The role of endoscopic imaging for an improved diagnosis of colorectal neoplasia
van den Broek, F.J.C.

Link to publication

Citation for published version (APA):
van den Broek, F. J. C. (2010). The role of endoscopic imaging for an improved diagnosis of colorectal neoplasia
CHAPTER 3
Valid and efficient study designs for the evaluation of new colonoscopic techniques: clinical and statistical considerations

Frank J.C. van den Broek
Evelien Dekker
Aeilko H. Zwinderman
Paul Fockens
Johannes B. Reitsma

Submitted
Introduction

Optical colonoscopy is considered the “gold standard” for the detection of premalignant lesions in the colon. However, polyps are frequently overlooked by standard colonoscopy and polyp miss-rates are estimated to be as high as 26% for lesions ≤5mm. In recent years, novel endoscopic imaging techniques have overwhelmed colonoscopic research in order to improve the detection of neoplastic lesions. In addition, these new imaging techniques may also be able to differentiate neoplastic from non-neoplastic polyps, thereby enabling the endoscopist to leave non-neoplastic lesions in situ and making colonoscopy more efficient.

Several imaging techniques (e.g. chromoendoscopy, narrow band imaging) have extensively been evaluated for their ability to detect colonic polyps on one hand, and for their diagnostic accuracy in differentiating neoplastic from non-neoplastic polyps on the other hand. Critical appraisal of these studies reveals remarkable differences in study design and statistical analysis, despite their similar objectives. The aim of this paper is to critically evaluate reported study designs, thereby focusing on validity and efficiency.

Regarding their ability to differentiate neoplastic from non-neoplastic polyps, new colonoscopic techniques can be compared using the classical design to evaluate diagnostic tests (i.e. comparing the index test result against a high quality reference standard, which is the histological examination of the polyp). The design, reporting and evaluation of such differentiation studies can methodologically be guided by the Standards for Reporting of Diagnostic Accuracy (STARD) initiative and the Quality Assessment tool for Diagnostic Accuracy Studies (QUADAS).

Methodological papers about studies evaluating new colonoscopic techniques with respect to their capability to detect polyps are lacking and therefore the focus of our paper is on detection. The detection of polyps is a methodologically challenging subject due to the lack of a high quality reference standard against which the new technique can be compared, and due to the possibility of multiple polyps within one patient. A colonoscopic technique should ideally detect all colonic polyps that are present within one patient, and hence the test result cannot be defined as a single dichotomous variable as in the classical evaluation of a diagnostic test.

In the next section we describe and critically compare the three most commonly reported study designs and discuss their respective methodology, outcome measures, and statistical analyses. In the final section, we present a flowchart that provides methodological guidance to researchers planning a colonoscopic detection study and to readers for critical appraisal of such studies.

Study designs to evaluate colonoscopic techniques for detection of polyps

A Parallel randomized design

Acknowledging that a traditional accuracy study for the purpose of detection is not possible because an acceptable reference standard does not exist, an alternative is to use a study with a parallel-randomized design (figure 1). In this design, patients are randomized to receive either
examination with technique A or technique B. At the end of the study we directly compare the number of detected polyps (*per-lesion analysis*) or the proportion of patients with ≥1 polyp (*per-patient analysis*) between both arms.8-11, 22-25

**Figure 1:** Parallel randomized study design comparing two colonoscopic techniques (A vs. B).

**Validity and efficiency**

Due to the randomization, the expected number of polyps or expected proportion of patients with ≥1 polyp will, on average, be the same in both arms of the study. Therefore, the 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 study, however, differences in the number of polyps between the two arms can be expected just by chance. Such differences have a strong impact on the final results because the presence of more polyps directly means that more polyps can be detected. So any mismatch in the number of polyps at baseline has a direct and large effect on the results.

The analogy with an intervention trial would be an imbalance in a major risk factor between treatment groups. Researchers will try to avoid such an imbalance by either stratified randomization or minimization. These techniques may be used in diagnostic colonoscopy studies as well, as the number of polyps in patients at baseline may partially be predicted by the age of the patient, gender, and colonoscopy indication.26, 27 However, the presence of polyps in patients is not fully understood, meaning that imbalances at baseline in the number of polyps cannot be avoided. This implies that results from one trial to another will vary just because of this random noise, and sample sizes need to be sufficient large to discover real differences in detection capability between techniques (see also results of our simulation study in the appendix). In addition, polyp detection rates may vary greatly between different endoscopists and between different examination times.28-31

**Outcome measures and analysis**

The results of the parallel, randomized design can be analyzed on a *per-polyp* basis or on a *per-patient* basis. The former has the advantage of more statistical power, since each polyp will generate information about the detection capability of a technique. Besides, the detection of each
individual polyp has clinical significance, as each polyp may be premalignant. In situation with multiple lesions there is the issue of correlation in detection within a patient. In general, correlation is ignored as the impact of patient factors on detection is considered limited.

This means that the number of detected polyps per patient should be used as primary outcome measure for comparison of the two techniques (with non-parametric Wilcoxon rank or Mann-Whitney U testing to detect a shift in distribution towards more polyps). Additionally, the proportion of patients with ≥1 polyp may be used as secondary outcome measure (Chi-square or Fisher’s exact testing).

Example from literature
A parallel randomized study by Inoue et al can illustrate the reporting and analysis of this type of study design.9 Patients planned for surveillance, screening or diagnostic colonoscopy were randomized to undergo either conventional colonoscopy (n=121) or narrow band imaging (n=122). The procedures were performed by one of six endoscopists who were allowed to select one out of two colonoscopes (PCF-240ZI or CF-H260AI, Olympus Inc., Tokyo, Japan) having different resolutions. The main outcome measure was the number of adenomas (e.g. mean number of adenomas per included patient), which resulted in 102 adenomas (0.84 per patient) for narrow band imaging vs. 65 (0.55 per patient) for conventional colonoscopy (Student’s t-test; p=0.046).

Although patients were well balanced for age, gender, indication, and level of bowel preparation, and examination times were comparable between both randomization groups, it was not reported that one of the six endoscopists performed significantly (p<0.001; calculated from their results with Chi-square testing) more procedures with narrow band imaging in this study. Consequently, the difference in performance between the two techniques may be attributed to the experience of this single endoscopist. Stratification could have prevented this imbalance in endoscopists.

The authors used the Student’s t-test was used for comparison but should be reserved for comparing normally distributed data. As the number of adenomas per patient varied greatly and most patients did not harbor adenomas at all, the adenoma distribution was skewed. Consequently, a non-parametric test would have been more appropriate and probably more powerful.

The main uncertainty remains whether the true number of polyps was equally distributed at baseline in this study. Substantial differences in the number of polyps can occur just by chance. These differences can be large in comparison to the expected differences in detection rate between the two techniques being evaluated. A relatively large difference at baseline will dominate the results of the study, basically making the study useless (uninformative). Our simulations show that to detect an absolute difference of 10% in detection rate (80 vs. 90%), the required sample size is around 2000 patients per group to achieve a power of 80% (see table 1, first row of results; or see appendix for further details). This is much higher than the number of patients required to detect a difference in a study where the proportion of patients with an event is 80 vs. 90%; a sample size of 200 per group would then be sufficient. Several adjustments in study design have been proposed to reduce the noise generated by an imbalance in the number of polyps at baseline.
B Sequential (back-to-back) design with fixed order

The most encountered attempt to increase the efficiency of a study comparing two colonoscopic techniques is using a sequential design with fixed order (i.e. back-to-back procedure). Technique A is applied first in all patients and any detected polyp is removed immediately. In a subsequent second examination, all patients are evaluated again but now with technique B (generally the more novel technique) to assess whether it can detect additional polyps, e.g. polyps missed by the first technique (see figure 2).32-38

![Sequential study design with fixed order](image)

**Figure 2**: Sequential study design with fixed order: A colonoscopic examination with technique A is followed by a second examination with technique B. Polyps detected by technique A are removed immediately, as a result of which only missed polyps can be picked up by technique B.

**Validity and efficiency**

The potential for increased efficiency by this design lies within the fact that each patient is examined by both techniques.39 However, this sequential design with fixed order is only informative if the underlying research question is whether the new technique should be added to the existing technique, i.e. the intended role of the new technique is that of an add-on test.40 New techniques are valuable as add-on tests when they are capable of identifying additional polyps.

In case the underlying question is whether either technique A or B should be used for the detection of polyps (replacement question), this study design does not provide informative evidence. The reason is that it remains unknown what would be the performance of technique B with respect to polyps that have already been detected and removed by technique A. The assumption that all these polyps would have been identified by technique B as well is untenable, given the fact that polyp miss-rates up to 48% have been reported for standard colonoscopic techniques.28 The bottom line is that for the decision whether to use either technique A or B for polyp detection, this study design is not providing the right evidence and should not be used.
In case the study objective is indeed that of an add-on diagnostic test (should we use technique B on top of technique A?), a study design with two arms using technique A first during both study arms and randomizing between technique A and B for the second examination would be a more informative design (A+A vs. A+B).

**Outcome measures and analysis**

When using this design, the correct outcome measure is the polyp miss-rate of technique A, given the fact that technique B is used during the second examination. This miss-rate is defined as the number of polyps detected during technique B, divided by the total number of polyps detected by A plus B (per-lesion analysis). Likewise, for a per-patient analysis the miss-rate is defined as the number of patients with ≥1 polyp during B, divided by the total number of patients with ≥1 polyp during either A and/or B. These proportions should only be described without comparative statistical tests.

A better alternative to answer the question whether technique B is really superior as an add-on test, is a design where the second round of examination is either by A or B, i.e. comparing A+A vs. A+B. The appropriate analysis would be the comparison of the two miss rates (proportions) using a Chi-square or Fisher's exact test.

**Example from literature**

Several studies may illustrate the problems in analyzing and reporting the results of this study design. East et al used this design in patients with hereditary non-polyposis colorectal cancer syndrome. High definition endoscopy was used first (technique A) followed by narrow band imaging (technique B); polyps were removed instantaneously during technique A. The proportion of patients with ≥1 adenoma during technique A (17/62 pts) was compared to the same proportion during both A and B (26/62) using a statistical test for paired data (p=0.004; McNemar’s test).

However, the data from such a design are not fully paired as the initial polyps detected by technique A have not been examined by technique B. It is therefore impossible for techniques A plus B to detect fewer lesions than technique A alone, meaning that the paired analysis is incorrect as it will always favor technique B. Most studies with this design have used the incorrect paired analysis. It is better to report the magnitude of the polyp miss-rate and possibly compare it with previously reported miss rates, as accurately done by Rastogi et al. A systematic review of back-to-back colonoscopy studies has demonstrated an estimated overall polyp miss-rate of 21% (95%-confidence interval: 14-30%). When comparing one’s own study results to this systematic review, one has to be aware of the fact that miss-rates may vary greatly by level of experience, examination time, bowel preparation and by utilized technique.

**C Cross-over (back-to-back) design with randomized order and direct removal**

A more informative and efficient study design to compare the accuracy of two techniques to detect polyps is one in which patients undergo back-to-back examinations with both techniques, but the order in which they receive both techniques is determined by randomization (see figure 3). Polyps detected by the first technique (either A or B) are removed immediately, as a result of which only missed polyps can be picked up during the second examination.
Validity and efficiency

The layout of this design is similar to that of a crossover trial in intervention research comparing two treatments. The first part of this design (before cross-over) is equal to the parallel-randomized design. This also means that the same outcomes and analyses can be used for this part. The strength of this design is that it greatly increases the power to detect a difference in detection for two reasons. Firstly, the information from the second round can strengthen the results from the first round because it is likely that the technique with the better detection in the first round will also detect a larger proportion of initially missed polyps in the second round. This increases the difference in detected polyps between techniques. Secondly, because patients in each arm have been examined by both techniques, the total number of polyps detected (the denominator) is a reflection of the number polyps present at baseline. This means that the impact of differences in polyps at baseline is reduced, making this design more efficient. From our simulations it is clear that the increase in power is large, often reducing the number of patients required by a factor of 3 or more. The contribution of the second mechanism is more important, as even in the situation where there is no difference in detection between techniques in the second round, there is still a large increase in statistical power (see table 1).

Furthermore, this design enables the evaluation of the characteristics of the missed polyps (additional outcome measure), revealing which types of polyps are more difficult to detect for one technique compared to the other.

**Figure 3:** Crossover study design with randomized order and direct removal of polyps. The first of the back-to-back examinations is determined by randomization.
Outcome measures and analysis
Analogous to the ‘sequential (back-to-back) design with fixed order’, the primary outcome measure can be defined on a per-polyp or on a per-patient basis. The proportion of initially detected polyps (or its complement: the proportion of initially missed polyps can be compared between the randomization groups using the Chi-square or Fisher’s exact test. Although these ratios are not true proportions from a statistical perspective, because the denominator is a number with uncertainty rather than a known number, the standard test for a difference in proportion (Chi-square test) performs well (see simulations).

Example from literature
We have previously used this randomized cross-over design to compare the detection of neoplasia between autofluorescence imaging and high resolution endoscopy among patients with ulcerative colitis.41 Twenty-five patients were randomized to autofluorescence imaging first and high-resolution endoscopy second, whereas 25 other patients were allocated to high-resolution endoscopy first and autofluorescence imaging second. Lesions detected during the first techniques were immediately sampled or removed. The neoplasia miss-rates (proportion of neoplastic lesions missed during the first examinations) were compared between the two techniques by using Fisher’s exact test for proportions.

D Cross-over design with randomized order and matching of polyps

Figure 4: Cross-over study design with randomized order without direct removal of polyps. The first of the back-to-back examinations is determined by randomization. Polyps detected by the first technique (either A or B) are left in situ, as a result of which these polyps can be picked up during the second examination as well.
Potential additional increment of efficiency

This cross-over design with direct removal of all detected polyps after each stage is still not optimal from a methodological point of view. The strength of a true cross-over trial is that each patient can serve as his own control (paired analysis). A key factor is therefore that patients are in similar health status at the start of the first and second part of the study. Thus, cross-over intervention trials are limited to chronic stable conditions where symptomatic treatments can be compared, e.g. patients are not cured. Furthermore, a period without any treatment is build after the first treatment, the so-called wash-out period, in order to bring patients back to their original health status. If this is feasible, the treatment effect in each part of the study can directly be compared within patients.

However, this situation does not arise in a cross-over colonoscopy study when polyps are removed during the first examination; patients then have a different number of polyps during the second examination. A further gain in efficiency would therefore be reached if removal of polyps could be postponed to after the second examination. This would allow a head-to-head comparison between the techniques for each polyp within a patient (figure 4). Our simulations show that such a true paired comparison has the greatest statistical power to detect a difference in detection (see simulation results).

A true paired comparison of detection requires several fundamental adjustments. First, two endoscopists are required: one performing technique A and another technique B. Second, they should be blinded for each other’s findings. Third, polyps detected during technique A must be left in situ, providing an opportunity for detection (and removal) by technique B as well. Fourth, in order to perform a paired analysis, each individual polyp should be matched between the two successive examinations.

Several methods exist to match identical polyps between the procedures: (1) linking polyp size, distance from the anus, colonic location, and shape; (2) comparing photographs of each polyp; and (3) using an independent observer of both examinations to determine whether polyps can be matched. The most reliable method, however, would be to perform all three methods to match polyps between techniques A and B. This impractical methodology may cause resistance among patients, endoscopists and researchers.

Example from literature

This paired variant of the randomized cross-over design has been used only once. This study compared standard colonoscopy to narrow band imaging for the detection of patients with neoplasia in ulcerative colitis. Two endoscopists blinded for each other’s results performed either technique A or B in the same patient. The McNemar’s test was used to compare the proportion of patients in which a neoplasia was detected. The exact information on individual lesions was not observed in a paired manner, preventing the possibility for a per-lesion analysis as well.

As uncertainty may still exist when matching individual lesions, one might prefer to compare only the total number of detected lesions within each patient using a paired Wilcoxon rank test.
Results of simulation study

The underlying research question in our simulations is to compare two colonoscopic techniques (A and B) and determine which one has the better detection capability. Across several clinical scenarios, we compared the three main study designs that have been used for such a study: the parallel randomized design (figure 1), the randomized cross-over design with (figure 3) and without direct removal of polyps (figure 4). Our focus was on difference in statistical power between the three designs. Detailed results of our simulations can be found in the appendix.

The final results of our simulations are summarized in table 1. Each scenario was repeated 1000 times across a range of sample sizes and we report the sample size that produced a significant result in about 80% (power) of the simulations. Scenarios differed in the number of polyps present, the absolute height and difference in detection capabilities. The simulation results indicate that the cross-over design has far greater power to identify differences in detection than the parallel design. In many scenarios, the randomized design requires 10 to 15 times more patients than the cross-over design with direct removal to reach the same statistical power of 80%. Even further reductions in sample size can be achieved if matching of individual polyps (i.e. a true paired analysis) would be possible.

In the three scenarios where there was no differences in detection capability, the frequency of significant results (type I error = finding a significant result when in reality there is none) were close to the expected nominal value of 5% (see Table). This was true for all three designs.

Recommendations for research

We have outlined the various advantages and disadvantages of commonly encountered designs in studies evaluating the detection capability of new colonoscopic techniques. Researchers have to balance these against each other to select the most adequate study design. Factors which have to be taken into account are validity, statistical power (efficiency), possibility to perform back-to-back examinations, and feasibility to match lesions in a cross-over design.

With respect to validity, the sequential design with fixed order should be abandoned in case the research question is that of a replacement question (should we use either technique A or B for the detection of polyps?). Only in case the objective is that of an add-on diagnostic test (should we use technique B on top of technique A?), this design provides useful information. In that case however, a study design where technique A is applied first in all patients and then patients are randomized for additional examination either by A again or technique B is more informative. Only then the additional value of technique B on top of A can be compared to the strategy in which technique A is used twice. Nevertheless, as in colonoscopic research the objective is that of a replacement question, a parallel randomized design or a randomized cross-over design should be selected.
Table 1: Summary table showing total sample sizes required to achieve 80% power for the three main designs applied in studies evaluating the detection capability of colonoscopic techniques. Scenarios differed in the number of polyps present, the underlying size and difference in detection capabilities.

<table>
<thead>
<tr>
<th>Mean number polyps</th>
<th>P1_A</th>
<th>P2_B</th>
<th>P1_B</th>
<th>P2_A</th>
<th>Parallel randomized</th>
<th>Cross-over with direct removal</th>
<th>Cross-over with matching</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.85</td>
<td>0.80</td>
<td>0.95</td>
<td>0.70</td>
<td>3980</td>
<td>280</td>
<td>125</td>
</tr>
<tr>
<td>2</td>
<td>0.85</td>
<td>0.80</td>
<td>0.95</td>
<td>0.70</td>
<td>2400</td>
<td>145</td>
<td>70</td>
</tr>
<tr>
<td>2</td>
<td>0.85</td>
<td>0.7</td>
<td>0.95</td>
<td>0.70</td>
<td>2400*</td>
<td>180</td>
<td>70*</td>
</tr>
<tr>
<td>1</td>
<td>0.75</td>
<td>0.60</td>
<td>0.90</td>
<td>0.50</td>
<td>930</td>
<td>130</td>
<td>50</td>
</tr>
<tr>
<td>Scenarios with no difference in detection</td>
<td>% significant studies</td>
<td>% significant studies</td>
<td>% significant studies</td>
<td>% significant studies</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SS=200</td>
<td>SS=800</td>
<td>SS=200</td>
<td>SS=800</td>
<td>SS=200</td>
<td>SS=800</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>5.2%</td>
<td>5.5%</td>
<td>5.6%</td>
<td>5.2%</td>
<td>3.9%</td>
<td>4.6%</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>5.9%</td>
<td>4.1%</td>
<td>6.1%</td>
<td>5.5%</td>
<td>6.5%</td>
<td>5.7%</td>
<td></td>
</tr>
</tbody>
</table>

P1_A sensitivity of technique A as first examination;  
P1_B sensitivity of technique B as first examination  
P2_A conditional sensitivity of technique A as second examination after technique B;  
P2_B conditional sensitivity of technique B as second examination after technique A  
* same sample size as row before as conditional sensitivities are not used in these designs
With respect to efficiency (or statistical power), there is the general issue that an analysis at the level of lesions is more powerful than an analysis at the level of patients. Besides, each polyp has clinical significance as each polyp may be premalignant and should therefore be detected. The power of the parallel-randomized design to detect differences in detection capability is very limited. The main reason is that differences in the number of polyps at baseline, although by chance, have a large impact on the results. Even with relatively large sample sizes, differences in the number of polyps are possible that will dominate the results. It means that only very large parallel-randomized studies have appropriate power. The crossover design with randomized order is much more powerful and should be considered more often. The main reason for the increase in power is that using the total number of detected polyps during both examinations as the denominator reduces the impact of any imbalance in the number of polyps at baseline. Back-to-back procedures may however be repellant and impractical for patients and endoscopists, but researchers need to be more aware of the low power of the parallel design.

The randomized crossover design with matching of polyps would be even more efficient, as a true paired analysis increases the statistical power even further. However, such a design would require two endoscopists and a time-consuming, technical challenging approach to match lesions. Despite extensive precautionary measures, there may still be a risk of incorrect matching of lesions. Lastly, some patients who will have polyps detected by the first examination that are subsequently missed by the second examination need a third colonoscopy to remove these potential harmful polyps.

In figure 5 a flow chart is presented that guides researchers in selecting the most appropriate study design for the evaluation of a new colonoscopic technique with respect to its ability to detect polyps.

Except for proper selection of study design, researchers have to be aware of several other possible factors, which may interfere with the outcome measures. One of those factor is the quality of bowel preparation, which can be measured quantitatively by using different classification schemes. Polyp detection rates may be higher in case the colon is perfectly prepared; however, this factor will be balanced by randomization. Another important factor is the experience of the endoscopist. Polyp detection rates or miss-rates may vary greatly between endoscopists. Therefore, the experience of the endoscopist should preferably be stated in terms of polyp detection rates or miss-rates; in case of absence of these quality indicators, the number of endoscopies performed in the past should be stated. A third possible factor is examination time, which has shown to be associated with polyp detection rates as well. Longer inspection times are associated with higher polyp detection rates. Lastly, some patient characteristics are associated with a higher prevalence of polyps, such as gender, age, race or indication of colonoscopy (e.g. polyposis patients). In case differences are found between two colonoscopic techniques, these ought not to be attributed to abovementioned confounding factors. Therefore, these factors have to be balanced between the two techniques by randomization or their impact should be examined at the analysis stage.
Summary

In summary, when evaluating new colonoscopic techniques with respect to their ability to differentiate neoplastic from non-neoplastic polyps a classical diagnostic accuracy study can be used by using histopathology of the removed polyps as reference standard. As no high quality reference standard exists for the evaluation of new colonoscopic techniques with respect to their ability to detect polyps, we critically evaluated the most commonly used study designs. The randomized parallel design has been frequently used and, although is free from bias, its power turns out to be disappointingly low. Researchers should carefully consider whether the cross-over design can be used instead. This design has far greater power, but will be more cumbersome for patients, endoscopists, and researchers. Outcome measures (e.g. polyp detection rates) should be properly defined, as well as the use of appropriate statistical tests.

Finally, reporting of possible confounders for the outcome measures (i.e. gender, age, race, indication of colonoscopy, experience of endoscopists, degree of bowel preparation, examination time, and type of endoscope used) are obligatory in detection studies. Detected differences in outcome measures may not be attributed to these confounders and should be balanced either through randomization or by having a strict protocol.

Figure 5: Guidance for researchers for selection of study design. Designs are listed in order of efficiency with the most efficient designs at the top.
The underlying research question in our simulations is to compare two colonoscopic techniques (A and B) and determine which one has the better detection capability. In a series of simulations we compared the three main study designs that have been used for such a study: the parallel randomized design (figure 1), the randomized cross-over design with (figure 3) and without direct removal (figures 4). Our focus was on difference in statistical power between the three designs across a range of clinical situations.

In our simulations we varied the following parameters: total sample size; the true number of polyps within each patient coming from a negative binomial distribution (characterized by the mean number of polyps and a overdispersion parameter); detection capability of technique A when applied first (sensitivity of A = p1_A); detection capability of technique B when applied first (sensitivity of B = p1_B); detection capability of technique A in the second round to find additional polyps missed by technique B (conditional sensitivity of A = p2_A); the detection capability of technique B at the second examination to find additional polyps missed by technique A (conditional sensitivity of B = p2_B).

To explain our simulations in more detail, we will illustrate the first scenario in table A by showing the intermediate results from a single study based on this scenario. We will use a study with total sample size of 200 meaning that in the parallel design 100 patients will be randomized to examination by technique A and 100 patients by the new technique B. The detection capability of technique A (representing conventional colonoscopy) is assumed to be 85%, which is 1 minus the reported miss-rate of at least 15%. The new technique has a better detection capability at 95%. In our simulations, we replicated such a study with these parameters 1000 times and we calculated how many times the three designs would produce a statistically significant result, i.e. the power of such a study. To actually generate detected polyps in patients we need to simulate the true, but unobserved number of polyps within each patient. To do this, we used the negative binomial distribution. This type of distribution is closely related to the Poisson distribution and is characterized by the mean number of polyps and an overdispersion parameter to account for the fact that the number of patients with no polyps and the number of patients with many polyps is often higher than expected under the Poisson distribution. In the basic scenario 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.
Chapter 3

Table A: Observed distribution of the number of polyps in a single simulation study with the parallel randomized design

<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 or more</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Total polyps detected</td>
<td>86</td>
<td>94</td>
</tr>
<tr>
<td>Mean number per patient</td>
<td>0.86</td>
<td>0.94</td>
</tr>
<tr>
<td>P-value</td>
<td>0.59</td>
<td></td>
</tr>
</tbody>
</table>

Total sample size = 200; mean number of true polyps = 1
Sensitivity of technique A = 85% and for B = 95%

The true number of all polyps (detected + undetected) among patients will never be observed (unless at autopsy), but after applying a specific detection rate we obtain an observed number of polyps in our simulations. In table A the distribution of detected polyps within patients is shown for the two randomized groups after they have been examined with two techniques that differ in their detection capability (technique A 85% vs. technique B 95%). In the analysis we compare the observed number of polyps in both arms of the trial. The non-parametric Wilcoxon statistic can be used to test whether the distribution has shifted towards more detected polyps in the group examined with the new technique. Our single study produces a non-significant p-value of 0.59. This is a typical value for the parallel-randomized design in this scenario because with a total sample size of 200 the parallel design has a power of only 10%. To achieve a power of 80%, the total sample size needs to be as high as 4,000 (see first row of table 1 in manuscript).

In the crossover design with removal of polyps, patients will be examined by the other technique to see whether any additional polyps can be detected. In our simulation study the number of missed polyps is known and we can simulate whether these will be detected in the second round using the conditional detection capabilities of technique A and B. The conditional (i.e. as observed in the second round) sensitivity of technique A (70% in first scenario) and B (80%) was assumed to be lower than their corresponding sensitivities in the first round as missed polyps are probably more difficult to detect. The results of the same single study are presented in table B. In the paired design with removal we can calculate for each strategy the ratio of the number of polyps detected in the first round by the total number of detected polyps in both rounds. A higher proportion indicates that the initial technique has better detection capabilities. A chi-square test can be used to test for a difference in the two proportions. This crossover design with the same sample size, but examining patients twice, now produces a significant result with a p-value of 0.015. To obtain a power of 80% for this design requires a sample size of only 280.
Table B: Number of detected polyps per strategy in the cross-over design with direct removal

In the crossover design without initial removal, both techniques are applied in the same patient and for each detected polyp we observe whether it was detected by only one technique (thus missed by the other) or detected by both techniques. The results of the 200 patients can be summarized in the following 2×2 table: The difference in detection capability manifests itself in the number of discordant results in this case 35 versus 14. These numbers can be analyzed by the McNemar test for paired observations, leading to a p-value of 0.0027. Because this is now a true paired analysis on each detected polyp, it increases the power even further. To achieve 80% power with this design using the same scenario now only requires 125 patients (see table 1 in the manuscript).

Table C: Number of polyps detected by only one or both techniques in the cross-over design with matching of lesions

The final results of all our simulations are summarized in table 1 of the manuscript. Each scenario was repeated 1000 times across a range of sample sizes and we report the sample size that produced a significant result in about 80% of the simulations. In many scenarios, the randomized parallel design requires 10 to 15 times more patients to reach the same statistical power of 80% than the cross-over design with direct removal. Even further reductions in sample size can be achieved if matching of individual polyps (i.e. a true paired analysis) is possible.
In the final three scenarios there were no differences in detection capability between the two techniques. In these simulations we report the number of significant results when using a total sample size of 200 and 800. In these simulations the frequency of significant results (type I error = finding a significant result when in reality there is none) were close to the expected nominal value of 5% (see final table). This was true for all three designs.
Reference List


11. Rex DK, Helbig CC. High yields of small and flat adenomas with high-definition colonoscopes using either white light or narrow band imaging. Gastroenterology 2007;133:42-47.


30 Chen SC, Rex DK. Endoscopist can be more powerful than age and male gender in predicting adenoma detection at colonoscopy. Am J Gastroenterol 2007;102:856-861.


