection. However, it is possible that prior experience of measurement during CT colonography may have conveyed an advantage.

However, one benefit of our approach is that the diameter and volume of each polyp was known with certainly. In reality, obtaining accurate measurements of real polyps is fraught with difficulty. Studies of CT colonography have generally used comparison with adjacent open biopsy forceps as a reference standard, a procedure routinely performed during colonoscopy. Nevertheless, it is well-recognised that this approach is often inaccurate. For example, an endoscopic study of polyp size determined by visual estimation, linear probe, and biopsy forceps found that the latter technique was the most inaccurate [161]. A study of artificial polyps of known size found that endoscopic estimates tended to be lower than the true size by approximately 30% [162]. Other studies have found colonoscopists' estimates at variance with those obtained immediately after polypectomy and also at subsequent histopathological examination [163;164]. These factors will conspire to frustrate meaningful analysis of
automated polyp measurement and phantom studies are thus particularly relevant to this field of research. It was unavoidable that the radiologist reader knew in advance the diameter of the polyps in the QRM phantom since they had previously used this phantom for studies of lung nodule detection. The possibility that this could have biased their manual estimates of polyp diameter in this phantom must be considered although this knowledge cannot have biased the semi-automatic and fully automatic measurements of both diameter and volume since these were generated by segmentation outside the control of the radiologist. In any event, despite this potential bias in favour of the manual estimate, intra-observer agreement for this reader and phantom combination was still worst for the manual estimates. It could also be argued that we should have investigated observers who were expert in CT colonography. However, although this was potentially possible, these authors were heavily involved in the design, acquisition and accumulation of data for this study and we felt the potential for bias was consequently too high. We would also argue that prior experience in CT colonography predominantly relates to sensitivity and specificity for polyp detection rather than measurement.

In summary, we have demonstrated that fully-automated measurement of polyp diameter and volume is technically feasible in vitro and results in superior inter- and intra-observer agreement when compared to conventional manual methods. Contingent upon further validation in-vivo, this approach has the potential to facilitate accurate measurements of polyp morphology and ultimately may have a role in accurate serial monitoring of those patients in whom small polyps have been identified using CT colonography.
6.4 CT colonography: Automatic measurement of polyp diameter compared with manual assessment: An in-vivo study

6.4.1 INTRODUCTION

This thesis has demonstrated previously in-vitro (Chapter section 6.3) that automated software can improve both the repeatability and reproducibility of polyp measurements when compared to manual assessment, of artificial polyps. However, in-vitro data may not be generalisable to clinical practice, for example due to variable polyp morphology in vivo. The purpose of this study was to compare automated polyp measurement with manual assessment in vivo.

6.4.2 MATERIALS AND METHODS

6.4.2.1 Dataset

The study dataset was accrued from two different institutions, both of whom had obtained institutional ethical review board approval for data sharing. In total there were 33 patients with 50 polyps: 17 patients with 26 polyps and 16 patients with 24 polyps from each institution respectively.

The median maximal diameter was 7mm (range 5mm to 12mm), determined in each case by colonoscopic assessment using either a linear measurement probe or comparison with adjacent open biopsy forceps; 11 polyps were 5mm; 12 were 6mm; 5 were 7mm; 6 were 8mm; 4 were 9mm; 8 were 10mm; 2 were 11mm and 2 were
12mm in maximal diameter. The segmental location of polyps was as follows: 6 rectal; 21 sigmoid; 6 descending colon; 6 transverse colon; 5 ascending colon; 6 cecal.

In each case patients had undergone CT colonography immediately prior to colonoscopy using a technique approved by a consensus of radiologists, experienced in CT colonography [25]. Prone and supine scanning was performed using multidetector row machines at a maximum individual slice collimation of 2.5mm. Full bowel preparation was used with one centre (16 patients) supplementing this with 500 mls of 2.1% barium suspension for positive fecal-tagging. One centre used room air as the insufflating gas (16 patients) and the other used carbon dioxide (17 patients).

The DICOM data of the 33 individual patient examinations were anonymized, a study number assigned, copied to CD, and transferred to the study office. The principal investigator of this thesis determined the segmental location and CT co-ordinates (both prone and supine when visible on both) for each of the known 50 polyps and recorded this on a study sheet, using the previous CT colonography and endoscopic reports for guidance. A commercially available workstation was used for visualization (Vitrea version 3.7, Vital Images, Inc, Minnesota, USA).

6.4.2.2 Polyp measurement

Two observers (a radiologist and radiographic technician), were provided with a study sheet that listed the 50 polyps, their CT coordinates (polyp detection was not an aim of this study), and the study number for each individual patient. Both observers were experienced in CT colonography interpretation and had read approximately 500 and
200 endoscopically validated cases respectively. The observers were unaware of the reference measurement for each polyp and also unaware of the distribution of polyp sizes in the dataset.

Each observer measured the maximal diameter of each polyp using a 2D display and the following four CT window settings: colon ‘modified lung’ (Width 1500, Level – 200, Fig. 24.a); lung (Width 700, Level – 600, Fig. 24b); abdominal (Width 400, Level 10, Fig. 24c); bone (Width 3500, Level 400, Fig. 24d). A fifth estimate was made using a 3D endoluminal view (Fig. 24e). Proprietary software (see above) and electronic software calipers were used. One observer was familiar with this platform while the other was trained by the principal investigator of this thesis, who was also available throughout for technical assistance. Observers were free to select which one of the prone and supine studies best depicted the polyp for measurement of its maximal diameter, as would happen in day-to-day clinical practice, and were free to choose which anatomical plane (or viewing angle in the case of 3D) from which to make the measurement. They were also aware of the need to magnify images to aid accurate measurement, especially for small polyps.
Figs 24 a-e

10mm maximal diameter tubular adenoma located in the sigmoid colon of a 62 years old man; displaying measurements found using colon CT windows (Width 1500, Level – 200, fig 1a); lung (Width 700, Level – 600, fig 1b); abdominal (Width 400, Level 10, fig 1c); bone (Width 3500, Level 400, fig 1d) and using a 3D endoluminal view (fig 1e).

Fig. 24a

Fig. 24b
After an interval of one month, to diminish recall bias, observers re-measured all fifty polyps using a second visualization platform that permitted automatic diameter measurements (MedicColon version 1.3, Medicsight PLC, London, UK). Both observers were trained on the software by the principal investigator of this thesis who was again available for assistance throughout the reading period. Using a mouse and cursor, observers placed two seed points opposite one-another, at the perceived junction between the polyp and the adjacent colonic wall on the 2D image that they felt best represented the maximal polyp diameter. They then initiated automatic segmentation and extraction of the polyp, which provided an assessment of diameter without the need for manual boundary identification (Fig. 25). In order to achieve this, the software automatically repositioned the subjectively placed cursors at the best local candidate for the polyp boundary in the immediate vicinity, and followed this with segmentation that was based on a region-growing scheme. This scheme used fuzzy logic to separate the polyp from surrounding normal structures. The result of this process was then mapped to contiguous slices, where two new seed points were initialised. This procedure was repeated until the entire polyp has been analysed. Results were copied to a data sheet and any failure to automatically segment a polyp was also recorded. As before, observers were unaware of the reference measurement or each other's responses.
Fig 25

Automatic segmentation and extraction of the 10mm polyp (colonoscopic diameter) shown in figure 1, which provided an assessment of diameter without the need for manual boundary identification.
6.4.2.3 **Statistical Methods**

Data were transferred to an electronic database (Microsoft © Excel 2002) for analysis using Stata7.0 (StataCorp, College Station, Texas, USA). Agreement between the two observers and between each observer and the reference measurement was assessed using the Bland-Altman method, constructing 95% limits of agreement around the mean difference between corresponding measurements [165]. Measurement error was determined for each observer and the Wilcoxon matched pairs test was used to compare these errors between observers. Subset analysis using the Mann-Whitney U test statistic was performed to determine any significant differences in agreement between tagged and non-tagged cases, for each observer.
6.4.3 RESULTS

There were 100 possible assessments of maximal polyp diameter (50 polyps measured by two observers) for each of the six methods of measurement. All manual assessments of diameter were successfully obtained whereas the automatic measurement software provided assessments for 90 (90%) polyps: Observer 1 and observer 2 obtained automatic measurements for 46 and 44 polyps respectively. Failure of polyp segmentation and extraction was cited as the reason for automatic measurement failure for all 10 polyps (median size 6.5mm, range 5-10mm).

For manual measurements using 2D displays, observer 1 (radiologist) chose the axial anatomical plane to measure 18 polyps, coronal for 17 and sagittal for 15. In contrast, observer 2 (radiographic technician) chose the axial plane to measure 44 polyps; coronal for 5 and sagittal for 1 polyp.

6.4.3.1 Inter-observer agreement

The mean difference between the two observers' estimates of polyp diameter, standard deviation of these differences, and limits of agreement for each of the six methods of polyp measurement are summarized in Table 25. For four of the measurement methods, observer 2 (radiographic technician) on average provided larger estimates of polyp diameter than observer 1 (radiologist). The largest mean difference between observers was found when making measurements using the 3D display. For both observers, the limits of agreement spanned a similar range for all measurement methods (6.3 to 8.2mm) but the widest span was obtained when using
either the 2D abdominal window display or the automatic measurement software. These data also reveal a difference of up to 5mm between measurements of the same polyp by the two observers.
Table 25. Inter-observer agreement.

<table>
<thead>
<tr>
<th>Measurement method</th>
<th>Mean difference (observer1–observer2) (mm)</th>
<th>Sd difference (mm)</th>
<th>95% Bland-Altman limits of agreement (mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2D Colon</td>
<td>0.0</td>
<td>1.7</td>
<td>(-3.3, 3.3)</td>
</tr>
<tr>
<td>2D Abdominal</td>
<td>-0.9</td>
<td>2.1</td>
<td>(-5.0, 3.2)</td>
</tr>
<tr>
<td>2D Lung</td>
<td>0.5</td>
<td>1.8</td>
<td>(-3.0, 3.9)</td>
</tr>
<tr>
<td>2D Bone</td>
<td>-0.2</td>
<td>1.7</td>
<td>(-3.6, 3.2)</td>
</tr>
<tr>
<td>3D</td>
<td>-1.1</td>
<td>1.6</td>
<td>(-4.3, 2.0)</td>
</tr>
<tr>
<td>Automatic</td>
<td>-0.5</td>
<td>2.1</td>
<td>(-4.6, 3.6)</td>
</tr>
</tbody>
</table>
6.4.3.2 Agreement with colonoscopy

The mean difference between each of the two observers' estimates of polyp diameter and the colonoscopy reference, standard deviation of these differences, and limits of agreement for each of the six methods of polyp measurement are summarised in Table 26. Both observers tended to overestimate polyp diameter except when utilising an abdominal window display. On average, overestimates were greatest when using either the 3D display or automatic software. The limits of agreement were relatively wide for all measurement methods (span ranging from 9.5 to 13.4mm), with the widest limits again observed when using the 3D display or automatic software.
Table 26. Agreement with colonoscopy.

<table>
<thead>
<tr>
<th>Measurement method</th>
<th>Observer</th>
<th>Mean difference (observer – colonoscopy) (mm)</th>
<th>Sd difference (mm)</th>
<th>95% Bland-Allman limits of agreement (mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2D Colon</td>
<td>1</td>
<td>0.7</td>
<td>2.6</td>
<td>(-4.4, 5.7)</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.7</td>
<td>2.4</td>
<td>(-4.1, 5.4)</td>
</tr>
<tr>
<td>2D Abdo</td>
<td>1</td>
<td>-1.0</td>
<td>2.6</td>
<td>(-6.2, 4.2)</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>-0.1</td>
<td>2.5</td>
<td>(-4.9, 4.8)</td>
</tr>
<tr>
<td>2D Lung</td>
<td>1</td>
<td>1.0</td>
<td>2.6</td>
<td>(-4.1, 6.0)</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.5</td>
<td>2.6</td>
<td>(-4.5, 5.5)</td>
</tr>
<tr>
<td>2D Bone</td>
<td>1</td>
<td>0.2</td>
<td>2.6</td>
<td>(-5.0, 5.4)</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.4</td>
<td>2.5</td>
<td>(-4.5, 5.3)</td>
</tr>
<tr>
<td>3D</td>
<td>1</td>
<td>1.3</td>
<td>2.7</td>
<td>(-3.9, 6.5)</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>2.5</td>
<td>3.3</td>
<td>(-3.9, 8.9)</td>
</tr>
<tr>
<td>Automatic</td>
<td>1</td>
<td>0.7</td>
<td>2.8</td>
<td>(-4.7, 6.1)</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1.2</td>
<td>3.4</td>
<td>(-5.5, 7.9)</td>
</tr>
</tbody>
</table>
When comparing the measurement error between each of the observers and colonoscopy, we found a significant difference when observers used either the 2D abdominal window (median difference 1 mm, interquartile range -0.3 to 1.3, p = 0.03) or 3D display only (median difference -1 mm, interquartile range -2.0 to 0: p<0.001) for observer 1 and observer 2 respectively.

6.4.3.3 Effect of faecal tagging on agreement

For observer 1, no significant effect of faecal tagging was identified for any estimate. However, observer 2 experienced significantly greater diameters compared to colonoscopy in untagged datasets when using the 3D display: Median measurement error of 3 mm (interquartile range 2, 5.5) for untagged cases versus 2 mm (1,3) for tagged cases (p=0.01).

6.4.4 DISCUSSION

This thesis, in concert with work by others [166], has demonstrated the feasibility of automatic polyp segmentation and extraction in vitro. The present study has confirmed that segmentation is also possible in vivo.

As with earlier studies in this chapter, the present study has shown that polyp measurement is prone to a degree of error when compared to colonoscopy, and that there is also variable disagreement between observers asked to measure the same polyp (and indeed between repeated measurements of the same polyp when made
Disagreement was encountered for all of the paradigms tested, including the automated method. This latter finding apparently contradicts the findings of the earlier in vitro study described earlier in this chapter (Chapter 6.3) in which the same automatic measurement software improved inter-observer agreement when compared to manual assessment. A possible explanation is that real polyps are rarely truly spherical and their boundary with the adjacent colonic wall is less well-defined than for artificial polyps sited in a phantom model. As a result, in-vivo measurement presents a greater challenge to accurate identification and segmentation. Supporting this hypothesis, segmentation failed in a greater proportion of polyps in the present study when compared to in-vitro work using the same software (10% versus 7% respectively), despite the smaller average size of the latter.

Once again, manual methods were also prone to disagreement, with 2D assessment using abdominal window settings and 3D endoluminal measurements providing the greatest variability, due to under- and overestimation respectively. These findings concur with an earlier study in this thesis, described in chapter 6, examining the effect of visualisation display on polyp measurement but are at odds with a study by Pickhardt et al [167], which concluded that 3D measurement was more accurate than multiplanar 2D measurement (which these authors found tended to underestimate polyp diameter). Notably, Pickhardt found no significant difference between 3D measurement and ‘optimised’ 2D measurement, which used the 2D multiplanar reformat that best depicted the maximal polyp diameter, with magnification prior to measurement [168]. The present study used experienced observers who were advised to optimise 2D measurement in this way. Pickhardt found that 3D generally provided larger estimates of polyp diameter than 2D [169], a finding in keeping with
our results, although we found that both 2D and 3D generally overestimated polyp
diameter. This might be explained by use of different software with different
calibration. For example, the software used for the Pickhardt study is optimized for
primary 3D reading and measurement [170].

Of the two observers, we found the radiologist most accurate, generally providing
smaller estimates of diameter and frequently choosing a non-axial plane for
measurement. This thesis has already suggested that measurement accuracy may be
improved by experience (Chapter sections 6.1 and 6.2). The radiologist observer was
experienced with using the software and making 2D and 3D measurement in routine
clinical practice whereas the radiographic technician was not. This may help to
explain the relative inaccuracy of observer 2 overall when compared to observer 1.
Experience may also facilitate accurate measurement when decisions related to polyp
morphology are necessary. For example, choice of anatomical plane for best
depiction of diameter; accurate delineation between the polyp boundary and adjacent
wall or haustral fold; exclusion of any adherent fecal matter; correct classification of
whether a polyp is sessile or pedunculated. Correct classification of morphology is
important since the stalk of a pedunculated polyp does not contain adenomatous
tissue and should be excluded from measurement. Endoluminal display usually best
depicts such a stalk and should probably be used routinely for characterization
whenever a polyp is found.

This study also has limitations. Observers were not blinded to their measurements
and it is possible that subsequent measurements of the same polyp may have been
influenced by those made in advance. As in earlier studies, we used a single software
platform for all manual measurements, and so did not investigate the effect of different systems. If such differences do exist, and this thesis suggests they might, they indicate a need for local calibration and polyp management strategies should take account of this. Also both observers were most experienced in 2D reading, a potential bias against 3D measurement, especially if there is a learning curve for 3D measurement, perhaps related to choice of optimal viewing angle for example. While our primary aim was to show whether automatic polyp segmentation and measurement was feasible in-vivo, we also attempted to assess agreement with a reference standard. However, as described above, there is considerable evidence that colonoscopic estimates are unreliable and, a priori, there is no reason to expect that colonoscopy will be any more accurate than CT. Interpretation of our results for agreement with colonoscopy should take this into account. Indeed, while it is important to demonstrate that automatic measurement is possible in vivo (since this is the clinical situation where it would be most used), it may be that accuracy is best determined using phantom data, where the diameter and volume of each individual polyp can be determined with absolute certainty.

In summary, automatic segmentation and measurement of polyp diameter is possible in-vivo and accuracy is comparable to manual methods. Further studies are needed to assess agreement with known estimates of diameter. Manual measurement using optimized multi-planar 2D displays (using either 'colon', bone or lung CT window settings) are more accurate than measurement made using 3D displays, for the platforms used in this study.
CHAPTER 7

CONCLUSIONS

7.1 Introduction

This thesis has attempted to provide evidence based information to inform and guide those interested in implementing CT colonography into routine clinical practice. Firstly, this thesis establishes how CT colonography is practised across the UK and the frequency of complications encountered. The performance of different observer groups who are likely to interpret CT colonography now and in the future has been investigated with the effect of dedicated training assessed. An emphasis on accuracy and report times is appropriate given the effect of implementing CT colonography on CT workflow in busy clinical practice. The accuracy of polyp measurement and categorization, necessary for subsequent polyp management strategies has also been investigated. Finally, variability in manual measurement accuracy prompted assessment of automated computer measurement software, which has been tested both in vitro and in-vivo.

7.2 Discussion of results

There are several potential barriers to the widespread adoption of CT colonography in the United Kingdom including lack of CT scanner capacity and level one evidence to support its general implementation. However CT colonography has attracted
considerable support from health professionals, political bodies and the general public alike and the survey presented in Chapter section 3.1 confirms that CT colonography is widely available in day-to-day practice across the UK; over one-third of departments offered a service. Indeed the survey of complications conducted two years later has shown that the frequency of examinations and the experience gained have increased further; the proportion of centres performing 300 or more examinations per year has increased from 12% in 2003 to 52% in 2005. While limited CT scanner capacity was cited as the main barrier to implementation in 2003, the 2005 survey findings suggest CT access may be improving.

In the UK, CT colonography was used most frequently for failed conventional colonoscopy and there is a good evidence base to support such a role. However, 74% of departments also used CT colonography for examining frail or immobile patients as an alternative to more established whole-colon techniques, despite a lack of direct evidence for this role and the potential disadvantages of using bowel preparation in these groups.

Since its inception, CT colonography has been promulgated as a safe technique. However, like conventional colonoscopy [77], there is concern that complications may occur more frequently in routine clinical practice. Reassuringly, the data presented in Chapter section 3.2 has revealed a relatively low symptomatic perforation rate (when compared to conventional colonoscopy) of only 0.03% with no deaths and only one of the nine patients with perforation requiring laparotomy. Moreover, several of the perforations were potentially avoidable with judicious use of rectal balloon catheters and knowledge of prior colonic surgery.
Highly experienced individuals agree that specific and supervised training is a prerequisite for acceptable CT colonography performance [16]. They specify such training should involve interpretation of forty to fifty endoscopically validated examinations. However, there is little evidence to support or refute this. The data presented in Chapter section 4.1 shows that the overall sensitivity of novices trained using this scheme is significantly inferior to experienced observers, for readers in an academic environment. For example, both trained groups detected approximately 70% cancers, significantly less than the 92% detected by experienced observers. This discrepancy suggests that, on average, training with fifty cases is inadequate if the aspiration is to match the performance of very experienced readers. However, this number may be sufficient for certain 'gifted' individuals as there is considerable overlap in individual performance between all groups; for example some individual radiographers exceeded the mean performance of more experienced radiologists. This latter finding has particular relevance for the UK, where barium enema is widely performed by radiographers and suitably experienced radiologists are in short-supply. Overall, we found no performance difference between trained radiologists and radiographers, suggesting that prior interpretation of conventional abdominal CT may not confer benefit for CT colonography.

While these findings above are important for guiding training schedules, this thesis has found that most patients undergoing CT colonography in the UK are examined in district general hospitals, outside of an academic environment. Because the technique has disseminated rapidly, most practitioners have had little or no specific training in either interpretation or technical performance. Despite this, the data
presented in Chapter section 4.2 have shown that, on average, sub-specialist radiologists in the UK interpreted CT colonography examinations as accurately as radiologists who had been trained using 50 colonoscopically validated cases. Once again, as above, we found wide variability in individual performance, but that this generally falls short of the average performance suggested by meta-analysis of published data.

For successful implementation both as a symptomatic and screening modality, CT colonography interpretation must be time efficient. This thesis examined multiple observers of differing experience across several European countries, and found that, on average, readers took less than fifteen minutes to interpret a case (Chapter 5). These interpretation times are similar to previous multi-observer studies utilising a primary 2D interpretation method and shorter than those employing a primary 3D approach [7;12]. Such interpretation times are comparable to optical colonoscopy, but prolonged when compared to barium enema. We found that experienced observers read significantly faster than others, taking on average ten minutes for normal studies, and encouragingly, the best performing experienced observers also reported fastest. In contrast, interpreting cases more slowly improves accuracy amongst radiographers. In addition, perhaps unsurprisingly given the busy NHS clinical environment, non-academic, sub-specialist radiologists interpreted as fast as experienced observers in academic centres, which has implications for workflow in busy clinical practice.

Accurate assessment of polyp size during CT colonography is fundamental for determining biological significance and therefore whether or not the patient should
proceed to colonoscopy and polypectomy. This thesis has shown that error occurs during CT colonography and is dependent upon the observer, their experience, and the viewing conditions used to make the measurement (Chapter section 6.1). Indeed limits of agreement as wide as 14mm, raise the possibility that a polyp whose true diameter was 10mm + could be assigned to the 'small' category as a result of measurement error and thus be either assigned to interval surveillance or even disregarded when the correct course of action is endoscopic polypectomy.

Conversely, by including 'clinically insignificant' polyps (2 to 5mm polyps inclusive) in our dataset, we showed that diminutive polyps could be potentially miscategorised as large and therefore inappropriately referred for polypectomy. In addition inaccurate measurements and reduced observer agreement were most likely using a 3D display, indicating it is a less reliable method with which to measure polyps. Although use of a 'colonography CT window display' generally overestimated polyp diameter, this is likely the best display for size assessment as it was associated with the narrowest limits of agreement and the least inter-observer error overall (irrespective of agreement with the reference values). That said, in day-to-day clinical practice, the window display and anatomical plane used for measurement should be documented to facilitate any subsequent interval comparisons.

Polyp management strategies are a sensible step towards standardisation of CT colonography reporting and the avoidance of unnecessary colonoscopy in patients whose polyps are diminutive. However, this thesis has shown wide variability amongst observers of different experience and from different centres when asked to measure the same polyp (Chapter section 6.2). Perhaps most importantly, in 20% of observations, large polyps were misclassified into a smaller category raising the
possibility that individuals with an advanced adenoma could be returned to surveillance alone. The difficulty of using a single discrete threshold to define an advanced adenoma has been highlighted, suggesting a colonographic diameter of 8mm might be a more reasonable threshold to trigger colonoscopy [7]. In our study, a threshold of 8mm would have meant correct categorisation of 99% of those polyps with a reference diameter of 1cm or more and a threshold of 7mm would have meant correct categorisation of all such polyps. In contrast to findings of the previous study, CT colonography tended to underestimate polyp diameter but different software was used raising the possibility that different software platforms may provide different measurements of the same polyp, for example due to different calibration methods, displays, or software caliper functionality.

This thesis also investigated the use of automated measurement software (Chapter sections 6.3 and 6.4), hypothesising that fully automated measurements would be less prone to inter- and intra-observer error than those that depended on subjective factors to a larger degree, such as cursor placement and boundary identification. For the in vitro study (Chapter section 6.4), automated diameter estimation had superior inter- and intra-observer limits of agreement when compared to the conventional manual measurement. The fully automated technique fared best, with inter-observer limits of agreement approximately 18 times narrower than the conventional manual method, indicating that this approach has the potential to facilitate accurate polyp measurements and ultimately may have a role in accurate serial monitoring of those patients in whom small polyps have been identified using CT colonography. However, the in-vivo study (Chapter section 6.5), comparing a number of measurement methods including the automated method showed no distinct advantage for any one
method. This latter finding apparently contradicts the findings of the in vitro study and shows that in-vivo measurement presents a greater challenge to accurate automatic polyp identification and segmentation. Manual methods were also prone to disagreement, with 2D assessment using abdominal window settings and 3D endoluminal measurements providing the greatest variability, due to under- and overestimation respectively. These findings concur with the earlier study examining the effect of visualisation display on polyp measurement.
7.3 Conclusion

CT colonography is a potentially useful diagnostic technique for accurately and safely examining the colon, which might adopt a complementary position alongside its therapeutic counterpart, conventional colonoscopy. However, successful and widespread implementation is contingent upon acceptable performance characteristics being generalisable to routine clinical practice.

This thesis has attempted to define the current status of CT colonography practise in the UK and has also provided evidence to help inform accurate and efficient interpretation. Evidence has been presented to inform interested parties that CT colonography is widely available in the UK NHS but that experience is variable; that CT colonography is generally safe and complications are potentially avoidable; experience improves accuracy but alone is insufficient to determine competence; fifty validated cases, on average, are insufficient to ensure competence but there is wide variability amongst observers and certain individuals are competent with this limited number; and that radiographers can achieve similar competence to radiologists suggesting potentially an important role in the future implementation of CT colonography.

Finally evidence showing variability in polyp measurement accuracy and revealing methods for optimising accuracy, reproducibility and repeatability have been revealed, including a potential role for automatic measurement methods in the future.
I believe that CT colonography is a powerful tool in the armamentarium for diagnosis of early colorectal cancer and polyp detection but this thesis reiterates the need for formal training to optimise interpretation techniques and indicates that specific accreditation needs be considered, particularly if the role of CT colonography is extrapolated to screening.
APPENDIX 1.

Dear Clinical Director / Gastrointestinal Radiology Lead,

Survey to determine CT Colonography ('virtual colonoscopy') practice in the UK

Please tick the appropriate boxes;

<table>
<thead>
<tr>
<th>Question</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. a. Does your department have a CT scanner?</td>
<td>☑</td>
<td>☑</td>
</tr>
<tr>
<td>b. Is it a multislice scanner?</td>
<td>☑</td>
<td>☑</td>
</tr>
<tr>
<td>c. Do you have 3D CTC software?</td>
<td>☑</td>
<td>☑</td>
</tr>
</tbody>
</table>

2. Does your department offer a CTC service in day-to-day clinical practice? | Yes | No |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>If you answered 'yes' to 2 please go to question 4. If 'no' then answer question 3.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

3. Why does your department currently not offer CTC? Please tick any applicable response.
   - Financial restrictions (hardware or software) ☑
   - Scanner capacity limited by other clinical demands. ☑
   - Lack of Radiological expertise/training in CTC technique ☑

If you answered question 3 please return the questionnaire in the enclosed S.A.E

4. Approximately how many CTC studies are performed in your department as a rough and ready estimate?
   - Approx.1 / month ☑
   - 1 / week ☑
   - 1 / day ☑
   - >1 / day ☑

5. Approximately how many CTC studies has your department performed over the last 5 years?
   - 0-20 ☑
   - 21-50 ☑
   - 51-100 ☑
   - 101-300 ☑
   - 300+ ☑
6. Please tick the following clinical circumstances for which you routinely offer CTC?

Incomplete colonoscopy [ ]
Failed barium enema [ ]
As an alternative to barium enema in frail or immobile patients [ ]

6. Regarding CTC technique;  
   Yes  No
   a. Do you generally use IV contrast? [ ] [ ]
   b. Do you generally do prone and supine scans? [ ] [ ]
   c. Do you always use full bowel preparation? [ ] [ ]
   d. Do you distend the colon with air or CO2? Air [ ] CO2 [ ]

7. Is there a facility for CTC to be performed by radiographers and reported by radiologists later?  
   Yes [ ] No [ ]

8. Who is reporting CTC? Please choose the most applicable from the two following suggestions (or both if also applicable):
   a. Radiologist(s) with a declared subspecialty interest in GI radiology, part of whose remit includes CT?  
      Yes [ ]
   b. Radiologist(s) with a declared subspecialty interest in CT (rather than GI), or a general radiologist whose does CT?  
      Yes [ ]
APPENDIX 2.

Telephone questionnaire used for a study of complications associated with CT colonography.

<table>
<thead>
<tr>
<th>Question</th>
<th>Possible responses</th>
</tr>
</thead>
</table>
| Approximately how many CT colonography studies does your department perform on average at the moment? | • More than one per day  
• One per day  
• One per week  
• One per month |
| Approximately how many CT colonography studies has your department performed in total? | • Total given. |
| How frequently does your department use inflated rectal balloon catheters for CT colonography? | • Never  
• Occasionally (please give an approximate %)  
• Always. |
| Does your department use an automated colonic insufflation device?       | • Yes  
• No |
| To the best of your knowledge, has there been any bowel perforation related to CT colonography? | • Yes (please give number)  
• No |
| To the best of your knowledge, has there been any other serious adverse event associated with CT colonography? For example, reactions to intravenous contrast or spasmolytic. | • Yes (please give number)  
• No |
APPENDIX 3.

LIST OF PUBLICATIONS ARISING FROM THIS THESIS

A. ORIGINAL ARTICLES


B. REVIEWS AND EDITORIALS


C. CHAPTERS IN BOOKS


D. LETTERS


E. PUBLISHED ABSTRACTS


2. ESGAR CT Colonography Study Group Investigators. CT Colonography: Effect of Directed Training on Reader Performance. European Radiology 2004;14 supp


F. PERSONAL PRESENTATIONS

Dec 2003: Great Western Hospital Grand round, Swindon.

Great Western Hospital Grand round, Swindon
CT Colonography

January 2004: Arab Health Imaging and Diagnostic conference
Dubai.
Multi-detector CT Colonography: Virtually replacing the barium enema?

June 2004: European Society Gastrointestinal Abdominal Radiology
Geneva, Switzerland.
CT Colonography: Effect of Directed Training on Reader Performance
European Multi-Centre Study

October 2004: St. Mark’s Association Annual Meeting
St. Mark’s Hospital, London
Virtual Colonoscopy

May 2005: Oxford Regional Radiologist Meeting
John Radcliffe hospital, Oxford.
CT colonography: only for experts?
October 2005: BIR colon imaging day
Royal Geographic Society, London
CT colonography as the first line colonic investigation.

October 2005: 6th International Virtual Colonoscopy Symposium
Boston, USA
CT colonography: Scan options.

January, 2006: ESGAR CT Colonography workshop
Edinburgh
Current UK implementation & complications

February 2006: St. Mark's Hospital Grand Round
St. Mark's Hospital, London
Reader performance at Virtual Colonoscopy

May 2006: M25 Surgical course
Pelican Centre, Basingstoke
Virtual Colonoscopy

June 2006: Barts and The London Radiology meeting
St Bartholomews Hospital,
Virtual colonoscopy

June 2006: European Society Gastrointestinal Abdominal Radiology
Crete, Greece

Co-presented with Dr H Fenlon: Virtual Colonoscopy: Difficult cases

October 2006: 7th International Virtual Colonoscopy Symposium

Boston, USA

Strategies of Current Implementation: U.K.
Ref Type: Abstract


Acceptance by patients of multidetector CT colonography compared with barium enema examinations, flexible sigmoidoscopy, and colonoscopy. AJR Am J Roentgenol 2003; 181:913-921


Ref Type: Abstract


71. Brooks D. American Cancer Society Position Statement on CT Colonography. In: Fifth International Symposium of Virtual Colonoscopy meeting program,
Ref Type: Abstract


75. Halligan S, Atkin W. Unbiased studies are needed before virtual colonoscopy can be dismissed. Lancet 2005; 365:275-276


Ref Type: Abstract


Ref Type: Abstract


103. Rex DK, Vining D, Kopecky KK. An initial experience with screening for colon polyps using spiral CT with and without CT colography (virtual colonoscopy). Gastrointest Endosc 1999; 50:309-313


Ref Type: Abstract


121. Read TE, Read JD, Butterly LF. Importance of adenomas 5 mm or less in diameter that are detected by sigmoidoscopy. N Engl J Med 1997; 336:8-12


127. Schoen RE, Gerber LD, Margulies C. The pathologic measurement of polyp size is preferable to the endoscopic estimate. Gastrointest Endosc 1997; 46:492-496


131. Schoen RE, Gerber LD, Margulies C. The pathologic measurement of polyp size is preferable to the endoscopic estimate. Gastrointest Endosc 1997; 46:492-496


137. Schoen RE, Gerber LD, Margulies C. The pathologic measurement of polyp size is preferable to the endoscopic estimate. Gastrointest Endosc 1997; 46:492-496

138. Schoen RE, Gerber LD, Margulies C. The pathologic measurement of polyp size is preferable to the endoscopic estimate. Gastrointest Endosc 1997; 46:492-496


140. Macari M, Bini EJ, Jacobs SL, Lange N, Lui YW. Filling defects at CT colonography: pseudo- and diminutive lesions (the good), polyps (the bad), flat
lesions, masses, and carcinomas (the ugly). Radiographics 2003; 23:1073-1091


Ref Type: Abstract


146. Rex DK. Endoscopists, polyp size, and post-polypectomy surveillance: making a mountain out of a molehill? Gastrointest Endosc 1997; 46:571-574


Ref Type: Abstract


Ref Type: Abstract


Ref Type: Abstract


Ref Type: Abstract

Ref Type: Report


163. Schoen RE, Gerber LD, Margulies C. The pathologic measurement of polyp size is preferable to the endoscopic estimate. Gastrointest Endosc 1997; 46:492-496


