Advancements in colorectal imaging
Kuiper, T.

Citation for published version (APA):
Kuiper, T. (2013). Advancements in colorectal imaging
Chapter 2

Study designs to compare new colonoscopic techniques: clinical considerations, data analysis, and sample size calculations

F.J.C. van den Broek and T. Kuiper, E. Dekker, A.H. Zwinderman, P. Fockens, J.B. Reitsma

Submitted
ABSTRACT

Background & aims
Novel imaging techniques need to be evaluated for their ability to improve the detection of polyps. Critical appraisal of reported studies reveals remarkable differences in study designs, despite their similar objectives. The aim of the current study was to compare frequently used study designs for their required sample size to detect relevant differences in polyp detection rates.

Methods
We compared three commonly reported study designs: the parallel randomized design (1), the randomized cross-over design with direct removal of polyps (2) and the randomized cross-over design without direct removal of polyps (3). A total of 5 different scenarios were analyzed per study design, representing a variety of clinical settings. Each scenario was repeated a 1000 times for each study design and the sample size that produced a significant result per study design in each scenario is reported.

Results
In many scenarios, study design 1 required 10 to 15 times more patients to reach the same statistical power of 80% compared to study design 2. Further reductions in sample size could be achieved if study design 3 was performed, although this design is limited by its impracticality.

Conclusion
The statistical power of the randomized parallel design to detect differences in detection rates between techniques is low. Researchers should carefully consider whether the cross-over design can be used instead. The cross-over design has much greater power, although it may be more cumbersome for patients, endoscopists, and researchers.
INTRODUCTION

Optical colonoscopy is considered the gold standard for detection of premalignant colorectal polyps.\(^1\) Polyps are frequently overlooked during standard colonoscopy however, and polyp miss-rates are estimated to be as high as 26% for diminutive lesions.\(^2\) In order to increase polyp detection, several imaging techniques have been developed.\(^3, 4\) In addition, novel imaging techniques can also aid the endoscopist in the differentiation of polyps.\(^5-8\)

The differentiation performance of any imaging technique can be assessed by using the classical accuracy design for evaluating diagnostic tests.\(^9\) This design consists of a comparison between the results of the technique under evaluation and a high-quality reference standard (i.e. histopathology). The reporting of such studies can be guided by the Standards for Reporting of Diagnostic Accuracy (STARD) initiative.\(^10, 11\)

Detection studies on the other hand are methodologically challenging as they lack a high-quality reference standard against which the new technique can be compared. Furthermore, methodological papers discussing the detection ability of imaging techniques are lacking. As a consequence, detection studies reveal remarkable differences. For example, detection studies on narrow-band imaging (NBI) have used different study designs and different statistical methods to analyze their data, showing large variation in sample sizes.\(^12-15\) So, despite their similar objectives, results from these studies are difficult to compare.

The aim of the current study was to compare frequently used study designs in detection studies of novel imaging techniques. Through a series of simulations we compared these designs for their required sample sizes to detect clinically relevant differences in polyp detection rates. Also, designs were compared with respect to the underlying clinical question they try to answer.

METHODS

Study designs

A total of three frequently used study designs were included: the parallel randomized design, the randomized cross-over design with direct removal of polyps and the randomized cross-over design without direct removal of polyps (figure 1).

In the parallel randomized design (1), patients are randomized to undergo an examination with either technique A or B.\(^12, 15-23\) In the second study design, the randomized cross-over design with direct removal of polyps (2), patients undergo back-to-back examinations with both techniques, but the order in which they receive both techniques is determined by randomization.\(^24-28\) Polyps detected by the first technique (either A or B) are removed immediately. Consequently, only missed polyps can be picked up during the second examination. The third study design is the randomized cross-over design without direct removal of polyps.
(3), which is similar to the second study design, although polyps are not removed during the first examination but rather during the second examination. This study design allows a comparison between the techniques for each polyp within a patient.\textsuperscript{15, 29}

Results of detection studies can be analyzed on a \textit{per-polyp} or a \textit{per-patient} basis. The former has the advantage of having more statistical power, since each polyp will generate information about the detection quality of a technique. Moreover, detecting each individual polyp has clinical significance as each polyp may be premalignant. Therefore, the number of detected polyps per patient was used as the primary outcome measure in our assessment of each study design.

\textbf{Simulations}

In each of the three designs two colonoscopy techniques (A and B) were compared. Technique A represented the conventional colonoscopy technique while technique B represented the novel technique. A total of 5 different scenarios (I-V) were analyzed representing a variety of clinical settings and differences between the two techniques which were characterized by 2 parameters: the number of polyps within each patient and the detection ability of each technique (table 1). Within each scenario, a comparison was made between the three designs across a range of sample sizes. Each scenario was repeated a 1000 times for each study design and the sample size that produced a significant result for each of the three study design in each scenario was reported.

The first parameter of the simulation (the number of polyps within each patient) was obtained by using the negative binomial distribution in order to generate the true, but
unobserved number of polyps within each patient. The negative binomial distribution is closely related to the Poisson distribution, but it has the flexibility to simulate situations. The number of patients with no polyps and the number of patients with many polyps is often higher than expected under the Poisson distribution, which is known as overdispersion. The negative binomial distribution is characterized by a mean and the overdispersion parameter. In the scenarios I, II and III the parameters of the negative binomial distribution were chosen such that the observed mean number of polyps would be 1 and that in 50% of the patients no polyp would be identified with conventional colonoscopy. Such numbers are often reported in the literature.\textsuperscript{30, 31} In the remaining scenarios IV and V the parameters of the negative binomial distribution were chosen similar, but the observed mean number of polyps was now set at 2, a number that has also been reported often.\textsuperscript{32, 33}

The second parameter of the simulation (the detection ability) was specified as the detection ability for each technique when applied during the first inspection (pA\textsubscript{1} or pB\textsubscript{1}) and the ability of each technique when applied in the second inspection to detect additional polyps missed by the other technique (pA\textsubscript{2} or pB\textsubscript{2}, needed in design 2 and 3 only). The detection ability of both techniques was assumed to be lower in the second round than in the first round as missed polyps are presumed more difficult to detect.

The detection rate of technique A (representing conventional colonoscopy) was set at 100% minus reported polyp miss-rates derived from a systematic review.\textsuperscript{2} Therefore, the detection rate of A was set at 75% (scenarios I, IV and V) or 85% (scenarios II and III) during the first examination. In order to include a range of differences between technique A and technique B, technique B was set at having a 10% (scenario I, II and III), a 15% (scenario IV) or a 20% (scenario V) higher detection rate compared with technique A.

In order to assess the frequency of a type I error (finding a statistically significant result when in reality there is none), two additional scenarios were simulated in which no difference in detection rate was present between technique A and technique B.

\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|c|c|}
\hline
Scenario & Mean & PA\textsubscript{1} & PB\textsubscript{1} & PA\textsubscript{2} & PA\textsubscript{2} \\
\hline
Scenario I & 1 & 0.85 & 0.95 & 0.70 & 0.80 \\
Scenario II & 2 & 0.75 & 0.90 & 0.50 & 0.60 \\
Scenario III & 1 & 0.75 & 0.95 & 0.50 & 0.60 \\
Scenario IV & 2 & 0.85 & 0.95 & 0.70 & 0.80 \\
Scenario V & 2 & 0.85 & 0.95 & 0.70 & 0.70 \\
\hline
\end{tabular}
\caption{The 5 different scenarios used in each simulation for the three study designs}
\end{table}
RESULTS

Parallel randomized design (design 1)
To explain our simulations and the results in more detail, each study design is illustrated by showing the intermediate results from one randomly selected, single simulation out of the 1000 simulations. The simulation example in table 2 is derived from scenario 1 using study design 1. In this example a sample size of 200 is used, meaning that 100 patients were randomized to examination by technique A (with a true detection rate of 85%) and 100 patients by the new technique B (with a true detection rate of 95%).

To test whether more polyps were detected in the group examined with the new technique B, the non-parametric Wilcoxon statistic was used. In this particular example, the result was a non-significant p-value of 0.59. This is a typical value for the parallel randomized design in this scenario because the power of a parallel design with a sample size of 200 is only 10%. To achieve a power of 80%, the total sample size would have to be as high as 3,980 (table 5, first row).

Randomized cross-over with removal of polyps (design 2)
The same scenario is now applied in a study using a randomized cross-over design with removal of polyps (design 2). The results of a single simulation are described in table 3. The conditional detection capacity (i.e. as observed in the second round) of technique A was assumed to be 70% and 80% for technique B.

In this design, the ratio of the number of polyps detected in the first round divided by the total number of detected polyps in both rounds were calculated for each study arm. A higher ratio indicates that the initial technique has better detection capabilities. A chi-square test was used to test for a difference in the two proportions. In comparison with the parallel randomized design in the same scenario with a similar sample size, the result with this design was significant with a p-value of 0.015. To obtain a power of 80% with this design a sample size of 280 was required (table 5, first row).

<table>
<thead>
<tr>
<th># of polyps</th>
<th>technique A (n=100)</th>
<th>technique B (n=100)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>49</td>
<td>46</td>
</tr>
<tr>
<td>1</td>
<td>27</td>
<td>28</td>
</tr>
<tr>
<td>2</td>
<td>16</td>
<td>12</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>14</td>
</tr>
<tr>
<td>≥4</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>86</strong></td>
<td><strong>94</strong></td>
</tr>
</tbody>
</table>

Mean # per patient: \[0.86, 0.94\] \(p\)-value: 0.59

* Scenario 1: Observed mean number of polyps per patient = 1
  Detection rate of technique A during first examination = 0.85
  Detection rate of technique B during first examination = 0.95
Randomized cross-over without removal of polyps (design 3)

The results of this design when applied in the same scenario are summarized in table 4. In design 3, each detected polyp is either detected by one technique only (thus missed by the other) or detected by both techniques. The results of the 200 patients can be summarized in a 2*2 table where a difference in detection capabilities will generate higher numbers of discordant results (e.g. missed by one technique, detected by the other). The difference in discordant results was analyzed by the McNemar test for paired observations, leading to a p-value of 0.0027. Power was further increased in comparison to study design 2 because this is a true paired analysis on each detected polyp. To obtain 80% power with this design in this scenario, 125 patients would be required (table 5, first row).

<table>
<thead>
<tr>
<th></th>
<th>Detected technique B</th>
<th>Not detected technique B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>total number of polyps</td>
<td>102</td>
</tr>
<tr>
<td># polyps first inspection</td>
<td>86</td>
<td>94</td>
</tr>
<tr>
<td># missed first inspection</td>
<td>16</td>
<td>7</td>
</tr>
<tr>
<td># polyps second inspection</td>
<td>12</td>
<td>3</td>
</tr>
<tr>
<td># polyps missed by both techniques</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Ratio (polyps first round / polyps first + second round)</td>
<td>86/(86+12) = 88%</td>
<td>94/(94+3) = 97%</td>
</tr>
</tbody>
</table>

**Scenario 1:** Observed mean number of polyps per patient = 1
Detection rate of technique A during first examination = 0.85
Detection rate of technique B during first examination = 0.95

Sample size required for each study design

The final results of all our simulations are summarized in table 5. For each study design, all five scenarios were repeated a 1000 time across a range of sample sizes and the sample size that produced a significant result in about 80% of the simulations is reported. In many scenarios, study design 1 required 10 to 15 times more patients to reach the same statistical power of 80% compared to study design 2. Further reductions in sample size could be achieved if study design 3 was performed.

Results of the two additional scenarios in which there was no difference in detection rate between technique A and technique B are described in table 6a & 6b. In the simulations of
these two scenarios we report the number of significant results when using a total sample size of 200 and 800. The frequency of significant results (type I error) was close to the expected nominal value of 5% in both scenarios for all three study designs.

**DISCUSSION**

The current study reports on three study designs that are frequently used in detection studies of novel endoscopic techniques. Our results illustrate that a cross-over study design with direct removal has greater power to identify differences in polyp detection compared with the randomized parallel design. In many scenarios, the randomized parallel design required 10-15 times more patients than the cross-over design with direct removal to reach the same statistical power of 80%.

In the randomized parallel design the expected number of polyps will, on average, be the same in both study arms. Therefore, results are unbiased, meaning that if such a trial would be repeated many times, the average result would be similar to the true effect. In a single trial, however, the number of polyps at baseline can differ between the two arms just by chance.
Such differences have a strong impact on the final results because the presence of more polyps at baseline directly means that more polyps can be detected in that arm. Even though the number of polyps in patients at baseline may partially be predicted by patient age, gender, and indication, imbalances in the number of polyps at baseline cannot be avoided as the presence of polyps in patients is not fully understood. Consequently, results from one trial to another will vary just because of random noise, and sample sizes need to be sufficiently large to overcome this noise and detect true differences in detection capacity.

The sequential design with fixed order is often being used to reduce the noise that is generated by an imbalance of polyp numbers at baseline. In this design technique A is applied first in all patients and detected polyps are removed. Subsequently, a second examination is performed with technique B in all patients (generally the novel technique) to assess whether it can detect additional polyps. However, if indeed the underlying question is whether the novel technique should be added to the existing technique, a more informative design would have two arms using the conventional technique at first and randomizing between technique A and B for the second examination.

A more efficient study design to compare two techniques is the randomized cross-over trial. This design greatly increases the power to detect a difference for two reasons. First, the technique with the better detection ability in the first round is likely to also detect a larger proportion of initially missed polyps in the second round. Therefore, information from the second inspection can strengthen the results from the first inspection. Second, patients in each arm have been examined by both techniques, making the total number of polyps detected (the denominator) a good reflection of the number polyps present at baseline. This means that the impact of differences in the number of polyps at baseline between the two arms (i.e. random noise) is reduced. Our simulations showed that the increase in statistical power with this study design is large, reducing the number of required patients. Furthermore, this design enables the evaluation of the characteristics of missed polyps, revealing which types of polyps are more difficult to detect for one technique compared to the other.

The strength of a true cross-over study design without removal is that each patient can serve as his own control (paired analysis). To perform a paired analysis, patients need to be in similar health status at the start of the first and second part of the study, which does not occur if polyps are removed during the first examination. Therefore, a true cross-over design requires that polyps are only removed during or even after the second examination. Our simulations showed that such a paired comparison had the greatest statistical power of all three study designs. However, several other adjustments are required to obtain a valid paired comparison. First, two endoscopists are required for each technique who should be blinded for each other’s findings. Second, polyps detected during the first inspection must be left in situ, providing an opportunity for detection during the second inspection. Third, each individual polyp should be matched between the two successive examinations. Last, patients whose polyps are detected by the first examination which are missed during the
second need a third examination to remove these polyps. Because of its impracticality, only a few detection studies have used this design.\textsuperscript{15, 29} Rather, the cross-over design with removal of lesions has substantially less practical limitations while its statistical power is still much greater than the parallel randomize design.

In addition to the selection of a study design, other factors should be taken into account which can interfere with observed results. First, the quality of bowel preparation, which can be measured quantitatively using classification schemes, is known to influence polyp detection rate.\textsuperscript{38-41} Second, the level of the experience of participating endoscopists, which is associated with polyp detection rates, may vary greatly between endoscopists.\textsuperscript{42} A third interfering factor is examination time, which has shown to be associated with polyp detection rates as well. Fourth, some patient characteristics such as age and indication of colonoscopy are associated with the prevalence of polyps.\textsuperscript{35, 43} Last, surveillance intervals of previously performed colonoscopies should also be equally divided amongst patients in the two study arms. In case differences in detection quality are found between two techniques, these ought not to be attributed to the abovementioned confounding factors, which can be balanced by stratification or use of the cross-over design.

In conclusion, we showed that the statistical power of the randomized parallel to detect differences in detection capabilities between techniques is low. Researchers should carefully consider whether the cross-over design can be used instead. The cross-over design has much greater power, although it may be more cumbersome for patients, endoscopists, and researchers. Finally, reporting possible confounders is obligatory in detection studies and detected differences in outcome measures should not be attributed to these confounders.
REFERENCE LIST


