2 The sensitivity and specificity of tests that locate multiple lesions; evaluating colonography


Abstract

There is ambiguity in the estimation of sensitivity and specificity in studies evaluating colonography in the screening for colon cancer. This ambiguity is due to conventional analysis methods that do not take into account multiple findings that can be obtained by colonography as well as by colonoscopy, the reference standard. In this paper the pitfalls of some frequently used methods will be discussed. In addition a new approach will be outlined that can be used to estimate sensitivity and specificity from such data. This method estimates the sensitivity and specificity, respectively, as the probability of correctly identifying true positive and true negative patients. Relative to other methods, this method takes all findings into account, including false positive and false negative lesions, as such no information is lost.
The sensitivity and specificity of tests that locate multiple lesions; evaluating colonography

Background
When evaluating a diagnostic test one is interested in how well the test is able to distinguish patients with the target condition from those without that condition. Different measures of test performance are available to express a test's performance. Well known properties are the sensitivity and specificity of a test. These two measures indicate the proportion of patients correctly classified by the index test in the population of diseased and non-diseased persons, respectively, as diagnosed by the reference standard.

In addition to making a distinction between diseased and non-diseased persons, tests can also be used to document the size, severity, location or number of target lesions. An example can be found in the evaluation of CT or MRI (2D or 3D) colonography (virtual colonography). Each technique can result in the detection of multiple polyps. In the analysis of such studies, lesions detected on colonography are compared with the findings of conventional colonoscopy. Such comparisons may or may not result in a match, whenever lesions are found on both tests.

It is not clear how the number of matches and mismatches can be transformed into one of the more conventional test statistics, such as sensitivity and specificity. By discussing various methods used in the literature to analyse data from such colonography evaluations, we will show that the estimation of sensitivity and specificity is not without pitfalls. In addition we present a new method that can be applied to this kind of data. The application of these methods will be illustrated with data from a recently performed study to evaluate the test characteristics of CT colonography (CTC) at different levels of radiation dose. In this study 50 patients at risk for colorectal cancer were evaluated for the presence of one or more polyps (or other findings e.g neoplasms). Colonoscopy was used as the reference standard. Herein only the data at mAs levels of 100 will be used and lesions of all sizes will be considered. Detailed information can be found elsewhere.

Methods

Per-polyp analysis
In many studies of CTC a per-polyp based approach is used in the evaluation, in which the polyps are the unit of analysis. In these studies, the researchers compare polyps found on conventional colonoscopy with the findings of CTC, estimating the per-polyp sensitivity of the latter.

In our study, for example, colonoscopy revealed 62 polyps found in 27 patients. Of these polyps 23 could also be detected by CTC. The sensitivity is
then estimated as 23/62 or 37% (table 1), indicating a probability of 0.37 that a polyp detected by colonoscopy can also be detected by CTC. Alternatively, one can start from the polyps found on CTC and see whether these can be verified by conventional colonoscopy, this way calculating the per-polyp positive predictive value.

One must consider that the estimated sensitivity can be biased and its statistical precision overstated because of clustering of polyps within patients. Such clustering can lead to correlation of diagnostic on multiple polyps in the same patient. In addition, individual patients with multiple polyps would effectively carry more weight in the calculation of sensitivity relative to patients with only few polyps. A correction for clustering of polyps within patients is needed to guarantee that k polyps in a single patient have less weight in the estimation of sensitivity than a series of single polyps in k patients. The per-polyp specificity of the test is a lot harder to define. In principle there is an infinite number of negative results: each potential location of a polyp in the colon. As a consequence the per-polyp specificity and negative predictive value are undefined.

**Table 1** The sensitivity and specificity as estimated by different methods. Data regarding polyps of all sizes.

<table>
<thead>
<tr>
<th>Method</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Uncorrected</td>
<td>Corrected</td>
</tr>
<tr>
<td></td>
<td>% (95% CI)</td>
<td>% (95% CI)</td>
</tr>
<tr>
<td>Per-polyp</td>
<td>37 (25 to 49)</td>
<td>37 (26 to 49)</td>
</tr>
<tr>
<td>Per-patient</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unmatched</td>
<td>93 (83 to 100)</td>
<td>-</td>
</tr>
<tr>
<td>FFP Conditional on index test</td>
<td>89 (74 to 100)</td>
<td>-</td>
</tr>
<tr>
<td>FFP Conditional on reference test</td>
<td>59 (41 to 78)</td>
<td>-</td>
</tr>
<tr>
<td>Combined per-polyp/per-patient</td>
<td>37 (25 to 49)</td>
<td>37 (25 to 49)</td>
</tr>
<tr>
<td><strong>FROC approach</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Per-polyp</td>
<td>49 (11 to 86)</td>
<td>-</td>
</tr>
<tr>
<td>Per-patient</td>
<td>79 (24 to 99)</td>
<td>26 (19 to 36)</td>
</tr>
</tbody>
</table>

*In the corrected results, clustering of polyps within patients is taken into account.*

**Per-Patient analysis**

A per-patient analysis is often done as an alternative. In this approach the patient is the unit of analysis. A patient is classified as positive by the reference test if at least one polyp of a pre-specified size is present and negative if no such lesion is detected. A true positive patient is a patient with at least one polyp detected by the reference standard and at least one lesion on the index test, regardless its location. In our study 27 of the 50 patients were found to have at least one polyp...
The sensitivity and specificity of tests that locate multiple lesions; evaluating colonography

by colonoscopy. Of these patients 25 also had at least one lesion at CTC, resulting in a per-patient sensitivity of 93% (25/27). Of the remaining 23 patients without lesions detected at colonoscopy 6 were also negative on CTC, giving a specificity of 26% (6/23).

This analytic approach is relevant in the context of clinical practice. CTC is mainly propagated as a screening tool and, if any lesion is found, that patient is subsequently scheduled for conventional colonoscopy, a technique that also allows removal of the polyps.\textsuperscript{10} The disadvantage of this approach is that all patients with a lesion on CTC and a polyp found by colonoscopy are counted as true positives. It is possible that these two lesions are found at different sites in the colon. For convenience we call this 5\textsuperscript{th} category of patients, the 'floating false positives' (FFP). Counting these FFP’s as true positive patients may seem arbitrary, as they reflect a lack of correspondence between findings on the index test and those of the reference standard. If within the study design lesions have not been matched, the FFP’s are inevitably regarded as true positive patients. If lesions have been matched, based on size and location, there are three cells within the 2x2 table possible for the FFP’s. In our study there were nine FFP patients, thus nine patients with on both test at least one lesions, but where none of these lesions matched. One possibility is regarding the lesions as unmatched, hence interpreting the FFP’s as discussed above.\textsuperscript{14,15} Alternatively, if one conditions on the results of the index test, that is by fixing the row totals, the FFP’s are considered as false positives (see table 2).

<table>
<thead>
<tr>
<th>Reference standard</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Index</td>
<td>test</td>
<td>pos</td>
</tr>
<tr>
<td>pos</td>
<td>16</td>
<td>T P</td>
</tr>
<tr>
<td>neg</td>
<td>2</td>
<td>F N</td>
</tr>
<tr>
<td>18</td>
<td>32</td>
<td>50</td>
</tr>
</tbody>
</table>

Estimated sensitivity = \( \frac{16}{18} = 0.89 \), specificity = \( \frac{6}{32} = 0.19 \).

If one conditions on the reference standard results (fixing the column totals), the FFP’s are considered as false negatives (see table 3).
The cell in which the FFP patients are put clearly affects the sensitivity and specificity (see table 1). The larger the proportion of FFP’s, the larger the difference in the value of these accuracy estimates.

An illustration of the ambiguity that can follow from these FFP results can be found in the report of a study performed by Fenlon and colleagues. In that study CTC and conventional colonoscopy were compared in 100 patients of whom 49 had polyps. The following results can be found in the paper: 42 true positives, 41 true negatives, 8 false positives and 9 false negatives. Adding up the these totals gives us 51 patients with polyps, not 49 as given in the beginning of the results section of that paper. This discrepancy is probably due to 2 FFP patients. Since the authors did not specify whether this type of positives were present and, if so, to which cell they were assigned, the way in which the results are presented impedes a clear interpretation of the test’s properties.

**Per-Segment analysis**

The disadvantage of the per-patient approach is that the multiple findings, and their location in a single patient are ignored, resulting in a loss of information in the analysis. Patients with multiple false positive results are treated in the same way as patients with a single false positive result without affecting the estimated specificity of the test.

Some authors propose to divide the organ of interest in a number of segments, - the ascending colon for example - and take the segments as the unit of analysis. This approach has been used, for example in studies of the performance of diagnostic modalities in staging colorectal cancer, detecting liver metastases, staging prostate cancer, detecting musculoskeletal cancer, and MR Angiography. In each patient, every segment is scored for the existence of polyps. A true positive segment is a segment with at least one positive lesion both on colonoscopy and on
The sensitivity and specificity of tests that locate multiple lesions; evaluating colonography

CTC. A true negative segment is one in which nothing is seen on both tests. The sensitivity can then be estimated as the number of positive segments as found on CTC relative to with number of segments with at least one polyp on colonoscopy. The specificity is estimated as the proportion of negative segments found by colonoscopy that were also negative on CTC.

The problem of FFP has not disappeared in this approach. The proportion of FFP segments now depends on the number of segments, with the probability of FFP decreasing with an increasing number of segments. Increasing the number of segments will increase the number of true negative segments and this will inflate the specificity. In addition cluster dependency will exist between segments from the same patient and has to be corrected for. The design of our study did not include the recording of segment specific information, so we cannot report on this kind of analysis.

**Combined per-polyp/per-patient analysis**

To bypass the impossibility of estimating a per-polyp specificity some authors have turned to a mixed approach: presenting a per-polyp sensitivity and a per-patient specificity.

If specificity is calculated as the proportion of patients without lesions in the group of patients who test negative for polyps on colonoscopy then the issue of FFP patients is avoided. FFP’s affect the sensitivity in a per-patient approach. Although definition and interpretation of the single sensitivity and specificity are straightforward in this mixed approach, one has to be aware that this strategy neglects a key aspect of diagnostic test evaluation: the close relation and implicit threshold balance between sensitivity and specificity. Likelihood ratio’s, well known in the Bayesian updating of prior probabilities, cannot be derived from these sensitivity and specificity values, where one is based on a per-patient analysis and the other is polyp based.

**FROC methodology**

About fifteen years ago Chakraborty introduced a method within the field of radiology to estimate free-response receiver operating characteristics (FROC) curves for tests that are able to locate multiple findings. The FROC method was originally developed for observer performance studies of imaging tests where the observer classifies each image with one of several (usually five) confidence ratings. These ratings reflect the observer’s confidence that the image is abnormal. This FROC method can be adapted to estimate sensitivity and specificity of tests that can locate multiple lesions whenever dichotomous (positive or negative) outcomes are preferred.
The FROC approach estimates a false positive rate, from which a specificity can be calculated. This approach assumes that the number of false positives follows a Poisson distribution, in other words, the probability of a specific number \( x_i \) of false positives in the i-th patient is given by the equation:

\[
Pr(X_i = x_i) = \frac{e^{-\mu} \mu^{x_i}}{x_i!}
\]

where the parameter \( \mu \) represents the expected number of FP lesions per patient. When there are \( n \) patients, the rate parameter \( \mu \) is estimated as

\[
\frac{1}{n} \sum x_i
\]

The specificity is then defined as the probability that no false positives are observed in a patient \( Pr(X_i=0) \). The probability that this occurs is equal to \( Pr(X=0)=e^{-\mu} \).

The assumption of a common FP rate across patients is somewhat too restrictive, and a common method to relax it is to assume that the rate parameter varies between patients, in which case:

\[
Pr(X_i = x_i) = \frac{e^{-\mu_i} \mu_i^{x_i}}{x_i!}
\]

The variation in the FP rates \( \mu_i \) between patients can be described by a distribution \( g(\mu) \). In this way the mean specificity is defined as \( Pr(X=0)=\int e^{-\mu} g(\mu) d\mu \). This integral can be interpreted as a weighted average of the specificity values across patients, with the weights given by the distribution function \( g(\mu) \). This function is unknown but can be estimated from the data. In order to do so, choices must be made about the functional form of \( g(\mu) \). Most common choices for \( g(\mu) \) are the gamma distribution, but we prefer a fully non-parametric distribution. The latter approach takes cluster variation of polyps between patients into account and is preferable in the presence of variability. The approach just presented allows the estimation of a per-patient specificity using all information available. As it is based on a joint analysis of all patients, those with and without proven polyps, the estimated per patient specificity is automatically corrected for FFP patients.

A somewhat similar approach can be used to model the sensitivity. Assume that the number of true positive polyps \( Y_i \) in patient i, having \( k_i \) polyps as defined by colonoscopy follows a binomial distribution with specificity \( P_i \). If the between-patient variation in \( P_i \) is described by a nonparametric distribution \( g(p) \), the probability of a true positive polyp in a randomly selected patient - the per-polyp sensitivity - can be expressed as a similar weighted average \( \int g(p) dp \).
The sensitivity and specificity of tests that locate multiple lesions; evaluating colonography

Notice that the per-patient sensitivity can be written as a function of the per-polyp sensitivity (\(\phi\)) and the mean number of polyps (\(K\)) per patient which approximately follows the binomial distribution: \(1-(1-\phi)^K\). In our data \(K=62/27=2.3\) and \(\phi=0.49\). The per-patient sensitivity can then be estimated as 0.79 (see table 1).

A disadvantage of this approach is that it analyzes sensitivity and specificity separately, without accounting for the tradeoff between the two quantities. The computations are somewhat complex, though available through routines in packages like SAS and Egret. A computer program that simultaneously estimates the specificity and sensitivities can be obtained free of charge from the first author.

**Discussion**

We have discussed various methods to express the results of evaluation studies of colonographic techniques. These methods can produce widely diverging estimates of test performance, as can be inferred from table 1. The variety of methods can be attributed to the strong inclination of investigators to force the results of such studies in a 2x2 table and to calculate sensitivity and specificity. With tests that are able to detect and locate multiple lesions, the binary patient-based nature of these measures is lost and the classical 2x2 table looses its straightforward appeal. Only with additional assumptions and simplifications can one arrive at a 2x2 table. However, the resulting sensitivity and specificity may no longer express a test’s true diagnostic performance.

In this paper, we have introduced a convenient approach that is able to incorporate all lesions detected in the analysis. As such our approach bypasses ambivalent FFP patients. If a random effects model is considered it directly takes into account dependency between lesions found in the same patient. However, the modified FROC approach analyzes sensitivity and specificity separately, without accounting for the tradeoff between the two quantities.

As far as we know, FFP patients have never been reported in the articles that used a per-patient based approach to estimate sensitivity and specificity. This omission might have led to ambiguity of results, and a hampered critical appraisal by interested readers. In addition it raises the question whether the today’s impression of the diagnostic performance of virtual colonography is correct. As a promising technique, virtual colonography deserves a proper evaluation. This can only happen if investigators look critical at the methodology used and turn to clear and complete reporting of their results.
References


2 The sensitivity and specificity of tests that locate multiple lesions; evaluating colonography


