Do’s and don’ts in evaluation of endoscopic screening for gastrointestinal cancers

Authors
Michael Bretthauer1,2,3, Mette Kalager1,4, Hans-Olov Adami1,3,5

Institutions
Institutions are listed at the end of the article.

Evidence in perspective

Introduction
Improved management of gastrointestinal cancers is one of the biggest challenges in cancer control. Colorectal cancer (CRC) alone is the second most common cancer globally, with an estimated 1.3 million new cases and 700,000 deaths every year [1]. Taken together, cancers in the colorectum, stomach, and esophagus are the most common malignancies worldwide [1]. The ongoing epidemiologic transition in low- and middle-income countries—encompassing an increase in life expectancy and a change toward a Western lifestyle—will predictably further increase this burden, especially for CRC.

Although different in their etiology, temporal trends for incidence and mortality, and symptoms and clinical management, the three main gastrointestinal cancers have two important features in common. First, the prognostic outlook following clinical diagnosis is gloomy. At best, long-term cure is achieved in 10% of patients with esophageal and 20% of patients with stomach cancer [1]. The prognosis among patients with CRC is better, but overall cure rates only marginally exceed 50% [1]. Second, therapeutic progress has been slow for all three sites. Metastatic disease remains incurable for most patients.

When treatment fails, a main option to reduce suffering and death from these gastrointestinal cancers is screening. The fundamental goal in endoscopic cancer screening is to reduce cancer incidence and mortality. Needless to say, this can only rarely be achieved among patients who appear in our offices with symptoms; success of screening relies on targeting the whole population before this symptomatic stage is reached.

Is endoscopy the magic bullet?
Population screening by endoscopy needs serious consideration as a strategy to reduce the enormous global burden from cancers of the esophagus, stomach, and colorectum. Screening is, however, distinctly different from clinical service, as it targets people who are not knowingly sick and who have a very low risk of getting the disease they are screened for. Thus, benefits and harms need to be carefully evaluated and quantified.

Fundamental principles of screening
There is a fundamental difference between diagnosing symptomatic patients who show up in our offices and detecting asymptomatic disease in the general population. In the former situation, the prevalence of disease is often high (e.g., 10%, 50%, or even higher), and there is an ethical imperative to diagnose and treat whenever possible. The evidence from everyday practice, showing that cancer patients diagnosed early fare better than those detected with locally advanced let alone metastatic disease, is compelling.

Unfortunately, it would be cavalier to generalize this experience to make the case for population screening. The primary ethical imperative here is...
not to treat, but rather to be confident that the overall benefit offered to those who accept the invitation to be screened is greater than the harm. The prevalence of cancer precursors in gastrointestinal cancer screening programs can be high, but the progression risk is low, often less than 1% [2, 3]. The lifetime risk of CRC, even in high-incidence countries such as Norway or the Czech Republic, is approximately 5%. Thus, the vast majority of individuals in the screening population will never get the disease, and have no theoretical possibility of benefiting from screening. Therefore, there has to be clear evidence that the benefits of screening outweigh the harms.

Cancer screening can be applied in a so-called “opportunistic” fashion or in the form of an organized screening program. The latter entails stringent organization, which not only includes the screening test itself, but also systems for quality control, management, invitation and recall, and treatment and surveillance after screening; opportunistic screening does not necessarily include these features. The success of screening thus depends not only on the screening test itself, but on well-organized comprehensive programs.

**Lead-time and length-time bias in screening**

Important principles of cancer screening are illustrated in Fig. 1. Understanding these principles is fundamentally important for the design of studies that aim to quantify the benefit of screening.

Regardless of whether screening is opportunistic or is delivered via an organized program with invitation of eligible individuals, the goal is to advance the time of diagnosis, to detect cancer at an earlier stage, and thus to reduce mortality. Sojourn time is the maximum time that screening can advance diagnosis of a precursor as well as an invasive cancer. The sojourn time begins when the lesion first becomes detectable and ends when it would have been diagnosed in the absence of screening. For lesions that would never have surfaced clinically, the sojourn time has no upper limit. For other lesions, the sojourn time has a distribution that is determined by features of the malignancy, growth rate, and the performance of the screening test.

The advancement of diagnosis achieved by screening is called the lead time. This time period starts when a lesion is first detected by screening during the sojourn time and ends when it would have been diagnosed clinically in the absence of screening (Fig. 1). Lead time directly affects outcome measures in screening studies, for example by biasing the result of survival estimates when comparing screened and nonscreened individuals (see below).

Another fundamental issue in screening is that fast-growing lesions are more likely than slow-growing ones to become symptomatic and diagnosed clinically before or between screening events. Hence, from the entire pool of incident lesions, screening undersamples those that grow rapidly, a phenomenon called length-time bias.

**The two approaches of cancer screening**

Cancer screening can reduce mortality through two different mechanisms.

1. Preventive screening: screening detects and prompts the removal of benign cancer precursor lesions that might otherwise have progressed to cancer.
2. Early detection screening: screening detects early-stage invasive cancer before it has become incurable with existing therapies.

**Target of endoscopy screening**

The main effect on cancer mortality of preventive screening methods targeting precursor lesions (as endoscopic screening) is through reduction of cancer incidence. Thus, the effect of endoscopic population screening on mortality from cancers with established precursors (such as Barrett’s esophagus or colorectal polyps) is primarily a function of reduced incidence because individuals who do not get cancer do not die from it. An exception is screening in populations with high cancer prevalence, where the direct effect on cancer mortality by early detection of prevalent early cancers at screening plays a larger role. Screening by fecal occult blood testing (FOBT) mainly facilitates early detection of invasive cancer, and thus does not aim primarily to reduce cancer incidence. However, high positivity rates in FOBT screening programs with subsequent high rates of colonoscopies may reduce cancer incidence as a result of removal of adenomas at colonoscopy irrespective of the overall accuracy of the FOBT.

A prerequisite for the preventive screening approach is obviously that a premalignant lesion exists as an intermediary step in the malignant transformation of a normal cell to a malignant phenotype. A second prerequisite is that the premalignant lesion can be detected by means of a screening test that has high sensitivity and specificity, and is feasible to use in thousands of healthy individuals.
Overdiagnosis and overtreatment

Neither preventive nor early detection screening approaches to prevent cancer incidence and/or death are a free lunch; the price is overdiagnosis and overtreatment of lesions that would never have progressed to a symptomatic or even lethal cancer during the patient’s remaining lifetime. The probability of a cancer precursor progressing is unknown and is methodologically difficult to study because it would be unethical in most instances to observe the natural history of such lesions without therapeutic intervention. Studies of the natural history of colorectal polyps have shown that many small- and medium-sized polyps actually regress over time [2], and even large polyps may have a low progression rate to cancer [3]. Most adenomas will never grow into cancer, and the progression time is long—presumably 10 or even 15 years. Modeling further indicates that progression probabilities and progression times from carcinoma in situ to invasive cancer of the cervix are low, in the order of 12% and 14 years, respectively [4]. If similar scenarios also apply to gastrointestinal cancers—which currently remain unknown—a combined benefit of reducing both incidence and mortality would obviously be achieved at a substantial price—both human and economic—of overtreatment.

The level of overtreatment resulting from removing polyps at endoscopy screening is high, because most removed polyps would not have progressed to cancer. The individual “price to pay” for overtreating polyps by polypectomy is small because complications are rare. However, for society, the price may be considerable because polyp removal by colonoscopy is resource-demanding. In addition, endoscopic screening is uncomfortable or painful for patients. For early detection screening, the amount of overdiagnosis (detecting cancer that would not have progressed without screening) may be smaller in number, but side effects and complications of cancer removal are more common and more severe compared with polyp removal. Overdiagnosis and overtreatment exist in both screening approaches and should be closely monitored in any screening program.

Surveillance after screening: from individuals to patients

Detection of precursor lesions might allow identification of individuals whose risk of developing cancer is so high that surveillance programs are justified and add benefit. However, this is a complex and underdeveloped clinical domain because individual risk prediction is often uncertain, surveillance programs are rarely evidence based, costs can easily escalate, and resources for more pressing needs in health care may become re-allocated to surveillance [5,6]. We also need a better understanding of the consequences (e.g. insurance implications) and possible negative impact on quality of life when screening programs define people to be at high risk and recommend surveillance.

Quantifying the benefits of screening

Cancer mortality, incidence, and survival are outcome measures that are frequently used to quantify benefits of screening, but some of them are uninformative or even misleading (Table 1).

Table 1 Outcome measures in endoscopic screening for gastrointestinal cancer and their validity for screening evaluation (+, advantages; –, disadvantages).

<table>
<thead>
<tr>
<th>Outcome measure</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cancer mortality</td>
<td>++</td>
<td>–</td>
</tr>
<tr>
<td>Cancer incidence</td>
<td>+</td>
<td>–</td>
</tr>
<tr>
<td>Survival</td>
<td>––</td>
<td>–</td>
</tr>
</tbody>
</table>

Cancer mortality

The widely accepted gold standard for the assessment of the benefit of screening is randomized trials in which the mortality from a particular cancer is compared between individuals invited to screening and those not invited, with both groups followed up for a sufficiently long period of time. This design eliminates the impact of lead-time bias, length-biased sampling, overdiagnosis, and confounding due to different baseline risks of disease among screened and unscreened individuals. The challenge is that randomized screening trials are difficult to undertake; they require long-term follow-up of large cohorts, which is often unaffordable or is unfeasible because screening contamination in the control group may entail underestimation of the benefit and compliance may be too low among those invited to screening. Furthermore, if previous trials have documented a benefit from screening, a new trial with an unscreened control group may be considered unethical. Increasingly, comparative effectiveness research is needed to accommodate this situation by comparing different screening modalities, and will often use noninferiority as an outcome criterion [7]. Novel approaches for unbiased assessment of screening without randomization are needed, but such assessments are methodologically challenging. For example, as has been shown for breast cancer and is presumably also applicable for gastrointestinal cancers, comparison of cancer mortality among those who attend and those who do not attend screening is largely uninterpretable because it remains unknown whether the baseline risk of dying from a particular cancer is identical in these two groups and whether early cancer symptoms influence motivation to attend screening [8]. Similar selection factors can also bias comparisons between those who attend and those who are not invited to screening.
Comparisons of mortality rates before and after introduction of screening in a population – although still often used – are also notoriously difficult to interpret. Bias can arise from numerous sources such as concomitant trends in cancer incidence and mortality, increased awareness and/or access to health care, and improved treatment over time [9]. However, if biases can be adequately controlled in the design or analyses, valid estimates of benefit and harm may be achievable, although they require sophisticated societal infrastructures [9].

Cancer incidence
Cancer incidence is an informative measure when screening entails detection and removal of precursor lesions. This is indeed an opportunity to estimate benefit without awaiting mortality data to become available because if some of the removed precursors would have progressed to invasive cancer, then a reduction in incidence must occur sooner or later. If no such reduction is recorded compared with an unscreened population, the concept of “cancer precursor” should be reconsidered.

When a screening test solely or predominantly detects invasive cancer early, monitoring incidence rates serves different purposes and is less informative as a measure of benefit. Because the test must achieve a lead time by advancing the time of diagnosis, an increase in incidence is expected after initiation of screening. During extended follow-up, the early increase is typically followed by a deficit caused by the cancers diagnosed earlier due to screening. Gradually, the incidence in the screened population will then approach that in an unscreened population. For once-only screening (such as in some colonoscopy screening programs), the period of time from initiation of screening until the incidence is similar in screened and unscreened individuals will correspond to the sojourn time (described above). For CRC screening by flexible sigmoidoscopy, 10 years or more of follow-up are required to observe an unbiased effect on incidence after screening [10]. Studies with shorter follow-up may show no screening effect and may falsely reject a screening benefit [11]. Interpretation of incidence rates becomes even more difficult when the screening test detects both cancer precursors and early invasive cancer, as is the case in most endoscopic screening programs. What is observed is the net-effect of two phenomena, namely reduced incidence due to cancer prevention and increased incidence because diagnosis is advanced in time (lead time). Further complications arise if screening entails overdiagnosis of invasive but indolent cancers. The different time frames for these effects on incidence may, however, facilitate interpretation; the excess incidence due to earlier diagnosis (and possibly overdiagnosis) will arise early during follow-up, whereas reduced incidence due to removal of precursors will appear later in time.

Because of these complicated and often unmeasurable (and thus uncontrollable) mechanisms, studies investigating incidence of cancer or cancer precursors in patients on different surveillance strategies (e.g. yearly vs. three-yearly) with precancerous lesions such as Barrett’s esophagus, atrophic gastritis, or colorectal adenomas are often biased due to lead-time effects, and are thus unreliable. The only exception is when patients under different surveillance strategies have undergone their last screening or surveillance examination at the same point in time before the end of follow-up; however, most often this is not the case.

Survival
Survival rates among patients with screen-detected invasive cancer – although still often used – are uninformative and often misleading. As illustrated in Fig. 1, a longer survival time for screen-detected cancers compared with cancers detected in the nonscreening group may be misleading because the increase may be due to the addition of the lead time, and occur even in the absence of a mortality reduction, as was recently shown for thyroid cancer screening in Korea [12].

Surrogate end points
The results of screening are often eagerly awaited both by the medical community and by society. The political pressure to roll out screening programs is indeed often substantial. Therefore, end points that presage a mortality reduction are attractive. Among such measures that are widely used are incidence or proportion of early-stage cancer among screened and unscreened individuals, and detection rates of cancer precursors such as colorectal polyps or Barrett’s esophagus. The validity of these measures, however, is difficult to ascertain. To rely on early-stage cancer as a proxy for screening effectiveness is naive and misleading because the incidence and proportion are predictably higher in the screened group as a result of lead-time bias and possibly also to overdiagnosis of indolent cancers and length-time bias (Fig. 1). Screening studies have repeatedly and convincingly shown excesses of early-stage cancer even when no subsequent mortality reduction was documented [13, 14]. With regard to precursors, only the adenoma detection rate has been shown to be associated with subsequent cancer risk [15]. This association may not be linear [16] and most certainly has a threshold. It also includes an element of overdiagnosis and overtreatment of polyps, which requires resource use and increases the number of complications.

Because the majority of future deaths are likely to occur among patients with these disease stages, however, because a lower occurrence of advanced disease following screening does not guarantee a mortality reduction, this surrogate end point should ideally be validated before it is used more widely. Such validation can best take place if randomized trials document a mortality benefit that is preceded by lower incidence of advanced disease.

Study design
As the gold standard for evaluation of the benefits and harms of screening (comparing individuals invited to screening with those not invited in randomized trials) is resource-demanding and requires a long follow-up, it is appealing to rely on observational study designs, notably case–control studies, which are cheaper and faster than randomized trials. However, such studies are problematic in screening, as they typically compare those who attend screening with those who do not. This approach is often flawed because cancer screening compliers have different mortality compared with noncompliers as a result of differences in background disease risk [14]. We agree with the World Health Organization, which concluded in their 2002 report on screening that “observational studies based on individual screening history, no matter how well designed and conducted, should not be regarded as providing evidence of an effect of screening” [17].
An example from CRC screening may help to illustrate the problems of applying case–control study designs in screening. During recent years, four large randomized trials have demonstrated that sigmoidoscopy screening reduces CRC mortality by 28% [10]. This is a remarkable effect and will presumably elevate sigmoidoscopy in future guidelines. However, before these results of randomized trials were available, it was believed that sigmoidoscopy would be considerably more effective because of a case–control study published in the New England Journal of Medicine in 1992, which indicated an effect of 60% to 70% [18]. There is certainly an important difference between 28% and 60% effectiveness. In the era of informed consent and shared decision making, it is important that the effectiveness estimates of screening interventions conveyed to patients, caregivers, and providers, are as valid as possible. Case–control studies have difficulties in delivering such estimates.

Conclusions

Evaluation of the benefits and harms of endoscopic screening for gastrointestinal cancers is important, but entails a number of specific challenges in study design and choice of end points. Pitfalls are many and frequently result in overestimates of benefits with too little attention paid to harms and overdiagnosis as a result of screening. Researchers performing studies on endoscopic screening for gastrointestinal cancers should avoid the following pitfalls because they entail a high risk of bias and thus may give incorrect study results:

1. using surrogate measures such as early-stage cancer incidence or detection rates of precursors (e.g. colorectal polyps or Barrett’s dysplasia)
2. using survival as an end point
3. comparing individuals who comply with screening with those who do not comply
4. comparing incidence of cancer or precursors in groups of individuals with different screening or surveillance schemes. Future studies should focus on randomized trial designs, using comparative effectiveness methodology if several screening tests are compared with each other, and should include only valid end points such as cancer-specific mortality and all-cause mortality. Such studies may be more resource-demanding and require longer follow-up and more sophisticated methodology than many uninformative studies performed today. But there are no short cuts if valid estimates of the benefits and harms of screening are to be achieved. And validity is what matters most.

Competing interests: None

Institutions
1 Department of Health Management and Health Economy, Institute of Health and Society, University of Oslo, Oslo, Norway
2 Department of Transplantation Medicine, Oslo University Hospital, Oslo, Norway
3 Department of Epidemiology, Harvard T.H. Chan School of Public Health Community, Boston, Massachusetts, USA
4 Department of Research, Telemark Hospital, Skien, Norway
5 Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden

References
4 Gustafsson L, Adami HO. Natural history of cervical neoplasia: consistent results obtained by an identification technique. Br J Cancer 1989; 60: 132–141
7 Bretthauer M, Hoff G. Comparative effectiveness in cancer screening programmes. BMJ 2012; 344: e2864
8 Jørgensen KJ. Flawed methods explain the effect of mammography screening in Nijmegen. Br J Cancer 2011; 105: 592–593
10 Holme Ø, Bretthauer M, Fretheim A et al. Flexible sigmoidoscopy versus faecal occult blood testing for colorectal screening in asymptomatic individuals. Cochrane Database Syst Rev 2013; 9: CD009259