
Downloaded from: http://insight.cumbria.ac.uk/id/eprint/2093/

Usage of any items from the University of Cumbria’s institutional repository ‘Insight’ must conform to the following fair usage guidelines.

Any item and its associated metadata held in the University of Cumbria’s institutional repository Insight (unless stated otherwise on the metadata record) may be copied, displayed or performed, and stored in line with the JISC fair dealing guidelines (available here) for educational and not-for-profit activities provided that

• the authors, title and full bibliographic details of the item are cited clearly when any part of the work is referred to verbally or in the written form
• a hyperlink/URL to the original Insight record of that item is included in any citations of the work
• the content is not changed in any way
• all files required for usage of the item are kept together with the main item file.

You may not

• sell any part of an item
• refer to any part of an item without citation
• amend any item or contextualise it in a way that will impugn the creator’s reputation
• remove or alter the copyright statement on an item.

The full policy can be found here. Alternatively contact the University of Cumbria Repository Editor by emailing insight@cumbria.ac.uk.
FULL PAPER

Do prevalence expectations affect patterns of visual search and decision-making in interpreting CT colonography endoluminal videos?

Objective: To assess the effect of expected abnormality prevalence on visual search and decision-making in CT colonography (CTC).

Methods: 13 radiologists interpreted endoluminal CTC fly-throughs of the same group of 10 patient cases, 3 times each. Abnormality prevalence was fixed (50%), but readers were told, before viewing each group, that prevalence was either 20%, 50% or 80% in the population from which cases were drawn. Infrared visual search recording was used. Readers indicated seeing a polyp by clicking a mouse. Multilevel modelling quantified the effect of expected prevalence on outcomes.

Results: Differences between expected prevalence were not statistically significant for time to first pursuit of the polyp (median 0.5 s, each prevalence), pursuit rate when no polyp was on screen (median 2.7 s\(^{-1}\), each prevalence) or number of mouse clicks [mean 0.75/video (20% prevalence), 0.93 (50%), 0.97 (80%)]. There was weak evidence of increased tendency to look outside the central screen area at 80% prevalence and reduction in positive polyp identifications at 20% prevalence.

Conclusion: This study did not find a large effect of prevalence information on most visual search metrics or polyp identification in CTC. Further research is required to quantify effects at lower prevalence and in relation to secondary outcome measures.

Introduction

If we are expecting an event, we are more alert to it and more likely to react when it occurs.\(^1\) We might expect that radiologists are more alert to the presence of an abnormality when given an indication that prevalence is particularly high and, conversely, be less alert when the chance of encounter is believed to be low, as in screening.

Interpretation of medical imaging occurs in three environments: the symptomatic population, the asymptomatic/screening population and the research setting. Expected levels of abnormality vary considerably between these settings and between different medical specialities.\(^2\) It follows that the effect of varying prevalence of abnormality on image interpretation is crucial to our understanding of how diagnostic accuracy and interpretative performance might change across reporting environments.

In 2011, a systematic review\(^3\) found only three medical imaging studies\(^4-6\) that assessed the impact of experimentally modified prevalence on reader diagnosis. Subsequent studies have been published,\(^7-10\) but the relationship between prevalence and interpretation accuracy remains unclear. Some studies report increased false negatives or reduced diagnostic confidence at lower prevalence levels, for example, for interpretation of pulmonary arteriograms,\(^4\) mammograms\(^8,11\) or ankle trauma radiographs.\(^7\) This "rare target" effect has also been reported in non-clinical scenarios, such as baggage scanning\(^12,13\) and artificial target search experiments.\(^14\) By contrast, in chest radiography, the evidence for a prevalence effect on diagnostic accuracy is weaker,\(^5,9\) although
two studies that used eye tracking to monitor visual search of experienced readers suggested a possible association between increased prevalence and the duration and pattern of image scrutiny.\textsuperscript{16,17}

Despite increasing use of CT colonography (CTC) in routine practice, there is little research describing the effect of abnormality prevalence on diagnostic performance.\textsuperscript{8} This is surprising because CTC is commonly applied across a wide range of expected prevalence, from asymptomatic individuals undergoing screening\textsuperscript{16–18} to symptomatic and high-risk patients.\textsuperscript{19–21} Establishing the presence or absence of a prevalence effect on reader attention, visual search and diagnostic performance is important both in understanding how CTC should be used in clinical practice and for designing future research studies.

The purpose of this study was to assess the effect of expected abnormality prevalence on visual search and decision-making in CTC.

METHODS AND MATERIALS

Research ethics committee approval was obtained to record eye-tracking data from consenting observers in this prospective study. Institutional review board and research ethics committee approval was granted to use anonymous CTC data collated in previous studies.\textsuperscript{22,23}

Participants and cases

13 radiologists (readers) were recruited from a UK training hospital over 2 days in July 2012. All provided written, informed consent. Readers (6/13 males; mean age 32 years, range 27–36 years) were trainees with 1–7 years’ experience as a radiologist and at most 50 cases CTC experience.

10 CTC endoluminal fly-through videos lasting 30 s each were generated (EH, PP) with dedicated CTC software on a medical imaging workstation (Vitrea®; Vital Images, Inc., MN) and exported for viewing. Navigation speed was fixed at approximately 1.5 cm s\textsuperscript{-1}. Five videos depicted a single colorectal polyp (true positive, 5–8 mm maximal transverse dimension), verified by three radiologists with >200 cases’ experience.\textsuperscript{23} To counteract recall, cases were excluded if they contained polyps within 5 s navigation of the caecal pole, rectal ampulla or insufflation catheter, or contained other distinctive characteristics, assessed by a radiologist with 6 years’ experience (EH). Polyps were on screen for between 2.4 and 11.1 s. The remaining five videos (true negative) were selected from different sections of the colon, containing no polyps, in the same patient group.

The sample size was based on practical considerations: the number of readers available and the number of cases that could be assessed comfortably in one sitting. As the primary outcome measures have not been used before in this context, no power calculation was performed.

Data collection

The group of 10 videos was presented to each reader three times in one sitting, with an optional break between the groups. The order of cases was randomized for each reader. Before viewing each group, readers were told that the videos in that group came from a population with known prevalence of abnormality—20%, 50% or 80%. The ordering of the three prevalence scenarios was varied between readers using block randomization. Readers were not told that the three groups actually contained the same 10 videos repeated three times and were therefore unaware that the true prevalence was identical (50%) and the declared 20% and 80% prevalence levels were incorrect. Information given to readers was worded as:

“We are going to show you 3 groups of 10 videos in a random order.

Each group is taken from a different population, each with a different prevalence of abnormality.

Before each group we will tell you the population prevalence, either 80%, 50% or 20%.”

Readers were asked to hold a computer mouse throughout and indicate with a click (polyp identification) when they saw a lesion that they considered highly likely to represent a real polyp or cancer. Readers were not required to specify polyp location and could not pause, rewind or review videos. They were not told which videos contained polyps and were given no feedback about their performance. Data collection took 20–30 min per reader.

Viewing conditions

Reading was conducted in a quiet room with constant, ambient light. A liquid-crystal display monitor, 1280 × 1024 pixel resolution, was used (SyncMaster 971P; Samsung, Suwon, Republic of Korea and Fujitsu E19-5; Fujitsu, Tokyo, Japan; 1 pixel = 0.29 mm). The screen was positioned 60 cm in front of the reader. Videos measured 512 × 512 pixels (14.8 × 14.8 cm), representing a visual angle of 14.1°. The eye position of readers was recorded using a Tobii X50 or X120 eye tracker (Tobii Technology AB, Danderyd, Sweden), sampling at 50 or 60 Hz, respectively, positioned beneath the screen. No headrest was used. Readers wore glasses or contact lenses as normal. They performed a nine-point calibration procedure prior to data collection and were excluded if this could not be completed. They then viewed a supplemental warm-up video prior to data collection. They were not asked to fixate a particular point before each video.

Data preparation

The eye position data were prepared for analysis as described elsewhere;\textsuperscript{24} a summary follows. True-positive polyps were approximated using a circular region of interest (ROI), manually overlaid onto each video frame-by-frame by a medical image perception scientist (PP). The centre and radius of this ROI were adjusted manually to match the polyp’s transition across the screen. Within each frame, the perpendicular distance between the recorded eye position and the edge of the ROI was calculated and used in outcome measures described below. Eye gaze falling within a 50-pixel acceptance radius from the edge of the ROI was considered to be within high visual acuity. For periods when
no polyp was visible, the \((x, y)\) eye position co-ordinates were retained for analysis. Co-ordinates located >100 pixels outside the screen area were excluded as recording errors.

**Outcome measures**

Eye co-ordinate data were used to derive three primary and six secondary pre-specified outcomes (metrics) (Table 1). Figure 1 shows an example eye-tracking trace (distance between eye position and ROI over time) to illustrate metric definitions. Detailed information about metric derivations has been reported previously. Metrics reflected three aspects of reader behaviour: eye position when a polyp was on screen; eye position when no polyp was on screen; and frequency and accuracy of polyp identifications. Primary outcomes were time to first pursuit of the ROI; pursuit rate in the absence of an ROI; total number of polyp identifications. The “screen coverage” measure was defined by the proportion of eye gaze falling into three regions: within, above or below a 256 \( \times \) 256-pixel square at the centre of the screen. “Any correct identification” and the “polyp on screen” metrics are defined only for true-positive videos. “Any incorrect identification” is defined only for the period before any polyp appeared, to prevent readers who delayed their decision after seeing a polyp being misclassified as making a false-positive identification.

**Statistical analysis**

Metrics were analyzed using multilevel modelling, incorporating independent random intercepts for reader and video, including prevalence level as a factor. Effects of prevalence expectation were expressed relative to the true 50% prevalence category. In a planned sensitivity analysis, to test whether results were altered by the order (first, second or third viewing) in which the prevalence categories were presented, this order was included as an additional factor variable.

Within this multilevel framework, proportional hazards, logistic and Poisson models were used, as appropriate for the data type. As most viewings had at least one missing eye position data point, short missing data runs were imputed, based on the recorded eye co-ordinates immediately before and after, and adding random measurement error. Estimates were combined using multiple imputation methods with 10 imputations.

Cases with >50% missing values or >50 consecutive missing values were examined individually by two authors (TF, AP) and removed if deemed likely to make the metric calculation highly unreliable. The electronic Supplementary material contains more details.

A different approach was adopted only for pursuit rate, which has no generally agreed definition. We used the number of pursuits calculated by Tobii Studio v. 1.7.2 (50-pixel dispersion, 100-ms minimum time threshold) throughout the period when no polyp was on screen, divided by the duration of this period. Time points when the Tobii software failed to identify whether a co-ordinate belonged to any particular pursuit were excluded, and the time denominator adjusted accordingly. Cases with >50% missing values of the pursuit classifier were excluded from analysis.

---

Table 1. Metric definitions. The identifying letters A, B etc. refer to time points indicated in Figure 1

<table>
<thead>
<tr>
<th>Group</th>
<th>Name</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>Polyp on screen</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Time to first pursuit*</td>
<td>Time between appearance of polyp (A) and start of first pursuit of polyp (B)</td>
</tr>
<tr>
<td></td>
<td>Total assessment time span</td>
<td>Time between start of first pursuit of polyp (B) and polyp identification (E)</td>
</tr>
<tr>
<td></td>
<td>Assessment pursuit time</td>
<td>Cumulative time in pursuit of polyp before polyp identification (B–C and D–E), expressed as a proportion of the total time when the polyp was visible (A–G)</td>
</tr>
<tr>
<td></td>
<td>Assessment pursuit rate</td>
<td>Number of separate pursuits of polyp before polyp identification, divided by the total time when the polyp was visible before polyp identification (A–E)</td>
</tr>
<tr>
<td></td>
<td>Pursuit rate*</td>
<td>Number of distinct eye pursuits, divided by the total time when the polyp was off screen</td>
</tr>
<tr>
<td></td>
<td>Screen coverage</td>
<td>Proportion of eye co-ordinates falling in to each of three regions of the screen display, “upper”, “central” and “lower” (Figure 2)</td>
</tr>
<tr>
<td></td>
<td>Total number of identifications*</td>
<td>Number of identifications recorded over whole video</td>
</tr>
<tr>
<td></td>
<td>Any correct identification</td>
<td>Binary indicator of whether an identification occurred while the polyp was visible (a reaction time of 0.5 s after the polyp left the screen was allowed)</td>
</tr>
<tr>
<td></td>
<td>Any incorrect identification</td>
<td>Binary indicator of whether an identification occurred before the polyp was visible (or at any time, for true-negative videos)</td>
</tr>
</tbody>
</table>

*Primary outcome.
Results are presented as point estimates with 95% confidence intervals (95% CIs) and p-values. A 5% significance level was used, unadjusted for multiple testing.

Statistical analysis used STATA® v. 12.1 for Windows (StataCorp, College Station, TX) and R version 3.1.1.28

RESULTS
Eye tracking was successful and 389 of the intended 390 viewings were completed. Seven (1.8%) of these were omitted from the analysis of one or more metrics (with the exception of pursuit rate) because patterns of missing data made calculation unreliable. For pursuit rate, 37 (9.5%) of the viewings were excluded.

Table 2 summarizes metrics across all readers within each prevalence scenario. Of the videos that contained a polyp, readers made at least one pursuit of the polyp for 185 of the 190 (97%) viewings with reliable data.

There were no statistically significant differences between expected prevalence levels in any metric relating to visual search while the polyp was visible (Table 3). In each prevalence scenario, readers took approximately half a second on average to direct their gaze to the ROI after the polyp appeared [hazard ratio 1.32 (95% CI 0.95 to 1.93, \(p = 0.14\)) for 20% vs 50% prevalence; hazard ratio 0.95 (95% CI 0.64 to 1.40, \(p = 0.79\)) for 80% vs 50% expected prevalence; Tables 2 and 3, Figure 3].

Average total assessment time span, assessment pursuit time and assessment pursuit rate were also similar in the three prevalence scenarios (Tables 2 and 3).

During the period when the polyp was not on screen, the average pursuit rate was approximately 2.7 pursuits per second at each of the three prevalence levels (Table 2), with no statistically significant differences (Table 3). There was a tendency for readers’ gaze to fall inside the central region of the screen less often at the 80% prevalence level than at the 50% prevalence level [odds ratio 0.82 (95% CI 0.72 to 0.95, \(p = 0.008\)), Table 3], with a concomitant increase in the upper region. This effect, however, was small, with on average 82% of gaze points falling in the central region at 80% prevalence compared with 84% at 50% prevalence (Table 2).

There were no statistically significant differences with respect to expected prevalence regarding the total number of identifications (Table 3). As expected, the average number of identifications was higher for videos that contained polyps than for those that did not (1.3 vs 0.4, Table 2). The sensitivity, or probability of a polyp being correctly identified, was higher at 50% prevalence (86%) than at 20% prevalence (71%). This difference was statistically significant (\(p = 0.01\), Table 3) but the trend did not persist at the 80%
prevalence level (75%). This metric was subject to an extremely high case-specific effect (Figure 4), as in three videos 1, 2 and 4 almost every reader identified the polyp at each prevalence level; the other two videos 3 and 5, for which the polyp was superficially more difficult to identify, are therefore likely primarily responsible for the differences in rates of correct identification.

Table 3. Comparison of metrics between prevalence levels: hazard ratio (HR), odds ratio (OR) or rate ratio (RR), as appropriate, with 95% confidence interval (CI) and p-value

<table>
<thead>
<tr>
<th>Metric</th>
<th>Measure</th>
<th>20% vs 50% prevalence</th>
<th>80% vs 50% prevalence</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Effect size (95% CI)</td>
<td>p-value</td>
</tr>
<tr>
<td>Time to first pursuit</td>
<td>HR</td>
<td>1.32 (0.95–1.93)</td>
<td>0.14</td>
</tr>
<tr>
<td>Total assessment time span</td>
<td>HR</td>
<td>0.74 (0.50–1.12)</td>
<td>0.15</td>
</tr>
<tr>
<td>Assessment pursuit time</td>
<td>OR</td>
<td>1.27 (0.87–1.84)</td>
<td>0.22</td>
</tr>
<tr>
<td>Assessment pursuit rate</td>
<td>RR</td>
<td>0.91 (0.70–1.18)</td>
<td>0.47</td>
</tr>
<tr>
<td>Pursuit rate</td>
<td>RR</td>
<td>1.01 (0.98–1.05)</td>
<td>0.39</td>
</tr>
<tr>
<td>Screen coverage</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Upper</td>
<td>OR</td>
<td>0.93 (0.78–1.12)</td>
<td>0.45</td>
</tr>
<tr>
<td>Central</td>
<td>OR</td>
<td>1.06 (0.92–1.23)</td>
<td>0.39</td>
</tr>
<tr>
<td>Lower</td>
<td>OR</td>
<td>0.96 (0.81–1.13)</td>
<td>0.63</td>
</tr>
<tr>
<td>Total number of identifications</td>
<td>RR</td>
<td>0.81 (0.62–1.06)</td>
<td>0.12</td>
</tr>
<tr>
<td>Any correct identification</td>
<td>OR</td>
<td>0.24 (0.08–0.73)</td>
<td>0.01</td>
</tr>
<tr>
<td>Any incorrect identification</td>
<td>OR</td>
<td>0.66 (0.37–1.19)</td>
<td>0.17</td>
</tr>
<tr>
<td>Videos with polyps</td>
<td>OR</td>
<td>0.86 (0.35–2.11)</td>
<td>0.75</td>
</tr>
<tr>
<td>Videos without polyps</td>
<td>OR</td>
<td>0.53 (0.24–1.17)</td>
<td>0.11</td>
</tr>
</tbody>
</table>

*Kaplan–Meier estimate, calculated without allowing for clustering, excluding viewings with immediate pursuit.*
The probability of an incorrect identification (false positive) ranged from 30% at 20% prevalence to 39% at 80% prevalence; this difference was also not statistically significant (Table 3). On average, incorrect identifications occurred with similar frequency for videos that contained no polyps and for videos that contained polyps during periods when the polyp was not visible, although there was considerable variability between cases (Figure 4). Some false-positive features were identified with a mouse click by several readers (e.g. Case 3 at 5 s, Figures 4 and 5).

In sensitivity analysis, including as an extra factor variable, the order in which the prevalence scenarios were presented did not affect the prevalence effect sizes shown in Table 3.

DISCUSSION

This study investigated the effect on visual search and decision-making for CTC of providing readers with substantially different expectations of the likely prevalence of abnormality in the population from which cases were drawn. We did not demonstrate a strong link between prevalence expectation and the pattern of search or decision-making.

Our conclusion differs from those of several studies using scenarios other than CTC that found increased false-negative rate at lower prevalence levels. Our study showed a statistically significant increase in the proportion of polyp identifications between 20% and 50% expected prevalence, but for three reasons this finding should be treated cautiously. First, it did not extend to the highest prevalence level, for which the proportion was similar to that at 20%, and a non-monotonic relationship seems implausible.

Second, the effect was driven by an increased true-positive rate in just two of the five cases with polyps; a consistent increase across all cases, which would have provided more convincing evidence, was not observed. Third, this was just one of several secondary analyses performed, and so it may be a chance result.

The existence of a prevalence effect is not a universal finding in image interpretation studies. For example, Gur et al found that varying prevalence levels between 2% and 21% did not affect the diagnostic accuracy of chest radiograph assessment. Likewise, we did not find a prevalence effect for our three primary outcomes, which were chosen to represent visual search and decision-making. Modality may therefore be an important determinant of prevalence effects.

We have shown previously that time to first pursuit of the polyp changes with reader experience and the presence of a computer-aided detection marker; in the present study, this metric was unchanged across prevalence scenarios. When no polyp was visible, readers tended to spend more time, proportionally, looking at peripheral screen regions in the 80% prevalence condition, but this effect is small and is not supported by changes in other visual search metrics. However, the finding requires further investigation as our measure is based on a simple square at the centre of the screen area, which may not adequately capture gaze narrowing effects.

In sensitivity analysis, including as an extra factor variable, the order in which the prevalence scenarios were presented did not affect the prevalence effect sizes shown in Table 3.

We used a common set of cases for each of the prevalence conditions to directly observe the effect of disclosing different prevalence information, as opposed to the effect of the true case mix. Lau et al13 claim that the latter may have a larger effect on decision-making, but testing this was not our...
objective. Indeed, it would have been infeasible for readers to make an assessment of the true underlying prevalence within a realistic time frame. It is possible that some readers realized that they had viewed videos more than once, but this is unlikely to have a major effect on our findings; the order in which the prevalence conditions were presented was determined randomly and this order was not strongly associated with outcomes. Enabling all cases to be viewed with comfort in a single sitting was an important practical consideration in our choice of the number of cases used. Despite the number of cases being moderately small, repeated viewings of the same case under different prevalence conditions enabled quantities of interest to be estimated with acceptable precision.

Future studies should assess further the possibility of a threshold effect in CTC. It is possible that the expected prevalence level needs to be <20% for an effect to be visible, as is usually the case in everyday clinical practice, except in very high-risk patient groups such as those examined following a positive faecal occult blood test. Evans et al found a marked reduction in sensitivity for breast cancer diagnosis using mammography during screening when the prevalence was extremely low (0.3%). Whether a similar effect applies to CTC remains unknown. Additionally, prevalence effects may vary according to the ease of visualization and identification of the cases chosen.

This study has limitations. This study was exploratory in nature, and therefore we may not have used enough cases for subtler prevalence effects to be detected. Endoluminal fly-through view was presented in automatic mode only, so readers could not adjust navigation speed as in usual practice. We were therefore unable to assess the effect of prevalence on the time the reader would spend scrutinizing each video; from laboratory experiments and some clinical studies, there is evidence that assessment time is affected by prevalence in static viewing modes. Mouse clicks are not synonymous with definitive decisions about the presence of polyps and thus can only be regarded as proxy measures of diagnostic accuracy. Readers were not asked to identify polyp locations and so, even with eye-tracking data, it is impossible to state with certainty the cause of any particular click. Readers were inexperienced in CTC, and so our findings are not directly generalizable to experienced radiologists using CTC in day-to-day clinical practice. Finally, we did not assess the effect of providing information about the spectrum of disease severity, since readers received prevalence information alone.

In summary, CTC readers were provided with different estimates of the prevalence of abnormalities from which cases were drawn, and study results did not demonstrate a strong link between prevalence information and the pattern of visual search or decision-making. Further research should investigate effects at lower prevalence levels, such as might be present in asymptomatic populations.

**FUNDING**

This work was supported by the UK National Institute for Health Research (NIHR) under its Program Grants for Applied Research funding scheme (RP-PG-0407-10338). A proportion of this work was undertaken at the University College London and University College London Hospital, which receive a proportion of funding from the NIHR Biomedical Research Centre funding scheme. The views expressed are those of the authors and not necessarily those of the National Health Service, the NIHR, or the Department of Health.

**REFERENCES**
