# Methodological aspects of emulating target trials for causal evaluation of cancer screening programs using health claims data

Thesis submitted to the Faculty of Mathematics and Computer Science, University of Bremen in partial fulfillment of the requirements for the degree of

Doktor der Naturwissenschaften (Dr. rer. nat.)

by

**Malte Braitmaier**

Leibniz Institute for Prevention Research and Epidemiology - BIPS

Braitmaier M (2024), Methodological aspects of emulating target trials for causal evaluation of cancer screening programs using health claims data, Dissertation, Faculty of Mathematics/ Computer Science, University of Bremen

Date of thesis submission: July 3rd 2024
Date of thesis defense: September 5th 2024

1st reviewer: Prof. Dr. rer. nat. Vanessa Didelez
Leibniz Institute for Prevention Research and Epidemiology – BIPS
2nd reviewer: Prof. Sonja Swanson, PhD
University of Pittsburgh, School of Public Health

# Acknowledgements

First and foremost, I want to thank my thesis supervisor Prof. Vanessa Didelez for her guidance throughout my PhD. She always found the time and patience to listen and assist with any issues and continually encouraged new ideas. I am honored to have her as my ”Doktormutter”.

I am thankful to Prof. Sonja Swanson, who has kindly agreed to review this thesis.

I also want to thank Prof. Ulrike Haug, who was always willing and able to find time for me and my projects, all the while also supervising the PhD candidates from her own department and juggling dozens of appointments every day. It was a pleasure and inspiration to be able to work with her.

Of course I also want to thank Prof. Iris Pigeot for continually guiding me, beginning with my Master’s through to my PhD. Her leadership and open communication style during the COVID pandemic were paramount to enduring those difficult times.

Furthermore, I want to thank Dr. Bianca Kollhorst. Her pragmatic leadership style was essential in order to navigate project work, teaching and finding time for my thesis.

I am grateful to everyone in the BIPS family and to all the external collaborators with whom I had the privilege of working with. While too numerous to list here, they all helped me learn and grow and I am thankful for the time spent together.

Ich bin unendlich dankbar für meine Familie, die immer für mich da ist und mich unterstützt. Ganz besonders dankbar bin ich für meine Mutter: Ohne dich wäre nichts möglich gewesen.

Lastly, and most importantly, I am thankful for my girlfriend, who shares all ups and downs and is the dearest and most special person to me.

# Contents

# Abstract

Cancer screening affects cancer-related outcomes by detection at an early – possibly even pre-cancerous – stage, thereby enabling timely treatment initiation or removal of precursors. Ideally, efficacy of cancer screening programs should be assessed in randomized controlled trials before population-wide implementation. In the absence of trial evidence, or when interest lies in real-world effectiveness, observational data must be used to assess the effectiveness of existing programs. However, limitations of observational study designs and data sources must be addressed.

Issues relating to unclear research questions or incorrect temporal alignment of study design elements have been identified as a common source of major bias in non-interventional research in recent years. Target trial emulation has been proposed as a framework to formulate clear and precise estimands by defining the study protocol of a hypothetical target trial that would answer the research question at hand and emulating said target trial as best as possible using observational data.

As part of this thesis, I developed a detailed study design for the evaluation of the German mammography screening program regarding its effect on breast cancer-related mortality. Furthermore, I conducted an extensive, realistic simulation study to assess the potential of residual immortal time bias due to the coarse granularity of discrete time available in the underlying database. Next, I emulated a target trial to assess the causal effect of screening colonoscopy on the incidence of colorectal cancer. Differing effectiveness by site of the tumor was reported in previous observational studies on screening colonoscopy. I showed in the present thesis that previous observational studies overestimated the effect of screening colonoscopy and that the difference by site was largely a result of design-induced bias. I extended the initial study design to more complex settings with a sustained no screening strategy and to strategies incorporating the quality of colonoscopy. Finally, I conducted substantive sensitivity analyses tailored to the specific study design and research question, e.g. concerning residual confounding bias, strengthening confidence in the validity of findings.

# Zusammenfassung

Krebsscreening zielt darauf ab, inzidente Tumorerkrankungen in einem möglichst frühen Stadium zu entdecken, damit eine Behandlung begonnen werden kann so lange die Prognose günstig ist, oder aber bereits Vorstufen zu erkennen und direkt zu entfernen. Vor Einführung eines Krebsscreening Programmes sollte dessen Wirksamkeit in randomisierten kontrollierten Studien (RCT) nachgewiesen werden. Ist dies nicht geschehen, oder soll die Wirksamkeit existierender Programme unter realen Bedingungen in der Bevölkerung untersucht werden, müssen dafür i.d.R. Beobachtungsdaten genutzt werden. Dies geht mit einer sorgfältigen Abwägung eventueller Limitationen nichtinterventioneller Studiendesigns und Datenquellen einher.

In den vergangenen Jahren wurden Probleme im Design von Beobachtungsstudien als Quelle starker Verzerrungen identifiziert. Hierbei sind insbesondere eine unklare Definition der Forschungsfrage und Selektionseffekte durch eine unlogische zeitliche Anordnung von Elementen des Studiendesigns zu nennen. Als Lösung bietet sich das Emulieren von Target Trials an, wobei zunächst das Studienprotokoll eines hypothetischen RCTs inklusive einer klar formulierten Forschungsfrage entwickelt wird, welches dann so exakt wie möglich mit Beobachtungsdaten umgesetzt oder emuliert wird.

Die vorgelegte Dissertation besteht aus mehreren eigenen Forschungsleistungen. Im Rahmen meiner Dissertation habe ich ein detailliertes Studienprotokoll zur Effektivitätsevaluierung des deutschen Mammographie Screening Programms bzgl. der Brustkrebsmortalität erarbeitet. Weiterhin habe ich in einer umfangreichen Simulationsstudie untersucht, ob die vergleichsweise starke Vergröberung diskreter Zeit in der verfügbaren Datengrundlage zu residualem Immortal Time Bias führen kann. Darüber hinaus habe ich einen Target Trial zur Beurteilung des kausalen Effekts der Screening Koloskopie auf die Darmkrebsinzidenz emuliert. Frühere Beobachtungsstudien hatten auf einen deutlich größeren Effekt im distalen Teil des Kolons hingewiesen. Ich konnte hingegen zeigen, dass das Studiendesign dieser früheren Berichte zu Verzerrungen geführt hat, wodurch die Effektivität von Screening Koloskopien insgesamt überschätzt wurde. Die unterschiedlichen Effektschätzer für

den distalen und proximalen Teil des Kolons sind auf diese Verzerrungen zurückzuführen. In Folgestudien konnte ich das ursprüngliche Design der Emulierung auf komplexere Expositionen ausweiten. Insbesondere habe ich den Effekt einer anhaltenden Nichtteilnahme modelliert, aber auch die Qualität der Koloskopie in die Expositionsdefinition einbezogen. Die Target Trial Emulierung habe ich mit vielfältigen und auf die Fragestellung, Datenquelle und Studiendesign maßgeschneiderten Sensitivitätsanalysen untermauert, um die Gefahr durch z.B. ungemessene Störgrößen beurteilen zu können. Die Ergebnisse dieser Sensitivitätsanalysen deuteten auf eine hohe Robustheit der Ergebnisse hin.

# CHAPTER 1

# Introduction

Health claims data and other routinely collected real-world data (RWD) provide a rich source of individual-level health information. These data cover diagnoses, medications and operations or other treatments [Haug and Schink, 2021; Pigeot and Ahrens, 2008]. Using them to assess the causal effect of exposures on health outcomes, however, poses unique challenges to study design and statistical methods [Schneeweiss and Avorn, 2005].

In the context of RWD and causal inference, a special focus must be put on the intersection of applied knowledge and statistical methods. In many cases it is not immediately clear how a subject matter question can be translated into statistical language and off-the-shelf methods may be inadequate. Collaboration between subject-matter experts and statisticians is required to clearly define the research question. Next, bespoke study designs and statistical solutions need to be tailored to fit to the target of inference.

## 1.1 Defining the target of inference

The first step in any causal analysis must be to define the target of inference, i.e. the estimand [Faries et al., 2020]. In the following, the estimand will refer to the unknown quantity which is to be estimated and the estimate will refer to the effect estimate produced by the statistical estimation procedure. Any systematic difference between the estimator and estimand will be referred to as bias. Specifying the estimand entails several steps, first of which is a clear and precise definition of an exposure of interest, corresponding to a realistic – possibly hypothetical – intervention, which must be represented in the available data [Hernán and Robins, 2020]. It must be defined whether the exposure is a once-only event, or is sustained over time. In the latter case, a strategy on how to address non-adherence during follow-up must be defined [Hernán and Robins, 2020]. Artificial censoring for non-adherence might be required, but is generally informative and leads to bias, which makes adjustment for time-varying confounders necessary [Joffe, 2001; Hernán and Robins, 2020]. Besides non-adherence, the definition of the estimand must also consider other intercurrent events, such as competing events. While these are often treated as censoring events in applied studies, this might not correspond to a meaningful causal effect given that it corresponds to a hypothetical scenario under which competing events can be eliminated. Instead, approaches that do not treat competing events as censoring events, but incorporate the effect of the exposure on the outcome mediated by the competing event in the estimand may be preferable [Young et al., 2020].

After the exposures of interest have been defined succinctly, the outcome and contrast of interest must be defined. The effect of interest can be expressed as an absolute effect such as the absolute risk reduction (ARR), or as a relative effect such as the relative risk (RR). Generally, contrasts based on risks should be preferred to contrasts of hazards for causal inference, since hazards are always conditional on the event of interest (and competing events) not having occurred yet and, thereby, have a ”built-in” selection bias [Hernán, 2010; Aalen et al., 2015].

The above-described choices must be made explicit. This can be achieved using the target trial emulation (TTE) framework, which aims at applying design elements of a randomized controlled trial (RCT) to observational studies. In this approach, the study protocol of the ideal randomized trial – the target trial – is defined first and then emulated using observational data [Hernán and Robins, 2016]. The study protocol contains information on the most important design elements of the target trial, such as eligibility, treatment strategies, follow-up, outcome variable, contrast of interest, and the statistical analysis. When specifying the emulated trial, any deviations from the target trial become immediately apparent and any impact on the estimates can be discussed or assessed in appropriate sensitivity analyses [Didelez, 2016]. Furthermore, the resulting study protocol for the emulated trial must contain information on which confounding variables are required to emulate randomization. While the TTE framework is a useful tool for causal inference in general, it also provides a structured template for applied researchers and statisticians to speak the same language.

## 1.2 Estimation of causal effects

Some features of the TTE framework need special emphasis. First, the concept of cloning is introduced to maximize statistical efficiency. While one person can only be included in an RCT once and be randomized into only one treatment arm, the same person can be included in an emulated trial repeatedly. For instance, sequential trials may be emulated over time to best use the information contained in a longitudinal database and individuals may be included in more than one emulated trial if they are eligible at the respective baseline. Furthermore, the exposure status at the baseline of an emulated trial might be consistent with more than one exposure strategy, in which case information from these individuals is copied and cloned into all exposure strategies that are consistent with the observed exposure. To distinguish identical individuals who were included in more than one emulated trial or under more than one exposure strategy, the terms person-trial and clone are used, respectively [Hernán et al., 2016; García-Albéniz et al., 2020; Danaei et al., 2013].

Next, artificial censoring and appropriate adjustment (e.g. via inverse probability of censoring weighting (IPCW)) may be used to adjust for non-adherence during follow-up, if the target of interest is a per-protocol (PP) effect. Since censoring is informative in the presence of time-varying covariates that affect both adherence and the outcome variable, appropriate confounder adjustment needs to be carried out. When using IPCW for adjustment, propensity scores (PSs) for the probability of adhering to the exposure strategy are estimated for each time point and weights are constructed by taking the inverse of the cumulative product of these PSs [Robins et al., 2000]. Alternatively, the parametric g-formula may be used to adjust for time-varying confounding, albeit at higher computational cost [Robins, 1986].

Regarding the type of outcome variable, TTE analyses most commonly feature time to event variables. As mentioned above, contrasts based on hazards are not ideal for causal inference purposes. As an alternative, pooled logistic regression is useful to estimate flexible functions of risk [D'Agostino et al., 1990]. This approach also allows estimation of contrasts at any point during follow-up, so that time-varying treatment effects can be visualized easily.

Another difficulty in causal inference from observational data is the selection of a sufficient set of confounders, i.e. a set of covariates that, when adjusted for appropriately, allows identification of the causal effect. While data-driven causal-discovery exists, the preferred way for covariate selection is via subject matter knowledge [Witte and Didelez, 2019]. When the causal relations between variables are known, confounders can be selected without further assumptions regarding any data-driven method. A tool to help with subject matter-motivated variable selection is the use of directed acyclic graphs (DAGs) [Pearl, 1995].

Finally, when cloning individuals and including them in the analysis dataset repeatedly, confidence intervals must be estimated using robust methods. For this, individual level bootstrapping is commonly used. It is important to note that samples must be drawn at the level of individuals and not at the level of clones to ensure that the assumption of random sampling with equal weights per individual is fulfilled [Efron, 1979]. Furthermore, the procedure of estimating inverse probability weights and obtaining marginal effect estimates must be repeated for each bootstrap sample [Murray et al., 2021]. The fact that bootstrapping is a computationally heavy procedure in combination with the large datasets commonly used for RWD studies poses further challenges to the statistical programming. For instance, a random subsample might be used instead of the entire study population to reduce the computational cost, if sample size allows it (see, e.g. García-Albéniz et al. [2017a] and Braitmaier et al. [2022b]).

## 1.3 Potential sources of bias

Several potential sources of bias must be considered carefully when planning a causal analysis of observational data. The most frequently discussed source of bias stems from a violation of the exchangeability assumption, i.e. from confounding. Relevant confounders need to be identified, preferably via subject matter knowledge about the causal structure between variables. Furthermore, these confounders need to be measured in the data. The so-called healthy screenee bias deserves special emphasis when evaluating cancer screening programs [Weiss and Rossing, 1996; Shrank et al., 2011]. Health-conscious individuals are both more likely to undergo voluntary screening and less likely to develop cancer, due to a healthier lifestyle. Health consciousness and related health factors are, therefore, important confounders. However, health consciousness itself is not an easily measured or quantifiable variable, and is not available in routinely collected data and, therefore, needs to be approximated as best as possible using proxy codes.

Besides confounding, great consideration should be given to self-inflicted biases resulting from inappropriate study design. In particular, time-related biases occur when basic elements of the study, particularly the assessment of eligibility, treatment assignment and start of follow-up are not aligned at a clear time zero. For example, immortal time bias results when the exposure assessment uses information from after baseline, since individuals assigned to the exposed strategy due to an exposure long after baseline cannot have died before their exposure, leading to an accumulation of early deaths in the control strategy. Bias also results when exposure assessment uses information from before baseline [Hernán et al., 2016]. For instance, if an analysis on the effectiveness of colonoscopy screening were to count individuals who underwent screening before baseline as exposed, but at the same time excluded individuals who had a colorectal cancer (CRC) diagnosis before baseline from the study, there would be a depletion of ill individuals among the screened but not the unscreened [García-Albéniz et al., 2017b].

## 1.4 Aim of this thesis

The thesis addresses methodological problems in the context of assessing the effectiveness of cancer screening programs. The German Pharmacoepidemiological Research Database (GePaRD) [Pigeot and Ahrens, 2008; Haug and Schink, 2021] was used to emulate target trials from observational data. The main and original contributions are:

1. I designed and conducted an emulated target trial on the site-specific effect of screening colonoscopy on CRC incidence (see Braitmaier et al. [2022b] and Chapter 4)
2. I carried out an extensive set of assumption checks and sensitivity analyses to assess validity of study results (see Section 4.4)
3. I extended the initial study design to a per-protocol analysis with sustained non-exposure (see Section 4.5)
4. I demonstrated that differences in site-specific effectiveness reported in previous work were a result of design-induced biases (see Braitmaier et al. [2024] and Section 4.6)
5. I extended the initial study design to include more than two exposure categories to contrast low and high quality colonoscopy (see Schwarz et al. [2024] and Section 4.7)
6. I developed a study protocol for an emulated target trial assessing the effectiveness of the German mammography screening program (see Braitmaier et al. [2022a])
7. I conducted a simulation study to quantify the potential of residual immortal time bias due to coarse granularity of discrete time in the context of the emulated trial on screening mammography (see Section 5.1)

## 1.5 Structure of this thesis

The thesis is structured as follows: Chapter 2 gives background on estimation of causal effects from observational data, with some general explanation of the estimation procedures used in the context of this thesis. Furthermore, some common sources of bias - such as confounding - are introduced in the chapter. Chapter 3 is an introduction to target trial emulation. The original contributions of the present work are covered in Chapters 4 and 5, with Chapter 4 focusing on work related to the evaluation of screening colonoscopy and Chapter 5 focusing on screening mammography. A discussion and outlook is given in Chapter 6. Finally, the publications contributing to this thesis are printed in Chapter 7.

## 1.6 Funding and competing interests

The studies relating to colonoscopy screening that are presented in this thesis were funded by BIPS intramural funds. The study protocol for the evaluation of the German mammography screening project was funded by the Federal Office for Radiation Protection from funds of the Federal Ministry for the Environment, Nature Conservation and Nuclear Safety, the Federal Ministry of Health and the Kooperationsgemeinschaft Mammographie (supported by the National Association of Statutory Health Insurance Funds and the National Association of Statutory Health Insurance Physicians), grant numbers 3617S42402 and 3617S42410.

There are no conflicts of interest to declare.

# CHAPTER 2

# Estimation of causal effects from observational data

## 2.1 What is a causal effect?

In an interventionist understanding of causal effects, the occurrence of $Y$ can be partially controlled by intervening on $A$ when $A$ causes $Y$ [Hernán, 2004]. One can imagine two alternative scenarios for an individual $i$ in which $A_i$ is either set to 1 or to 0. When $A_i$ is set to 0, the potential outcome is given by $Y_i^{A=0}$, while the potential outcome in the opposite scenario is given by $Y_i^{A=1}$. An individual-level causal effect is present if $Y_i^{A=1} \neq Y_i^{A=0}$. Only one potential outcome can be observed in one individual [Hernán and Robins, 2020].

Medical research generally aims to estimate group-level rather than individual-level causal effects. The average potential outcome had the entire study population (or some subgroup of interest) been exposed is compared to the average potential outcome had the entire study population not been exposed, i.e. a causal effect of $A$ on $Y$ is present, if $\mathbb{E}\left[Y^{A=1}\right] \neq \mathbb{E}\left[Y^{A=0}\right]$. The challenge arises to estimate causal effects from observed data, as only one potential outcome is observable per person. Since exposed and unexposed individuals might differ in variables other than exposure, tools are needed to disentangle the causal effect of exposure from non-causal associations due to confounding or inappropriate study design. This is the objective of causal inference and the fundamental assumptions required for the most frequently used methods will be laid out in this chapter.

## 2.2 Directed acyclic graphs

Directed acyclic graphs (DAGs) are used to visually illustrate the (assumed) causal structure between variables [Pearl, 1995]. See Figure 2.1 for a simple DAG with three nodes. DAGs are a valuable tool to reach clarity regarding the causal relations between variables and enable the identification of bias sources due to confounding or design-induced selection effects. This latter property was exploited in this thesis when exploring design-induced biases in the context of screening colonoscopy (see Sections 4.6 and 7.6). DAGs do not display information on the strength of associations, but give a qualitative representation of causal relationships. A DAG consists of nodes and directed edges or arrows connecting these nodes. While the presence of a directed edge between nodes indicates that one directly causes the other, the absence of edges is equally relevant for the analyst, as it indicates (conditional) independence and absence of direct causation (by the causal Markov property) between these nodes. The causal Markov property states that a node is independent of any node that is not its descendant, if conditioned on all of its direct causes [Hernán and Robins, 2020]. Undirected or bidirected edges are sometimes used to illustrate the presence of further, unmeasured variables that have a causal relationship with both nodes connected by the edge.

Some nomenclature is required: In the graph $\mathcal{G} : A \to Y$, $A$ is a parent to the child $Y$. Any node $A$ that precedes a node $Y$ on a directed path is called an ancestor to $Y$, while $Y$ is a descendant of $A$. On a path containing node $C$ with arrows converging in it $(... \to C \leftarrow ...)$, node $C$ is called a collider.

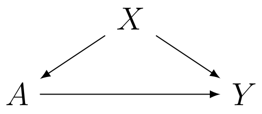


**Figure 2.1:** Basic DAG showing the relations between an exposure $A$, an outcome $Y$ and a confounding variable $X$.

An important concept to read off conditional independence from graphs is $d$-separation. If a graph contains the nodes $A$ and $Y$ and a set $Z$ is considered, $A$ is d-separated from $Y$ by $Z$ if a node $w$ fulfilling one of the two following criteria exists on every path from $A$ to $Y$:

1. Node $w$ is a collider, meaning that it has converging arrows into it and neither $w$ nor its descendants are contained in $Z$.

2. Node $w$ is not a collider and $w$ is contained in $Z$.

If the criteria for $d$-separation and the causal Markov property are fulfilled, $A$ is independent of $Y$ conditional on $Z$, i.e. $A \perp\!\!\!\perp Y|Z$.

Based on a DAG and using the concept of $d$-separation, a valid adjustment set $Z$ can be identified. In this context, the so-called *backdoor criterion* becomes relevant: A set $Z$ is said to fulfill the backdoor criterion regarding the causal effect of $A$ on $Y$, if $Z$ does not contain descendants of $A$ and if all paths between $A$ and $Y$ with an arrow into $A$ are blocked by $Z$ [Pearl, 1995; Peters et al., 2017].

## 2.3 Clear definition of exposure, outcome and intercurrent events

While seemingly obvious, it is important to stress that the target of inference, i.e. the causal estimand, needs to be specified before any effect estimation can be conducted. The estimand is defined as that which is to be estimated [Hernán and Robins, 2020; Rubin, 2005]. Any ambiguity in the definition of the estimand potentially leads to inappropriate analyses or misinterpretation. The complexity of clearly defining the estimand received more attention in the biostatistics community after the addendum to the ICH E9 (R1) guideline on estimands was published, requiring great care to define estimands in clinical research [ICH E9 (R1), 2020], although the concepts contained in the guideline have been known for a much longer time. The guideline mentions several aspects of estimands that need to be clearly specified at the planning stage of a study, namely the treatment strategy under investigation, the population to which the clinical question relates, the clinical endpoint of interest, intercurrent events and how they are incorporated in the research question and, lastly, the effect measure to be assessed.

Relating to the definition of treatment or exposure under investigation it is important to note that beyond specifying the medicinal product or potential health hazard itself, one must also define in what way study participants shall be exposed to it. Exposure could, for instance be the mere offer of receiving a medicinal product, uptake of at least one dose at baseline or sustained exposure over a certain amount of time. Both the interpretation of results and the statistical methodology appropriate for the study will differ depending on the exact definition of the exposure of interest [Goetghebeur et al., 2020].

Next, the population of interest must be defined. This has two-fold relevance: On the one hand, at the planning stage of the study individuals must be recruited into the study so as to be representative of this target population. On the other hand, if some patient groups cannot be included in the study for any reason and the trial eligible group differs from the target population, the latter needs to be clearly defined to assess under which circumstances study results apply to the target population. If the study population and the eligible population do not substantially differ, it is reasonable to assume that the study results can easily be applied to the entire eligible population, i.e. the study results are likely to be generalizable [Dahabreh et al., 2019]. If, furthermore, study eligible population and target population are identical, the study results are directly relevant to the target population. When transporting (referring to transportability as opposed to mere generalizability) study results beyond the trial eligible population, additional statistical care is needed [Dahabreh et al., 2019].

While the ICH E9 (R1) guideline relates to RCTs, further specifications regarding the estimand are generally required in observational studies. For instance, when obtaining marginal, population-level effect estimates, the population of interest needs to be specified. If marginal estimates are obtained in relation to the covariate distribution of the entire sample, these estimates relate to the average treatment effect (ATE). If, however, marginal estimates are obtained in relation to the covariate distribution of the subset of individuals who received treatment, these estimates relate to the average treatment effect on the treated (ATT). In an RCT without differential non-compliance, ATE and ATT are not expected to differ due to the randomization process, but in observational studies they are likely to differ [Li et al., 2022]. Similarly, the marginal effect in the population of untreated individuals, i.e. the average treatment effect on the untreated (ATU) might be relevant in many settings as well [Wang et al., 2017].

Next, an intercurrent event occurs after treatment initiation and possibly affects the occurrence or observability of the outcome of interest. As such, the term intercurrent event describes a wide variety of events, such as exposure to drugs other than the one being studied, treatment discontinuation due to adverse events, development of contraindications, experience of competing events and many more. Given that the occurrence of intercurrent events may be affected by exposure and may in turn affect the outcome of interest, careful consideration is required when defining the target of inference. Different strategies of treating intercurrent events yield answers to different causal questions.

The ICH addendum, which was specifically developed for randomized trials, was preceded by a rich literature on causal inference from observational data, one fruit of which is target trial emulation. In this framework, estimands are clearly defined by explicitly specifying the ideal randomized trial with corresponding intervention and then emulating this target trial as closely as possible using available observational data. Sometimes, the available observational data is not sufficient for reliable estimation, necessitating the choice of a different estimand. Target trial emulation will be covered in depth in Chapter

3.

## 2.4 Identifying assumptions

Some strong and often unverifiable assumptions are needed in order for the causal effect to be identifiable from the available data. While different methods require different assumptions, the below identifying assumptions need to be fulfilled when using the methods described in e.g. Braitmaier et al. [2022a] and Braitmaier et al. [2022b].

Under consistency the exposure of interest is well-defined, is observed in the data and can be intervened upon. More formally, if the observed exposure of an individual is $A = a$, then the consistency assumption states that $Y^{A=a} = Y$, i.e. the potential outcome under the exposure value that was indeed observed is consistent with the realized outcome value observed in the data. If the exposure of interest is a static strategy sustained over multiple time points $k \in \{1, ..., K\}$, consistency assumes that $Y^{\bar{A}=\bar{a}} = Y$ for individuals with the observed exposure history $\bar{A} = \bar{a}$. The term ”well-defined” in the context of consistency also means that the exposure observed in the data does not consist of different versions. When studying the effect of a reduction in body mass index (BMI) without defining which intervention leads to this reduction, problems in the interpretation arise. Some individuals might achieve the reduction by means of bariatric surgery, while others do so via lifestyle changes. The results of an analysis assessing this poorly defined exposure would not be informative, since it is unclear how much of the effect was achieved via which version of exposure. Consistency, then, can be achieved by careful and diligent planning of the study [Hernán and Robins, 2020].

The aspect of consistency relating to multiple versions of treatment also relates to the assumption of no interference, which states that the potential outcome of one individual is not affected by the treatment of another [VanderWeele and Hernán, 2013]. If we assume that the no interference assumption does not hold, there is a near-infinite number of alternative versions of treatment for each individual, depending on the exposure values observed in the rest of the study population. In most settings, it is implicitly assumed that no interference is present and no explicit mention of it is made. However, in some particular settings this might not be the case. In studies on the effectiveness of vaccines, vaccination of other members of the study population affects an unvaccinated persons risk of infection via increased herd immunity.

Positivity describes the assumption that all levels of exposure are observed in all strata of covariates used for confounder adjustment, or more precisely

$$\mathbb{P}\left[A = a|X\right] > 0$$

for all $a$ in all strata of the covariate vector $X$ observed in the data (i.e. all $x$ with positive probability density $f(x) > 0$). For exposures sustained over time,

$$\mathbb{P}\left[A_k = a_k|\bar{A}_{k-1}, \bar{X}_k\right] > 0$$

for all $\bar{a}_{k-1}$ and $\bar{x}_k$. Unlike with consistency, positivity can be empirically verified using the data. For instance, the occurrence of any level of a covariate can be assessed in each exposure group. Furthermore, overlap of the PS distributions should be checked (more details on PSs will be given below) [Hernán and Robins, 2020].

Exchangeability is formally defined as

$$Y^{A=a} \perp\!\!\!\perp A$$

for all $a$, i.e. the potential outcome had exposure $A$ been set to $a$ is independent from the observed value of $A$. This form of exchangeability, sometimes referred to as ”full” exchangeability, is achieved via randomization in an RCT with $A$ being assignment to treatment (rather than treatment received). It is usually not fulfilled in observational studies, assuming that both outcome and exposure are affected by covariates $X$. In an observational setting, conditional exchangeability holds, if $X$ forms a sufficient adjustment set, i.e.

$$Y^{A=a} \perp\!\!\!\perp A|X$$

holds for all $a$. If covariates other than $X$ exist that are not observed in the data and that confound the relationship between $Y$ and $A$, this assumption is violated. The conditional exchangeability assumption, therefore, is sometimes referred to as the ”no unobserved confounding” assumption [Hernán and Robins, 2020]. If the exposure is sustained over time, adjustment for time-varying covariates is typically necessary. In this setting, exchangeability is extended to the time-varying setting and is referred to as sequential exchangeability, since conditional exchangeability must hold at every time point. First, the time-varying exposure strategies under investigation must be clearly defined. Exposure at time $k$ might depend on past exposure and past and concurrent covariates, i.e. a strategy $g$ might be defined as $A_k = g(\bar{A}_{k-1}, \bar{X}_k)$, where the overbar represents the history of a variable. Then, the sequential exchangeability assumption states that

$$Y^{G=g} \perp\!\!\!\perp A_k | \bar{A}_{k-1} = g(\bar{A}_{k-2}, \bar{X}_{k-1}), \bar{X}_k$$

holds for all $g$ and $k$ [Hernán and Robins, 2020]. Exchangeability cannot be tested using the observed data, since the key assumption of no unobserved confounding refers to things that are not available to the analyst. Instead, the plausibility of this assumption being fulfilled needs to be judged on subject matter knowledge regarding potential confounders and outcome predictors. Additionally, sensitivity analyses such as negative control analysis can detect violations of this assumption (see Section 2.7.1). It is noteworthy that methods such as instrumental variable analysis exist that do not make the conditional exchangeability assumption. These methods, however, make other strong and unverifiable assumptions.

## 2.5 Estimation procedures

In contrast to RCTs, in observational studies one must generally assume that at least some variables that predict the outcome of interest are not distributed equally across comparison groups. In this case, a naive estimator that does not appropriately adjust for the confounding influence of covariates will be biased. While methods exist to obtain unbiased estimators when (some) confounders are not observed in the data (e.g. instrumental variable methods), most approaches assume that sufficient confounder information is measured so as to adjust for confounding in the analysis. Methods that use observed covariates for adjustment, relying on the exchangeability assumption described above, can be broadly divided into methods that model the outcome, such as regression adjustment, and methods that model the exposure, i.e. propensity score methods. Doubly or multiply robust methods that combine the two approaches exist [Goetghebeur et al., 2020].

Regression adjustment is arguably the most commonly taught method of confounder adjustment and follows the philosophy of outcome modeling. Covariates are included in the model equation of a parametric model and their influence on the outcome is modeled jointly with the effect of exposure [Goetghebeur et al., 2020]. For a binary outcome variable $Y$, consider the following logistic model

$$\mathbb{P}\left[Y = 1|A, X\right] = \text{logit}^{-1}(\beta_0 + a\beta_A + x\beta_X). \tag{2.1}$$

The above model represents the simplest case without interactions between covariates and exposure, in which $\beta_A$ is an effect measure for exposure $A$.

An alternative approach to adjusting for unbalanced covariates is the use of PS methods, such as PS matching or inverse probability weighting (IPW). The PS is defined as the probability of an individual in the study population, conditional on their observed covariates, to experience the exposure [Rosenbaum and Rubin, 1983], i.e.

$$\text{PS} = \mathbb{P}\left[A = 1 | X\right]. \tag{2.2}$$

It is important to note that the goal of the PS is not to perfectly predict the observed exposure. Instead, it has a dual property as a balancing score, meaning that conditional on the PS, the distribution of baseline covariates $X$ will be balanced between exposure categories, i.e., $X \perp\!\!\!\perp A|\text{PS}$. If the PS is estimated via a misspecified model, it might not fulfill this balancing criterion [Wyss et al., 2014]. To the contrary, Imai and Ratkovic [2014] note that even mild misspecification of the propensity model can lead to substantial bias and the propensity model should be selected so as to maximize covariate balance. The performance of the PS needs to be assessed, e.g. via balance checks using the absolute standardized mean difference [Zhang et al., 2019]. If the PS does not balance covariates, a different model specification must be chosen [Wyss et al., 2014]. While the PS is commonly estimated via logistic regression, other approaches including machine learning methods have been proposed [Lee et al., 2010; Pirracchio and Carone, 2018], each with their own strengths and limitations.

Many methods for confounder adjustment based on the propensity score have been developed. Propensity score matching gained particular popularity in applied work, presumably because of its ease of use. In 1:$n$ PS matching, one exposed individual is matched to $n$ unexposed individuals based on their PS. Since no exact match based on the real-numbered PS is to be expected, matching is either done on PS strata (e.g. quintiles) or uses a caliper width [Austin, 2011]. Both optimal and greedy matching algorithms are available, but given that optimal matching usually does not perform substantially better than greedy matching and that optimal matching can become computationally prohibitive, greedy matching is used in most real-life studies [Austin, 2014; Rosenbaum, 1989]. A limitation of PS matching is that data from some of the study population will not (or only to a small extend) be considered in the analysis, if no or few corresponding matches can be found. More than that, the covariate structure of the matched population will correspond to the covariate structure of the treated population, since matches are selected for all exposed individuals, but not necessarily for all unexposed individuals. Because of this, PS matching is often used to estimate the ATT, but estimation of the ATE requires further methodological considerations.

An alternative PS method that uses data from all individuals in the study population is IPW. When adjusting for baseline confounding, inverse weights are constructed as

$$\frac{1}{\mathbb{P}\left[A=1|X\right]}$$

for the treated and

$$\frac{1}{1-\mathbb{P}\left[A=1|X\right]}$$

for the untreated. In this approach, data from all individuals in the study population is considered and estimation of the ATE, among others, becomes straightforward. IPW can be extended to adjust for time-varying confounding in sustained or time-dependent exposures whereas PS matching is applicable to point exposures only. Furthermore, trimming of weights at the extremes or use of stabilized weights, especially when adjusting for time-varying confounding, limits instability due to extreme weights [Goetghebeur et al., 2020].

Outcome models using inverse probability of treatment weights are also called marginal structural models (MSMs). They are considered ”marginal” in that they model the marginal distribution of potential outcomes. While MSMs can be used to model point exposures, they are especially useful when modeling exposures sustained over or varying with time, in which time-dependent confounding plays a role [Robins et al., 2000].

One example in which time-varying confounding is particularly relevant is a per-protocol study design for sustained exposures, where artificial censoring is used to adjust for non-adherence during follow-up, i.e. individuals are artificially censored if and when they stop adhering to their assigned exposure strategy. If, however, a covariate that affects both exposure or treatment adherence and the outcome of interest changes during follow-up, this covariate affects both the probability to be artificially censored from the dataset due to non-adherence and the probability of experiencing the outcome. This leads to time-dependent confounding. Let $\text{Cens}_t$ be the censoring status at time $t \in \{1, ..., T\}$. Then, a time-varying inverse weight, which considers observed, time-dependent covariates, is given by

$$w_t = \frac{1}{\prod_{l=1}^{t} \mathbb{P}\left[\text{Cens}_l = 0|\overline{\text{Cens}}_{l-1} = 0, \bar{X}_l\right]}. \tag{2.3}$$

Given that these weights, which build the inverse of a cumulative product of probabilities, can become very large, the resulting weighted models may become unreliable. Therefore, stabilized weights are usually used instead. These are defined as

$$sw_t = \frac{\prod_{l=1}^{t} \mathbb{P}\left[\text{Cens}_l = 0 | \overline{\text{Cens}}_{l-1} = 0\right]}{\prod_{l=1}^{t} \mathbb{P}\left[\text{Cens}_l = 0 | \overline{\text{Cens}}_{l-1} = 0, \bar{X}_l\right]}. \tag{2.4}$$

These weights are easily obtained using standard software by generating a modified dataset with one entry per individual and time point and then applying e.g. pooled logistic regression to this modified dataset [Robins et al., 2000], i.e. time is included in the model applied to the longitudinal dataset. Furthermore, separate models are often fitted for each exposure level, which allows covariates to affect the censoring probability in different ways (see e.g. Murray et al. [2021] for a tutorial and Dickerman et al. [2023] for an example of this).

## 2.6 Time-to-event analysis

Assume a study in which time $T$ to an event of interest $Y \in \{0,1\}$ is measured in days, i.e. $T \epsilon \mathbb{N}$. For now, also assume that no other event can prevent the event of interest, e.g. in a study on overall mortality. However, individuals may drop out of the study prematurely or still be event-free at the end of the study period, in which case they are censored at the end of their available follow-up. This is called administrative censoring and is a form of right-censoring [Joffe, 2001]. The censoring status is indicated using a binary variable Cens $\epsilon \{0,1\}$. To explain the analysis of right-censored data, we assume for simplicity's sake that individuals in the study are randomly assigned at baseline to either one of two arms, i.e., $A \epsilon \{0,1\}$.

Let $t$ be the observed realization of the random variable $T$, with the cumulative distribution function, also called cumulative incidence function (CIF), as

$$F(t) = \mathbb{P}\left[T \leq t\right]. \tag{2.5}$$

The survival function is simply the complement of the cumulative distribution function, i.e.,

$$S(t) = 1 - F(t). \tag{2.6}$$

If no loss to follow-up occurred, $F(t)$ is simply estimated by dividing the number of individuals who experience the outcome event by time $t_h$ by the number of individuals at risk at time $0$, where $h \,\epsilon\, \{1, 2, ..., u\}$ indicates the ordered survival times. However, since loss to follow-up does occur in realistic studies, methods to account for this type of censoring need to be used. These methods usually assume that censoring is independent, meaning that individuals censored at time $t$ should not be systematically different from individuals not censored regarding the risk of experiencing the outcome event [Andersen et al., 2012]. Under the assumption of independent censoring, the Kaplan-Meier estimator is used to estimate the survival function as

$$\hat{S}(t) = \prod_{t_h \leq t} \frac{n_h - d_h}{n_h}. \tag{2.7}$$

In equation 2.7, $n_h$ is the number of individuals at risk at time $t_h$ and $d_h$ is the number of events at time $t_h$. While direct adjustment for baseline confounders by including covariate information in a model equation is only applicable to (semi-)parametric models, adjusted or standardized non-parametric Kaplan-Meier curves can be obtained by IPW [Cole and Hernán, 2004]. However, non-parametric estimates of survival curves tend to become unstable especially at later time points, because only few or no events are observed per time point and the resulting step function is often constant over some time points before substantially changing when events are observed at a subsequent time point. Parametric models, such as the pooled logistic regression approach described below, have the advantage of smoothing survival curves over time [Hernán and Robins, 2020].

The discrete-time hazard is

$$h(t) = \mathbb{P}\left[T = t | T \geq t\right]. \tag{2.8}$$

In the absence of competing events, the survival probability is a function of the discrete-time hazards [Suresh et al., 2022] and is defined as

$$S(t) = \prod_{u=0}^{t} 1 - h(t). \tag{2.9}$$

Non-parametric, semi-parametric and fully parametric methods exist to estimate hazards and survival probability. (Semi-)parametric methods have the advantage that they allow incorporating various confounder adjustment methods for scenarios in which random assignment is not given, such as regression adjustment where covariates are included in the outcome model itself. The most common semi-parametric method to model the hazard function is the proportional hazards Cox model. It consists of an infinite-dimensional, non-parametric part (the baseline hazard) and of a $p$-dimensional, parametric part, with $p$ being the number of variables in the covariate matrix $X$ [Tsiatis, 2006].

Even though the proportional hazards Cox model is popular, hazard ratios (HRs) are difficult to interpret causally due to a built-in selection of individuals who did not experience the outcome by time point $t$ [Hernán, 2010]. Complications arising in the context of hazard ratios are discussed more in depth in Section 2.8.

An alternative approach that circumvents the built-in selection bias of the hazard ratio is to estimate and contrast CIFs. By assessing the time-specific cumulative risk, time-varying effects become apparent. Pooled logistic regression is a parametric method of estimating discrete-time hazards, which in turn can be transformed into an estimate of CIFs [D'Agostino et al., 1990]. For the pooled logistic regression to approximate discrete-time hazards well, one must assume that less than 10 % experience the outcome at any given time point [Murray et al., 2021].

A modified dataset is generated in which every individual has one row per observed time point. If, for instance, individual $i = 1$ experienced the outcome at time point five, the modified dataset for this individual would contain five entries:

| $i$ | $t$ | $A$ | $Y$ | Cens |
|---|---|---|---|---|
| 1 | 1 | 1 | 0 | 0 |
| 1 | 2 | 1 | 0 | 0 |
| 1 | 3 | 1 | 0 | 0 |
| 1 | 4 | 1 | 0 | 0 |
| 1 | 5 | 1 | 1 | 0 |

The pooled logistic model is fitted on this modified dataset and time $t$ is included in the model equation. A simple model can be defined as

$$\mathbb{P}\left[Y_t = 0 | Y_{t-1} = 0, A\right] = \text{logit}^{-1}\left(\beta_1 t + \beta_2 a + \beta_3 t a\right). \tag{2.10}$$

From the above we obtain the cumulative risk by building the cumulative product

$$1 - \prod_{l=1}^{t} \mathbb{P}\left[Y_l = 0 | Y_{l-1} = 0, A\right].$$

By including various transformations of $t$ and possibly interaction terms with the exposure variable $A$ or other variables, this approach allows to model dynamic CIFs (see e.g. Braitmaier et al. [2022b] for crossing CIFs). Furthermore, the flexibility achieved by including various transformations of time is needed to compensate for the lack of a non-parametric baseline hazard, which is one aspect that makes Cox models appealing [Murray et al., 2021].

The above pooled logistic regression model is not adjusted for confounding by $X$. To do so, baseline covariates $X$ are either included in the model itself or confounder adjustment is achieved by IPW.

Note that the reason for censoring individuals when they drop out of the study is that researchers want to make inferences about a real-life population in which ”drop-out” does not exist. This same rationale does not apply when dealing with competing events, as discussed in the next section.

### 2.6.1 Competing events

A competing event is any event that prevents the event of interest from happening. That is, if a competing event occurs, the event of interest is not only not observed, it is not defined. Different approaches have been proposed to conduct statistical analyses in the presence of competing events with some authors recommending cause-specific hazard ratios when interested in etiology [Lau et al., 2009]. In the remainder of this thesis, the term ”event-specific” will be used instead of ”cause-specific” to avoid confusion with causal inference terminology. Young et al. [2020] proposed a framework that formalizes causal estimands in a competing events setting.

In studies on non-mortality outcomes, death is a necessary competing event. Let $T > 0$ be the event time and $E$ an indicator for the type of event (e.g. $E = 1$ for the event of interest and $E = 2$ for competing death). Various approaches are available in competing events settings, with two being described in depth in Young et al. [2020]: When estimating the controlled direct effect, one aims at estimating the effect of exposure on the outcome event under a hypothetical scenario in which the competing event was eliminated, i.e. under which the competing event cannot occur. Methodologically, this elimination is achieved by treating competing events as censoring events, which corresponds to setting the event-specific hazard to zero. This is the default in many applied fields. If, however, the competing event is death and given that most inference aims at answering questions in real-world populations, this makes the controlled direct effect difficult to interpret for most settings, because death cannot be eliminated. The total effect, on the other hand, denotes the effect of exposure on the event of interest, while allowing for competing events to occur. This means that the total effect also includes the effect of exposure on the outcome that is mediated by the competing event. Therefore, caution is also required when interpreting the total effect: If the exposure has a strong negative effect on the competing event, the total effect might indicate a spurious beneficial effect on the event of interest, as seen with the example whether smoking can prevent dementia [Rojas-Saunero et al., 2023]. Hypothetically, cancer screening would appear overly beneficial regarding cancer occurrence, if it had a strongly harmful effect on death. It is recommended to assess the total effect on both the event of interest and the competing event [Latouche et al., 2013].

The total effect is defined as a contrast of potential outcomes, where $Y_t^0$ denotes the potential outcome by time $t$ had all study participants been unexposed and $Y_t^1$ had all study participants been exposed [Young et al., 2020]. A relevant contrast could then be the risk difference based on event-specific cumulative incidence functions, given under randomization as

$$\mathbb{P}\left[Y_t = 1 | A = 1\right] - \mathbb{P}\left[Y_t = 1 | A = 0\right]. \tag{2.11}$$

In this case, two approaches are available: One using the event-specific hazard [Putter et al., 2020; Lau et al., 2009], the other using the subdistribution hazard [Fine and Gray, 1999; Lau et al., 2009].

Let $F_1(t)$ be the event-specific cumulative incidence for experiencing the outcome of interest (i.e. $E = 1$) before time $t$ and, accordingly, let $F_2(t)$ be the event-specific cumulative incidence of experiencing the competing event (i.e. $E = 2$) before time $t$. The event-specific hazard for event type $E = 1$ (in discrete time) is given by the conditional probability

$$h_1(t) = \mathbb{P}\left[T = t, E = 1 | T \geq t\right]. \tag{2.12}$$

The event-specific hazard for the competing event, $h_2(t)$, is defined accordingly as

$$h_2(t) = \mathbb{P}\left[T = t, E = 2 | T \geq t\right]. \tag{2.13}$$

There is no one-to-one relationship between a single event-specific hazard and cumulative incidence in the presence of competing events. The event-specific cumulative incidence for the outcome of interest is a function of both event-specific hazards $h_1(t)$ and $h_2(t)$ and is given by

$$F_1(t) = \sum_{s=1}^{t} h_1(s)S(s-1). \tag{2.14}$$

In Equation 2.14, $S(t) = \prod_{s=1}^{t}(1 - h(t))$ is the overall survival function, which depends on the overall hazard $h(t)$, which in turn is a function of all event-specific hazards and is defined as

$$h(t) = \sum_{e=1}^{\max(E)} h_e(t). \tag{2.15}$$

Accordingly, the event-specific hazards of all event types are needed to estimate the event-specific CIF [Schmid and Berger, 2021]. When treating competing events as censoring events instead (i.e. for the controlled direct effect), the resulting cumulative incidence function will always be larger than or equal to $F_1(t)$.

An alternative method of estimating the event-specific cumulative incidence function does not require information on the event-specific hazard of all event types: Individuals who experience the competing event are not censored, but remain in the risk set and are assigned a virtual end of their observation period, e.g. the end of the study period if only administrative censoring occurs [Putter et al., 2007]. The so-called subdistribution hazard [Fine and Gray, 1999; Lau et al., 2009; Schmid and Berger, 2021] for event type E = 1 is then given by the conditional probability

$$\lambda_1(t) = \mathbb{P}\left[T = t, E = 1 | (T \geq t \cap E = 1) \cup (T < t \cap E \neq 1)\right]. \tag{2.16}$$

The event-specific CIF for event type $E = 1$ as a function of the subdistribution hazard (see e.g. Putter et al. [2020]) is defined as

$$F_1(t) = 1 - \prod_{s=1}^{t}(\lambda_1(s)). \tag{2.17}$$

The advantage of the subdistribution approach is that only one hazard function needs to be estimated. In cases were computation takes a long time (e.g. big data), the lower computational burden is particularly appealing, even though it is best practice to also assess the effect of exposure on the competing event [Latouche et al., 2013].

## 2.7 Confounding bias

If $f(x)$ is an estimator for the estimand $\theta$, then it is considered biased if $|\mathbb{E}\left[f(x)\right] - \theta| > 0$. If an experiment were repeated many times, a biased estimator would return estimates that are systematically different from the estimand. Importantly, bias can arise from a multitude of issues such as the definition of the estimator, the data itself or the study design, among others.

The most frequently discussed source of bias in observational studies is confounding due to the lack of baseline randomization. When speaking of confounding, one must first define what is meant by the term ”confounder”. Often, a confounder is defined as a variable that is associated with both exposure and outcome. However, this simplistic definition is not sufficient. For example, if an exposure affects an outcome solely through a mediator, the mediator would be associated with both exposure and outcome, but we would not generally wish to control for the mediator as this would mask the effect. Throughout this dissertation, a confounder is defined as a variable that reduces confounding bias when adjusted for appropriately. More formally, a confounder is a variable on an open backdoor path from exposure to outcome which blocks this backdoor path when controlled for (see section 2.2 for definition of $d$-separation).

While elaborate methods exist to identify sufficient adjustment sets from the data (under various assumptions, see e.g. Witte and Didelez [2019]), the choice of confounding variables to be included in the adjustment set commonly relies on subject-matter knowledge. If the true DAG of the causal relationships between exposure, outcome and any other variables is known, a sufficient adjustment set can be read off from the DAG without the need for data-driven selection. While the assumption of no unmeasured confounding cannot be tested directly, sensitivity analyses are commonly used to collect evidence regarding the plausibility of this assumption, given the observed data.

### 2.7.1 Negative control outcome analysis

The underlying idea of negative control outcome analyses [Lipsitch et al., 2010] is to apply the data analysis framework of the study's main analysis, but to substitute the outcome variable with one that is known to be causally unaffected by the exposure of interest. After adjusting for observed confounders, the analysis should return a null-effect. If, in contrast, the analysis returns a non-null effect estimate for the negative control outcome, unobserved confounding may be at play.

In Figure 2.2, let $A$ be the exposure, $X$ the observed confounder, $U$ the unobserved

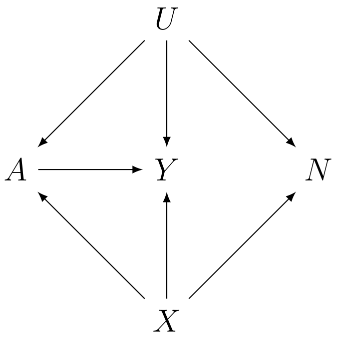


**Figure 2.2:** DAG of U-comparability of negative control outcome $N$

confounder, $Y$ the the outcome of interest and $N$ the negative control outcome, all binary. Naturally, some assumptions must be made. In particular, the assumption of "$U$-comparability" must be met: Lipsitch et al. [2010] define $U$-comparability of $N$ with $Y$ as the degree of overlap of the set of unobserved common causes of $A$ and $Y$ with the set of unobserved common causes of $A$ and $N$, with complete overlap indicating perfect $U$-comparability. Furthermore, one needs to assume that $A$ does not cause $U$ [Lipsitch et al., 2010].

When the effect of interest is the causal effect of $A$ on $Y$ e.g. expressed as

$$\theta = \mathbb{P}\left[Y^{A=1} = 1\right] - \mathbb{P}\left[Y^{A=0} = 1\right],$$

and the estimator $g(a, x)$ uses information on the observed covariates $x$, but not $u$, we would expect the estimator to be biased, i.e.

$$\theta \neq \mathbb{E}\left[g(a, x)\right].$$

This is obvious from the DAG in Figure 2.2, given the open backdoor path $A \leftarrow U \rightarrow Y$. More formally, $Y^{A=a} \not\perp\!\!\!\perp A|X$, because $Y^{A=a}$ and $A$ are only $d$-separated when also controlling for $U$. However, given that there is a causal effect from $A$ to $Y$, the observed (biased) effect estimate is a mixture of a true causal effect and confounding bias and the two cannot easily be disentangled.

In contrast, no causal effect of exposure on the negative control outcome exists. If variable $U$ were eliminated from the DAG in Figure 2.2 and if there was no model misspecification, we would expect the probability of observing $N = 1$ to be similar among exposed and unexposed individuals within strata of $X$, i.e.

$$\mathbb{P}\left[N = 1|A = 0, X\right] \approx \mathbb{P}\left[N = 1|A = 1, X\right].$$

Any deviation of the observed effect from the null, then, would be indicative of confounding bias due to $U$.

The above procedure of conducting negative control outcome analyses relies on the assumption of U-comparability being met. If, for instance, the directed edge from $U$ to $Y$ in Figure 2.2 were removed, the negative control analysis would indicate the presence of bias, but an estimator for the effect of $A$ on $Y$ conditional on $X$ would be unbiased, because no open backdoor path remains between $A$ and $Y$. Similarly, if the directed edge from $U$ to $N$ were removed, the negative control analysis would fail to identify the residual confounding bias present in the analysis of interest.

Negative control analyses are, therefore, only applicable if a suitable negative control outcome is available in the measured data. As discussed in Lipsitch et al. [2010], a perfect negative control outcome will rarely be available. However, a similar confounding structure and differences only in weak confounders may be sufficient in many cases.

Importantly, negative control analyses cannot generally be used to estimate the direction and magnitude of bias, unless one makes additional assumptions regarding the strength of association between variables. If, for instance, $U$ is a weak predictor of $N$, but a strong predictor of $Y$, the results of the negative control outcome analysis cannot be used to calibrate the effect estimate for the $X$-$Y$ relation [Lipsitch et al., 2010]. Only if the confounding structures relating to outcome of interest and negative control outcome are identical also with regard to direction and strength of association is it possible to quantify the magnitude of bias due to unobserved confounders $U$ and use it for calibration of the effect of interest.

In the context of the present thesis, negative control outcome analysis was used in the context of screening colonoscopy to investigate possible unobserved confounding. Here, the study outcome of incident CRC diagnosis was replaced by the negative control outcome of incident pancreas cancer diagnosis. Importantly, pancreatic cancer shares many risk factors with colorectal cancer, but the strength of the association cannot be assumed to be identical. For instance, stronger effects of tobacco smoke have been reported for pancreatic cancer [Maisonneuve and Lowenfels, 2015] than for colorectal cancer [Hannan et al., 2009]. For details on this application, see Section 4.4.1 and the corresponding paper in Section 7.2.

In many cases, confounding bias might only play a minor role while other biases, sometimes induced by the study design, are often not acknowledged appropriately (see, for instance, the example of menopausal hormone therapy and coronary heart disease discussed in Hernán et al. [2008]). These biases are referred to as ”self-inflicted”, because they arise purely from an inappropriate study design [Hernán et al., 2016]. Often, they arise due to non-alignment at time zero, which is discussed in depth in Chapter 3.

## 2.8 The built-in selection bias of the hazard ratio

The applied studies conducted as part of this thesis (Section 7) deviated from much of the published literature in that they did not report HRs as effect measures, but instead reported cumulative incidence curves showing the absolute cumulative risk in each group at any time during follow-up. Furthermore, a relative risk was estimated at the end of follow-up. The decision to estimate the risk over the entire follow-up rather than a single HR was intentional: The hazard at a given point in time is conditional on not having experienced the outcome previously. While groups may be comparable at baseline - either via randomization or adjustment - the survivors will systematically differ at a later time whenever the effect of exposure on outcome is non-zero [Hernán, 2010]. It has been argued that the hazard ratio from a Cox model cannot be interpreted as a causal effect measure, unless exposure has no effect or unless no factors other than exposure have any effect on the outcome [Hernán, 2010; Martinussen, 2022; Young et al., 2020].

## 2.9 Collider-stratification bias

While adjustment for confounding variables is usually required in observational studies, adjusting for the wrong variables can also introduce bias. One example of this is over-adjustment, where the adjustment set includes a variable on the causal path from exposure to outcome and, thereby, masks the effect of interest [Schisterman et al., 2009]. Another example would be that of collider stratification bias [Greenland, 2003; Hernán and Monge, 2023].

As discussed in section 2.2, a backdoor path from exposure to outcome is blocked by the empty set, if it contains a collider. This means for the analysis that neither the collider nor any of its descendants are to be adjusted for in the analysis. Conversely, if the analysis is adjusted for the influence of a collider or its descendants, the backdoor path is open, leading to bias. While collider-stratification bias may arise due to conditioning on a collider variable in the analytical model, it can also arise as a consequence of selection. This becomes particularly important in the context of non-alignment at time zero, which is discussed in depth in Chapter 3.

# CHAPTER 3

# Target trial emulation

## 3.1 Motivation

While RCTs are often considered a gold standard in medical research, they are often not feasible or even appropriate to answer specific research questions. Especially in cases where a high generalizability is paramount, or e.g. for assessing off-label use of medications, RCTs are not suitable. However, they possess some properties that are particularly advantageous to investigate cause-effect relations.

The most obvious of these properties is that there is on average no imbalance of baseline characteristics due to the randomization process. Randomization might not achieve perfect balance for every covariate in a specific trial, but conceptually covariates will approximate balance as sample size approaches infinity. Importantly, this applies to both measured and unmeasured baseline characteristics.

An often neglected property of RCTs that aids an unbiased assessment of causal effects is the temporal ordering of central design elements, which need to be aligned at time zero [Fu, 2023; Braitmaier and Didelez, 2022]. In an RCT potential study participants are first screened regarding their eligibility. Next, the eligible ones are randomly assigned to the treatment arms after signing an informed consent form and are invited to the baseline examination and first treatment. The follow-up, then, only starts after this initial visit and follow-up variables and the outcome of interest are measured at the subsequent follow-up visits or at anytime during follow-up when allowing for electronic patient reported outcomes. This temporal alignment is illustrated below in Figure 3.1 and is referred to as alignment at time zero in this thesis. Whenever there is misalignment of these three design elements, substantial bias may be the consequence [García-Albéniz et al., 2017b; Hernán et al., 2016]. As discussed in Section 2.9, this issue can be expressed as a special form of collider stratification bias.

A particular focus of the present thesis was bias due to non-alignment at time zero in observational studies on screening colonoscopy. In that context and considering prospective study designs, one mechanism merits special emphasis: Exposure definition based on pre-baseline information leading to a prevalent user type bias. Considering retrospective case-control designs, bias may arise due to a post-baseline exposure definition. Each will be introduced briefly here and a detailed assessment in the context of screening colonoscopy is given in Chapter 4 and Section 7.6.

Bias due to exposure assessment using information from **before** time zero can have various manifestations depending on the studied indication. In pharmacoepidemiological research, it is to blame for the so-called ”prevalent user bias” or bias due to ”depletion of susceptibles”: Considering a study on suspected adverse events of a medication, a comparison of current or prevalent users with never-users would be problematic. Any potential study participants who were treated with the drug in the past, but stopped taking the drug due to the adverse event, would not be included in either the current user group or the never user group, leading to a depletion of individuals susceptible to the adverse event among the previously exposed. Those who (still) take the drug at baseline are then more likely to respond well to the drug and be resistant against the adverse event. Due to this mechanism, menopausal hormone therapy was linked to a decreased risk of coronary heart disease in an observational study [Grodstein et al., 2006], even though an RCT indicated an increased risk [Manson et al., 2003]. A later study by Hernán et al. [2008] found no decreased risk for coronary heart disease when alignment at time zero was ensured by the study design. A similar bias arises in the context of screening colonoscopy, when individuals with a history of CRC are excluded while previous exposure to screening colonoscopy is used to define the comparison groups. A structural exploration of this scenario, together with a proper study design using TTE is given in Chapter 4 and Section 4.6, where it is shown that the resulting bias is a form of collider stratification bias. A special focus is set to site-specific effectiveness of screening colonoscopy.

Conversely, defining exposure based on information from **after** time zero leads to immortal time bias [Suissa, 2008; Hernán et al., 2016], which can also be understood as a form of collider stratification bias [Shrier and Suissa, 2022]. The mechanism at work is as follows: If exposure uses post-baseline information, e.g. compares ever users to never users, individuals exposed late during follow-up cannot, by definition, have died previously - hence the term ”immortal time”. Early deaths (or any outcomes), therefore, accumulate in the unexposed group, creating a false impression of the exposure being overly protective. Many examples have been published over the years: Suissa and Azoulay [2012] explore this bias in the context of metformin therapy and cancer risk, whereas García-Albéniz et al. [2017b] show the potential for immortal time bias when evaluating screening colonoscopy. It is noteworthy that these time related biases are not unique to prospective study designs, but also permeate case-control studies [Dickerman et al., 2019; Rasouli et al., 2023]. Further sources of bias, such as inappropriate adjustment for confounding, also affect case-control designs [Rasouli et al., 2023].

Bias due to non-alignment at time zero was discussed in depth for the use case of site-specific effectiveness of screening colonoscopy in Sections 4.6 and 7.6. As discussed in Section 4.6, TTE is a simple solution to avoid this type of bias.

Finally, RCT results naturally lend themselves to an interventionist interpretation of a causal effect, since they study the effect of a well-defined intervention on a subsequent outcome. If the randomization process is successful in eliminating baseline confounding and if there is no differential loss to follow-up, the only aspect in which the study arms differ is the treatment. This means that any difference between the treatment arms regarding the outcome is attributable to the intervention, which makes it easy to translate the effect into a recommendation for policy makers or regulators. If a beneficial effect is observed, it would be unethical to withhold treatment from the public and if, conversely, no beneficial effect or even a harmful effect is observed it would be unethical to offer the intervention to patients. In observational studies, on the other hand, exposures of interest are sometimes ill-defined. While it is, for instance, possible to estimate an ”effect” of BMI on health outcomes using association measures, no recommendation for policy makers would be possible based on a study that does not define how a change in BMI should be achieved. Causality is therefore usually defined as a contrast of potential outcomes under different interventions [Rubin, 2005].

## 3.2 Basic procedure

With the above points in mind, it seems plausible to apply some of the design elements of RCTs to observational studies, while avoiding some of the weaknesses of RCTs, such as low generalizability due to a highly restricted study population. This is achieved in the so-called target trial emulation framework, which has gained popularity especially for studies using RWD [Hansford et al., 2023b]. As noted by Labrecque and Swanson [2017], TTE is particularly suitable for teaching causal inference concepts, because it applies existing and well-known study design aspects in a new way rather than requiring researchers to understand a completely new method. The basic principles of target trial emulation, while not always referred to as such, have been established several decades ago. Dorn [1953] lists several questions that a researcher planning an investigation of causal effects using observational data should answer as to minimize the risk of mistaking association for a causal effect. One of these questions is: "How would the study be conducted if it were possible to do it by controlled experimentation?" Later, Robins [1986] defined a method to assess exposures sustained over time, in which he defines an observational cohort so as to mimic the data one would obtain from an RCT, if information on treatment assignment was missing. Even though the ideas behind target trial emulation have been published many decades ago, popularity of the methods only started to increase relatively recently, due to pioneering work such as the above-mentioned study by Hernán et al. [2008]. Since then, many target trial emulations on different research questions have been published (see for instance Hernán and Robins [2016]; Caniglia et al. [2019]; Danaei et al. [2013]; García-Albéniz et al. [2017a]; Petito et al. [2020]; Chiu et al. [2024]) and several studies demonstrating certain perils of observational studies that can be circumnavigated by target trial emulation are available (Hernán et al. [2016]; García-Albéniz et al. [2017b]; Dickerman et al. [2019]; Didelez [2016]; Emilsson et al. [2018]). Much work has been done on replication of existing RCTs as to identify scenarios in which TTE is either particularly suitable or faces substantial challenges [Franklin et al., 2021; Heyard et al., 2024; Hoffman et al., 2022; Wang et al., 2023, 2024]. Furthermore, guidance on reporting of emulated target trials is now available (see Hansford et al. [2023a,b]).

Target trial emulation is a two-step process. First, the study protocol of the ideal hypothetical trial, i.e. the target trial, is defined. Second, an emulation of this ideal trial using observational data is defined, so that the observational study is as similar to the target trial as possible. The goal of this two step process is to ensure that central elements of the study are clearly defined. For instance, which population should be studied, which (hypothetical) interventions should be compared using what contrast, or how the outcome of interest is defined, also considering intercurrent events [Hernán and Robins, 2020]. However, the study protocol of the target trial is usually not defined in as much detail as would be required in a real RCT seeking ethical approval, but instead is sketched out in tabular form [Braitmaier and Didelez, 2022; Hernán and Robins, 2016]. An example table, adapted from Hernán and Robins [2016] is given in Table 3.1.

Any observational study – even when emulating a randomized trial – will differ in some aspects from an RCT. Figure 3.1 illustrates the alignment of study design elements at time zero of a hypothetical RCT on the left side and the design of a corresponding observational, emulated trial on the right side. In an RCT, the time window for eligibility assessment and assessment of baseline variables ends before baseline, i.e. in the randomized trial before

**Table 3.1:** Tabular overview of the study protocol elements of emulated target trial studies, adapted from Hernán and Robins [2016].

| **Component** | **Description** |
|---|---|
| Study aim | Definition of the research question |
| Eligibility | Eligibility criteria might differ between RCT and observational study. For the observational study one might require individuals to be observable in the data source for a minimum lookback period to ensure that other eligibility criteria can be assessed accurately. Conversely, some criteria required in an RCT may be omitted in the emulation. For instance, pregnant women are routinely excluded from pre-marketing RCTs due to ethical concerns. Off-label use among pregnant women might, then, be assessed in observational studies. It must be kept in mind that any modification of eligibility criteria may affect transportability of study results. |
| Treatment strategies | Treatment strategies must be defined clearly. It is not sufficient to specify e.g. which drug should be investigated, but also over which time period treatment must be sustained and which deviations from prescribed treatment should be allowed or not allowed per protocol. If a treatment should be changed dynamically based on e.g. blood testing this needs to be pre-specified. |
| Treatment assignment | The treatment assignment in an RCT would be done randomly. In observational studies, treatment assignment corresponds to the observed treatment behavior. Randomization is emulated in observational studies by adjusting for a sufficient set of covariates. Adjustment covariates should ideally be selected based on subject matter knowledge and using causal reasoning. Adjustment variables should be listed in the protocol and a method of confounder adjustment be specified. |
| Follow-up | Clear definition of when follow-up starts and ends. |
| Outcome | The outcome variable must be clearly defined. One should consider whether the outcome was reliably measured in the data, or if there might be issues with measurement error and misclassification. Strategies for intercurrent events must be defined. |
| Causal contrast | A clear definition of the causal contrast of interest is required. It should be clearly stated how non-adherence is handled when the treatment of interest is dynamic or sustained over time (e.g. ”intention-to-treat” vs ”per-protocol”) as this depends on the target of inference and affects the statistical methods used. |
| Statistical analysis | When the causal contrast is a per-protocol effect, artificial censoring and adjustment for time-dependent confounding are required. |

visit 0 at which informed consent is obtained and randomization is conducted. The first treatment might not occur immediately at visit 0, but instead at visit 1 shortly thereafter. Follow-up starts after visit 0, i.e. after baseline. It is common for RCTs to estimate an intention-to-treat (ITT) rather than a PP effect, i.e. the target of inference is the effect of being assigned to a treatment arm rather than receiving treatment. As illustrated in the right-hand side of Figure 3.1, an emulated trial similarly obtains information regarding eligibility and baseline variables from before time-zero. However, since no randomization is conducted, group assignment is then based on the observed exposure during the time-zero time interval. Emulated target trials typically treat time as a discrete entity, so that this first time interval may correspond to e.g. a week, month or quarter. Since group assignment is based on observed exposure, no ITT effect regarding treatment assignment can be estimated in an emulated trial. Instead, studies often estimate an ”observational analog” of the ITT effect, namely the effect of being exposed at time-zero. This, however, corresponds more closely to the PP effect reported in many RCTs, where adjustment for non-adherence at baseline is done.

Note that in Figure 3.1 the term ”baseline” is used for the RCT, while the term ”time zero” is used for the emulated trial. In an RCT, one clear baseline is defined, namely the day at which a study participant signs their informed consent form and is randomized into one study arm. Often, no single baseline exists per person in an emulated trial. If, for instance, the emulated trial entails a control group not receiving treatment, it is unclear from the observed data when follow-up should start for this person. Some studies in the past have then declared one fixed baseline and assessed exposure either before (leading to prevalent user-type biases) or after (leading to immortal time bias) baseline. The solution in the TTE framework is usually to emulate multiple sequential trials, one at the beginning of each discrete time interval. Each of these emulated trials has its own respective baseline and all individuals eligible at that baseline are included in the respective trial. As a result, the same person is included in multiple trials with differing baselines. However, time-zero alignment of eligibility assessment, treatment assignment and start of follow-up is ensured in all these emulated trials.

Similarly to the above point, emulated trials may, in contrast to randomized trials, assign the same person to more than one treatment strategy. If, for instance, one were to compare one dynamic treatment strategy that adapts medicine dose to some biomarker value with one static treatment strategy that does not adapt medicine dose, all initiators would qualify for both these strategies. In an RCT such an individual would be randomly assigned to one strategy. In an emulated trial this person could also be randomly assigned to one strategy. However, it is more efficient to clone the data from this person, assign one clone to each

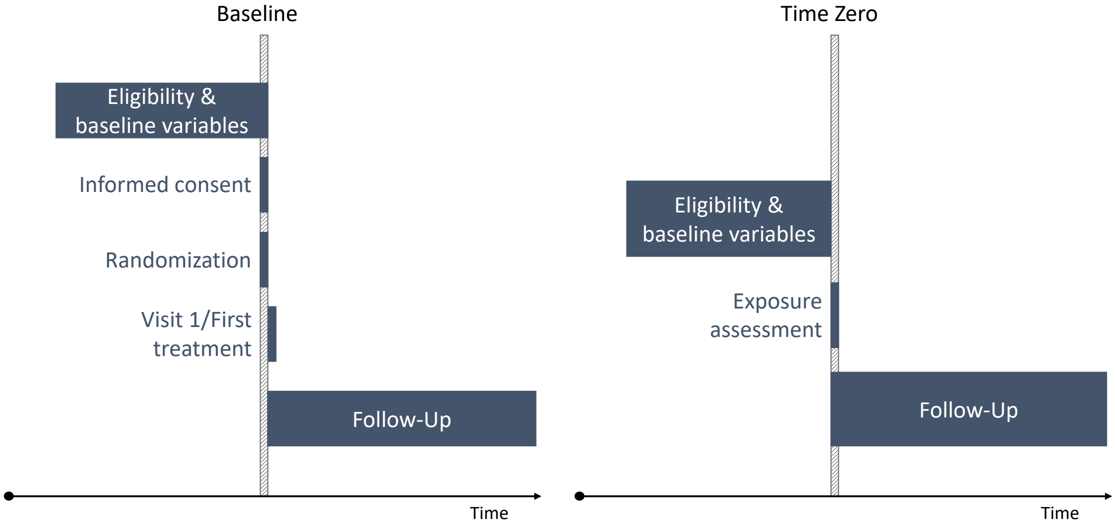


**Figure 3.1:** Left side: Temporal ordering of design elements of an example RCT aiming to estimate an intention-to-treat effect. Right side: Temporal ordering of an emulated trial aiming at estimating the observational analog of an intention-to-treat effect.

strategy and censor the clone from a strategy when the observed treatment exposure during follow-up deviates from the assigned strategy. As this artificial censoring introduces bias, adjustment for time-dependent confounding, e.g. via IPCW becomes necessary [Robins, 1986; Hernán, 2018]. This approach is sometimes called the "clone-censor-weight approach" [Zhao et al., 2021].

Importantly, duplicated data due to cloning or sequential trial emulation needs to be considered when estimating confidence intervals. In this context, bootstrapping is commonly used. A brief introduction to bootstrapping is given in Appendix A, while bootstrapping in the emulated target trial on screening colonoscopy is described in Chapter 4.

## 3.3 Data sources

The thought experiment of formulating the ideal trial to answer a given research question can be instructive in any observational study on causal effects [Didelez, 2016]. However, when emulating said ideal trial using observational data, certain criteria need to be fulfilled.

The observational data must contain sufficient information to fulfill the following: As mentioned in Franklin et al. [2019], the information must allow the identification of the target population via eligibility criteria, contain sufficient information to adjust for confounding bias, contain reliable information on exposure and outcomes and must contain information as to judge the generalizability of the study results. Importantly, the temporal ordering of events must be clear from the data. These requirements are not unique to TTE.

Next, data must be collected longitudinally, ideally without gaps. The TTE framework relies on sensible temporal ordering of eligibility assessment, treatment assignment and start of follow-up. Therefore, the data must contain temporal information on a granularity that allows this alignment. If continuous information is available over a long time period, sequential trials can be emulated to make the best use of the available information.

When aiming to answer medical research questions, the real-world data most commonly used are pseudonomized health claims data, electronic health records (EHR) or disease registries [Franklin et al., 2019; Haug and Schink, 2021]. Even though claims data and EHR are not collected for research purposes, they contain rich medical data, are readily available and possess large sample sizes, making them interesting for research questions that cannot easily be answered in RCTs.

Observational cohort data collected at subsequent visits is less ideal for TTE than routinely collected health data, given the large gaps in between visits and the often limited number of visits. TTE from such data must make stronger assumptions (see e.g. Chiu et al. [2021] for an application using cohort data).

## 3.4 Strengths and limitations

RCT evidence is usually regarded as the most reliable basis for decision making regarding medical interventions. However, RCTs have limitations of their own, cannot answer all relevant research questions and often need to be supplemented or even replaced by observational studies using RWD in specific settings. In observational studies, however, there is large variation regarding quality of both study design and underlying data. As part of the present thesis, Braitmaier and Didelez [2022] established a German language tabular overview of the limitations of RCTs and observational studies with and without TTE, which was adapted to English in Table 3.2.

**Table 3.2:** Limitations of various study designs, adapted from Braitmaier and Didelez [2022]

| ***Risk*** | ***RCT*** | ***Observational studies with RWD*** | |
|---|---|---|---|
| | | ***With TTE*** | ***Without TTE*** |
| Prevalent user bias [Ray, 2003] | Low; randomization marks treatment start | Low; avoided by alignment at time zero | High, if exposure assessment uses pre-baseline information |
| Immortal time bias [Suissa, 2008] | Low; assignment to arms with randomization | Low; avoided by alignment at time zero | High, if exposure assessment uses post-baseline information |
| Unclear research question [Didelez, 2016] | Low in both, due to definition of (hypothetical) intervention | | High, if exposure does not correspond to an entity that can be intervened on |
| Baseline confounding | Low, avoided by baseline randomization | High in both; can be corrected for by appropriate adjustment (if data is sufficiently informative) | |
| Time-dependent confounding [Hernán and Hernández-Díaz, 2012; Howe et al., 2016] | High in all, if (differential) loss to follow-up, treatment switching (non-adherence), artificial censoring, among others; can be corrected for (e.g. using inverse probability of censoring weights), if data is sufficiently informative | | |
| Low external validity | High; trial population is highly selected and possibly differs substantially from target population | Depends on data source; registry or claims data contain information on subgroups routinely excluded from RCTs and are more informative for real-life use/exposure. However, volunteer bias (among others) can occur in other data sources | |
| High economic and time costs | High | Low in both, if existing (secondary) data can be used | |

# CHAPTER 4

# Contributions to the field of screening colonoscopy

The development of bespoke study designs and statistical analysis methods for specific applications is a focal area of this dissertation. In this context, the effectiveness of two cancer screening programs - mammography screening for early detection of breast cancer and colonoscopy screening for early detection and prevention of colorectal cancer - was assessed. Contributions to the field of screening colonoscopy are described in the current chapter 4, while contributions to the field of screening mammography are described in chapter 5.

Colonoscopy screening for prevention and early detection of colorectal cancer is offered in Germany since 2002. While the reduction of CRC-related mortality is the ultimate goal of colonoscopy screening, CRC incidence is another important outcome. A cancer diagnosis and subsequent diagnostic procedures and curative treatments have a significant impact on a patient's live. Screening colonoscopy is thought to affect CRC incidence in two major ways: 1) Early detection of asymptomatic cases leads to an increased incidence early after screening, but also to earlier treatment initiation, which in turn improves survival 2) Detection and removal of precursor stages during the screening examination reduces CRC incidence [Bretthauer et al., 2022]. This assumed mechanism is illustrated in Figure 4.1, where presence of cancer precursors $P$, undiagnosed cancer $C$, colorectal cancer diagnosis $Y$ and exposure to screening colonoscopy $A$ is captured at two time points. While exposure at time point 1, $A_1$, leads to an increase of cancer incidence at the same time point ($Y_1$) by detecting prevalent cases ($C_1$), it also leads to a decrease of later cancer incidence $Y_2$ by removing precursor stages that are present at the time of screening ($P_1$).

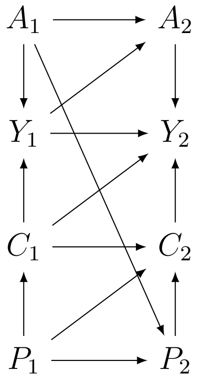


**Figure 4.1:** DAG representing the causal structure of screening colonoscopy and colorectal cancer incidence

Importantly, no RCT evidence regarding colonoscopy screening's effectiveness in reducing colorectal cancer incidence or mortality was available at the time of its introduction. However, its effectiveness was implied by RCT evidence on sigmoidoscopy, a less invasive, endoscopic screening tool that functions in a similar way, but screens only the distal as opposed to the entire colorectum [Elmunzer et al., 2012]. Furthermore, observational studies conducted after the introduction of screening colonoscopy in Germany suggested a strong effect on both CRC incidence and mortality, although some of these studies found considerably stronger effects for the distal colorectum [Baxter et al., 2009, 2012; Brenner et al., 2011, 2014a; Doubeni et al., 2013; Guo et al., 2021; Kahi et al., 2018; Mulder et al., 2010; Nishihara et al., 2013]. The first RCT results were published in 2022 [Bretthauer et al., 2022], i.e. after the study by Braitmaier et al. [2022b]. However, Bretthauer et al. [2022] focused on overall incidence of CRC, as they did not have sufficient sample size to obtain site-specific estimates.

García-Albéniz et al. [2017a] was the first to study the effectiveness of screening colonoscopy using a target trial emulation approach. In a companion paper, they discuss how commonly-applied approaches lead to non-alignment at time zero and consequently to self-inflicted biases [García-Albéniz et al., 2017b]. They did not, however, study whether these self-inflicted biases differ between CRC sites and if they could potentially explain the difference in effect estimates reported by previous observational studies. It was, therefore, the objective of our study [Braitmaier et al., 2022b] to extend the framework of García-Albéniz et al. [2017a] and use a target trial emulation design to study site-specific effectiveness of colonoscopy screening in reducing CRC incidence. The details of the study design including a tabular study protocol of the target trial and its emulation using observational data are given in the paper (see section 7.2). Briefly, we emulated sequential trials – one per calendar quarter – from 2007 to 2011. Calendar quarters were chosen as the unit of discrete time, because some information in the underlying GePaRD is only available on a quarterly basis. The strategies to be compared were to either undergo colonoscopy screening in the baseline quarter or not. Individuals who underwent screening colonoscopy in the baseline quarter were assigned to the screening strategy, while individuals not undergoing screening colonoscopy in the baseline quarter were assigned to the control strategy. This assignment process results in a trial population with zero non-adherence at baseline.

While the goal of the initial analysis described in Braitmaier et al. [2022b] was to estimate the effect of baseline exposure to screening colonoscopy regarding the site-specific effectiveness, several extensions and additional data years were added to the project later and are described in separate sections. First, section 4.1 introduces the methodological framework for the target trial emulation on screening colonoscopy. Section 4.2 describes the process used to find a suitable parameterization of the pooled logistic model used to estimate the discrete-time hazards. Next, Section 4.3 gives a brief summary of the main results from Braitmaier et al. [2022b], while an extensive description is given in the paper itself, which is included in Section 7.2. Sensitivity analyses are described in Section 4.4 and extensions to the original study design are described in the subsequent sections.

The only RCT evidence regarding screening colonoscopy's effect on CRC incidence [Bretthauer et al., 2022], which was published after Braitmaier et al. [2022b], assessed an intention-to-screen effect, i.e. the effect of being invited to screening at baseline. They furthermore conducted a per-protocol analysis by adjusting for non-adherence during baseline. While this is more comparable to the effect estimate reported in Braitmaier et al. [2022b], there is a key difference: The RCT by Bretthauer et al. [2022] was conducted during a time in which screening colonoscopy was not available to the broader public, i.e. there was no contamination of the control arm during follow-up. To make our results more comparable with theirs and to fully evaluate the effect of screening colonoscopy in Germany, we added a per-protocol analysis, censoring in the control arm at the earliest screening colonoscopy during follow-up. This is described in Section 4.5.

Section 4.6 gives a structural explanation of bias arising in the context of screening colonoscopy due to violations of alignment at time zero, with empirical results again reserved for the corresponding publication in Section 7.6.

Further extensions included the assessment of the effect of quality of screening colonoscopy. The quality was defined based on polyp-detection rate and the methodological framework was extended for three instead of two exposure strategies - no screening colonoscopy at baseline, low-quality screening colonoscopy at baseline and high-quality screening colonoscopy at baseline. Further extensions and additional analyses are described in detail in Section 4.7 below.

## 4.1 Methodological framework

### 4.1.1 Target trial emulation

Data was collected from an underlying cohort of $n$ individuals. Each individual, $i = 1, ..., n$, was characterized at time $t = 1, ..., T$ by covariates $X_t$, a binary exposure status $A_t$ and an outcome indicator $Y_t$. Individuals of this cohort were assumed to be independent. Let overbars indicate the history of a variable. An emulated trial might investigate what some authors have called the observational analog of the ITT effect (i.e. without censoring for non-adherence during follow-up) of exposure strategies $Q$ on the outcome $Y$, where $Q = 0$ is the strategy of not being exposed to colonoscopy screening at baseline and $Q = 1$ is the strategy of being exposed at baseline, both without restrictions regarding further screening during follow-up. A sequence of $r$ trials was emulated, in this case by starting one emulated trial at each calendar quarter from 2007 to 2011. The $r$-th emulated trial started at (calendar) time $t_r$ and follow-up time of the $r$-th trial is denoted by $k_r = 1, ..., K_r$, with $K_r = T - t_r + 1$. Selection of individuals into emulated trials was based on eligibility at time $t_r$, $E_{t_r} = 1$. Assignment to exposure strategies was based on observed exposure at time $t_r$, i.e. for the $r$-th emulated trial, person-trial $j$ has the assigned exposure strategy $Q_j = A_{j,t_r}$, where the subscript $j = 1, ..., m$ with $m \geq n$ refers to ”non-unique” person-trials. The same individual $i$ may be eligible for multiple trials.

### 4.1.2 Effect estimation

A pooled dataset of all emulated trials contained information on all $j = 1, ..., m$ person-trials regarding person-trial specific information on e.g. baseline covariates $X_{j,t_r}$ and information on the outcome of interest starting with baseline and continuing through follow-up as $Y_{j,k_r}$. While the main focus was on the event-specific cumulative incidence over time under each strategy, the summary measure of interest was the causal relative risk (CRR) at the end of follow-up given by

$$\mathrm{CRR}_T = \frac{\mathbb{P}\left[Y_T^{Q=1} = 1\right]}{\mathbb{P}\left[Y_T^{Q=0} = 1\right]}. \tag{4.1}$$

Estimand 4.1 does not refer to a hypothetical scenario controlling the occurrence of competing events $D$, i.e. the aim of this analysis was to estimate the total effect without elimination of competing events as defined in Young et al. [2020].

Estimation of 4.1 is based on a pooled logistic model applied to the pooled dataset to estimate the discrete-time subdistribution hazard for time point $k_r$ in a first step. Subdistribution hazards, rather than event-specific hazards, were used here, because this way only the model for the event of interest needed to be fitted. When using event-specific hazards instead, models for both event of interest and competing event need to be fitted to derive the event-specific cumulative incidence functions, i.e. an approach using event-specific hazards would have been computationally more costly. Instead, individuals experiencing the competing event are not treated as censored, but instead remain in the risk set until the time when they would have censored, had they not experienced the competing event (i.e. until the end of the study period). In the absence of censoring, these individuals would remain in the risk set indefinitely. With this, the at-risk set at time point $k_r$ comprises individuals who have experienced the competing event in addition to those who have not yet experienced any outcome event [Putter et al., 2007]. The pooled logistic model then takes the form

$$\mathbb{P}\left[Y_{j,k_r} = 0 | \bar{Y}_{j\,k_r - 1} = 0, Q_j\right] = \text{logit}^{-1}\left[\eta(q_j, k_r)\right]. \tag{4.2}$$

A suitable parameterization of $\eta(q_j, k_r)$ depends on the functional shape of the cumulative incidence functions and will differ from study to study. The process of finding a suitable parameterization in the example of the emulated target trial on screening colonoscopy is described in Section 4.2 below.

Inverse probability of treatment weighting (IPTW) was used in the above model to adjust for baseline confounding. For this, a set of baseline covariates $X_{j,t_r}$ was selected based on subject-matter knowledge (see Braitmaier et al. [2022b] for details). These covariates were included in a main effects logistic model estimating the probability of being exposed to screening colonoscopy in the time-zero discrete time-interval as $\mathbb{P}[Q_j = 1 | X_{j,t_r}] = \text{logit}^{-1}\left[x_{j,t_r}\beta\right]$. The predicted probability extracted from this fitted model ($\widehat{\text{PS}}$) was used in the denominator of stabilized inverse probability weights as

$$sw_j = \frac{\hat{p}\left[Q_j = 1\right]}{\widehat{\text{PS}}} \tag{4.3}$$

for exposed person-trials and as

$$sw_j = \frac{1 - \hat{p}\left[Q_j = 1\right]}{1 - \widehat{\text{PS}}} \tag{4.4}$$

for unexposed person-trials, with $\hat{p}[Q_j = 1]$ being the predicted probability of being exposed, extracted from a logistic model fitted without predictor variables. These weights were truncated by setting weights above the 99th percentile to the 99th percentile of the observed weight distribution, as is common practice [García-Albéniz et al., 2017a; Goetghebeur et al., 2020]. A sensitivity analysis was later conducted to assess the impact of truncation (see Section 4.4).

Once Model 4.2 was fitted on the data, using the above inverse weights, the predicted probability of not experiencing the outcome event by time $k_r$ under screening strategy $Q = q$, denoted here as $\hat{p}\left[Y^q_{k_r} = 0|\bar{Y^q}_{k_r-1} = 0, Q\right]$ , was extracted. This was achieved by generating a dataset with one entry per time point and screening strategy and using it as input to extract predicted probabilities from the fitted model. No person-trial level information or inverse weights are needed for this step. With these predicted probabilities, an estimate for the marginal, event-specific cumulative incidence was derived as

$$\hat{p}\left[Y^q_{k_r} = 1\right] = 1 - \prod_{l=1}^{k_r} \hat{p}\left[Y^q_l = 0|\bar{Y}^q_{l-1} = 0, Q\right]. \tag{4.5}$$

Finally, the summary measure of the effect of interest, in this case the CRR at time $k_r$, was derived as

$$\widehat{\text{RR}}_{k_r} = \frac{\hat{p}\left[Y^{q=1}_{k_r} = 1\right]}{\hat{p}\left[Y^{q=0}_{k_r} = 1\right]}. \tag{4.6}$$

While (4.6) was estimated as a summary measure of the relative effect at the end of follow-up, the cumulative incidence curves given by (4.5) should be considered the main output. The CIFs allow a visual assessment of temporal effects and are useful for risk prediction under hypothetical intervention by estimating them for specific subgroups of interest. In general, reporting CIFs is preferable to summarizing the temporal effect in a single measure, such as an average RR or HR.

In the above analyses, death is a competing event for CRC incidence, since CRC incidence after death is not defined. At the same time, CRC incidence is per design a competing event for death, because follow-up is terminated at any incident CRC diagnosis; any events after the event of interest are not pertinent to the research question at hand. Furthermore,

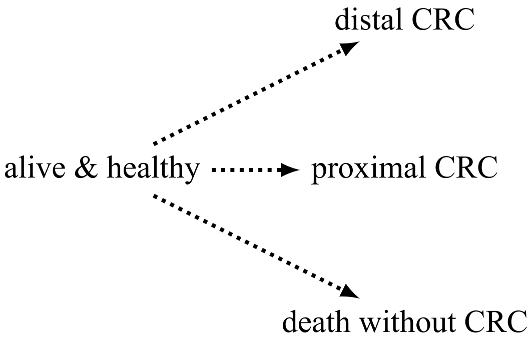


**Figure 4.2:** Multi-state model representation of site-specific CRC incidence and competing death

CRC can occur at different sites in the colon (distal versus proximal). Disregarding, for simplicity's sake, the rare cases in which the location of the tumor is unknown or tumors occur simultaneously at both sites, distal CRC is then a competing event for proximal CRC and vice versa, if interest lies in the first CRC diagnosis overall. In the analyses described in Braitmaier et al. [2022b], separate models were fitted for distal and proximal CRC, respectively, using the above methods, i.e. treating CRC at the respectively other site as competing events and not treating competing events as censoring events. Instead, event-specific CIFs were estimated via the intermediate step of estimating discrete-time subdistribution hazards. When using Cox PH models, this approach is often referred to as Fine-Gray approach [Putter et al., 2020]. Notably, the event-specific CIFs could have also been obtained using event-specific hazards, as is common in a multi-state representation of competing events [Putter et al., 2007]: The initial state is enrollment into the study population by meeting all eligibility criteria. Given that the outcome of interest is CRC incidence, follow-up is terminated at the time of diagnosis (because anything after that point is not of interest to this particular research question). With that and as illustrated in Figure 4.2, three absorbing states exist: CRC in the distal colon, CRC in the proximal colon and death. In the multi-state representation of the competing events model, the transition intensities from the initial state to each absorbing state are given by the respective event-specific hazard. The event-specific hazards then need to be estimated for all competing events, even if interest is only in one event type, because the event-specific cumulative incidence function depends on all of them. Therefore, the subdistribution approach is computationally faster.

### 4.1.3 Confidence intervals

No simple, parametric solution is available for obtaining confidence intervals when using pooled logistic regression to approximate discrete-time hazards, with repeated recruitment of the same individuals in more than one sequential trial. Bootstrapping was therefore used to obtain robust confidence intervals for CIFs and RRs.

One bootstrap sample of size $n$ is obtained by randomly sampling with replacement from the original cohort data. Each bootstrap ”individual” $i^*$ is characterized at time $t = 1, ..., T$ by covariates $X_t^*$, a binary exposure status $A_t^*$ and an outcome indicator $Y_t^*$. Note that an individual $i$ can be represented by more than one bootstrap individual $i^*$. Based on this bootstrap sample, the target trial emulation and accompanying analytical process described in Section 4.1.2 was repeated to obtain the first bootstrap estimate, e.g. $\widetilde{\mathrm{RR}}_{1,k_r}$ for the relative risk at time $k_r$. This process was repeated a total of $B = 250$ times to obtain $250$ bootstrap estimates, e.g. $\widetilde{\mathrm{RR}}_{1,k_r}, ..., \widetilde{\mathrm{RR}}_{B,k_r}$ for the time-dependent relative risk. While $B = 500$ bootstrap samples are often used in the literature, only $250$ samples were used here, due to computational limitations and the number of analyses conducted. However, $500$ samples were used in a sensitivity analysis to assess whether results would have been much different. This analysis is described in detail in Section 4.4.7.

While various bootstrap methods are available, the most common method of calculating confidence intervals in the target trial emulation literature is that of percentile based bootstrap intervals. For 95 % confidence intervals, these are defined as the 2.5 % and 97.5 % percentiles of the distribution of bootstrap estimates.

## 4.2 Functional shape of time

In the above analysis, pooled logistic regressions were fitted to obtain an estimate of the cumulative incidence functions. While a central aspect of this model is the covariate adjustment via inverse probability of treatment weighting, a necessary first step is the definition of an appropriate parameterization of the pooled logistic model itself. This model takes as input a modified dataset with one entry per discrete time point at which a person-trial was under observation. Time itself is then included in the model equation. An unadjusted, non-parametric method (e.g. Kaplan-Meier in absence of competing events or Aalen-Johansen when competing events exist) is used first to assess the shape of the CIFs. The pooled logistic model then features discrete time, transformations of time and possibly interaction terms with the treatment indicator. Depending on the nature of the exposure and outcome, further variables might need to be included e.g. if repeated exposure takes place during follow-up (see García-Albéniz et al. [2020] as an example). In the emulated target trial described in Braitmaier et al. [2022b] and its extensions, the parameterization was found via the following procedure: First, a non-parametric estimate of the cumulative incidence functions was obtained via Kaplan-Meier methods, without any covariate adjustment. Competing events were treated as censoring events in this model selection step, after it was shown that the difference between direct and total effect was minor, as discussed in Section 4.4.2. Second, candidate model specifications were selected as described below. Third, the resulting parametric estimates of the CIFs were compared to the Kaplan-Meier curves to assess fit, both visually and using a numeric measure, namely the Kolmogorow-Smirnov statistic. This measure – for exposure $q$ in this case – is defined as

$$KS^q = \sup_{k_r} |\hat{F}^q_{[\text{Kaplan-Meier}]} - \hat{F}^q_{[\text{Pooled logistic}]}|, \tag{4.7}$$

with sup being the supremum and $\hat{F}^q$ the observed cumulative incidence function for strategy $q$. The closer the Kolmogorow-Smirnov statistic is to zero, the less deviation exists between the cumulative incidence functions being compared. To assess the fit of the candidate functional shape, the maximum $KS$ statistic observed over both exposures $q$ was determined as

$$KS \text{ measure} = \max_q KS^q. \tag{4.8}$$

Candidate functional shapes were identified by the following steps:

1. The following transformations of time $k_r$ were used: $k_r$, $\sqrt{k_r}$, $(k_r)^2$, $\log k_r$, $\exp k_r$ and $\frac{1}{\exp k_r}$ (selection of subsets of these is defined in step 4 below).
2. Each candidate model was required to include a linear time term, because the Kaplan-Meier curves showed that for the control group a linear function would yield a good approximation.
3. Each candidate model was required to contain an interaction term between each transformation of time and the screening indicator as well as the main effects of only screening indicator and transformation of time, so that the shape of the curves could vary between exposure strategies.
4. In addition to linear time, combinations of at least two and at most three other transformations of time were included. This restriction to a limited number of time variables controlled model complexity, given that the $KS$-measure does not feature a penalty term for model complexity.

Considering all possible permutations of the above pre-selected transformations of time, 20 candidate models were identified using the above steps. All models are summarized

**Table 4.1:** Candidate functional shape of a pooled logistic model to estimate CRC incidence

| **Model identifier** | **Linear predictor $\eta$** | **KS measure * 100** |
|---|---|---|
| 1 | $Q + Qt + Q\exp(t) + Q\frac{1}{\exp(t)} + t + \exp(t) + \frac{1}{\exp(t)}$ | 0.085 |
| 2 | $Q + Qt + Q\exp(t) + Q\log(t) + t + \exp(t) + \log(t)$ | 0.158 |
| 3 | $Q + Qt + Q\exp(t) + Qt^2 + t + \exp(t) + t^2$ | 0.472 |
| 4 | $Q + Qt + Q\exp(t) + Q\sqrt{t} + t + \exp(t) + \sqrt{t}$ | 0.223 |
| 5 | $Q + Qt + Q\frac{1}{\exp(t)} + Q\log(t) + t + \frac{1}{\exp(t)} + \log(t)$ | 0.127 |
| 6 | $Q + Qt + Q\frac{1}{\exp(t)} + Qt^2 + t + \frac{1}{\exp(t)} + t^2$ | 0.098 |
| 7 | $Q + Qt + Q\frac{1}{\exp(t)} + Q\sqrt{(t)} + t + \frac{1}{\exp(t)} + \sqrt{(t)}$ | 0.121 |
| 8 | $Q + Qt + Q\log(t) + Qt^2 + t + \log(t) + t^2$ | 0.115 |
| 9 | $Q + Qt + Q\log(t) + Q\sqrt{(t)} + t + \log(t) + \sqrt{(t)}$ | 0.134 |
| 10 | $Q + Qt + Qt^2 + Q\sqrt{(t)} + t + t^2 + \sqrt{(t)}$ | 0.112 |
| 11 | $Q + Qt + Q\exp(t) + Q\frac{1}{\exp(t)} + Q\log(t) + t + \exp(t) + \frac{1}{\exp(t)} + \log(t)$ | 0.088 |
| 12 | $Q + Qt + Q\exp(t) + Q\frac{1}{\exp(t)} + Qt^2 + t + \exp(t) + \frac{1}{\exp(t)} + t^2$ | 0.071 |
| 13 | $Q + Qt + Q\exp(t) + Q\frac{1}{\exp(t)} + Q\sqrt{t} + t + \exp(t) + \frac{1}{\exp(t)} + \sqrt{t}$ | 0.084 |
| 14 | $Q + Qt + Q\exp(t) + Q\log(t) + Qt^2 + t + \exp(t) + \log(t) + t^2$ | 0.080 |
| 15 | $Q + Qt + Q\exp(t) + Q\log(t) + Q\sqrt{t} + t + \exp(t) + \log(t) + \sqrt{t}$ | 0.092 |
| 16 | $Q + Qt + Q\exp(t) + Qt^2 + Q\sqrt{t} + t + \exp(t) + t^2 + \sqrt{t}$ | 0.078 |
| 17 | $Q + Qt + Q\frac{1}{\exp(t)} + Q\log(t) + Qt^2 + t + \frac{1}{\exp(t)} + \log(t) + t^2$ | 0.063 |
| 18 | $Q + Qt + Q\frac{1}{\exp(t)} + Q\log(t) + Q\sqrt{t} + t + \frac{1}{\exp(t)} + \log(t) + \sqrt{t}$ | 0.074 |
| 19 | $Q + Qt + Q\frac{1}{\exp(t)} + Qt^2 + Q\sqrt{t} + t + \frac{1}{\exp(t)} + t^2 + \sqrt{t}$ | 0.061 |
| 20 | $Q + Qt + Q\log(t) + Qt^2 + Q\sqrt{t} + t + \log(t) + t^2 + \sqrt{t}$ | 0.045 |

in Table 4.1, where beta coefficients are omitted and follow-up time is referred to as $t$ instead of $k_r$ to improve readability. A pooled logistic model was then fitted for each of these candidate models without covariate adjustment and the $KS$ measure was obtained. Furthermore, the parametric curves resulting from each candidate model were plotted against the non-parametric Kaplan-Meier curves for visual assessment. The $KS$ measures are given in Table 4.1. The visual comparison of the candidate models is given in Figure 4.3.

Model 20 yielded the smallest $KS$ measure. Visual assessment also confirmed that model 20 approximated the non-parametric curves very well, with no major deviations at any time of follow-up.

Non-parametric methods, i.e. the Aalen-Johansen estimator, could have been used instead of pooled logistic regression, when only baseline adjustment was necessary. For this, IPTW could have been used to obtain a weighted Aalen-Johansen estimator, which would have led to faster computation times when compared to the pooled logistic regression approach. However, the parametric methods were used here, because they can easily be extended to more complex settings. For instance, IPCW for artificial censoring in a per-protocol analysis is easily integrated in the parametric approach (see Section 4.5). Furthermore, the logistic regression approach can include covariates directly in the outcome model, as was done in a sensitivity analysis using the g-formula (see Section 4.4.4). Finally, the parametric approach yields smoothed curves that are less volatile in small sample sizes when compared to non-parametric methods.

## 4.3 Main results of Braitmaier et al. [2022b]

For the results of the original analysis, the reader is kindly referred to the publication Braitmaier et al. [2022b], which is printed in Section 7.2. As discussed in the paper, no relevant difference in effectiveness according to the site of CRC were found. Braitmaier et al. [2022b] contribute the differences reported in the literature to self-inflicted biases introduced by inappropriate study design.

Braitmaier et al. [2022b] was the first study to use TTE to investigate site-specific effectiveness of screening colonoscopy. The site-specific results, therefore, cannot be compared to other observational studies with similar methodology. However, the estimates for the effect on overall CRC incidence can be compared to other studies. García-Albéniz et al. [2017a] reported results for a US sample in the age group of 70 to 74 with an eight year follow-up. The shape of the adjusted cumulative incidence curves was very similar to the curves reported in Braitmaier et al. [2022b]. Overall, the incidence of CRC was slightly

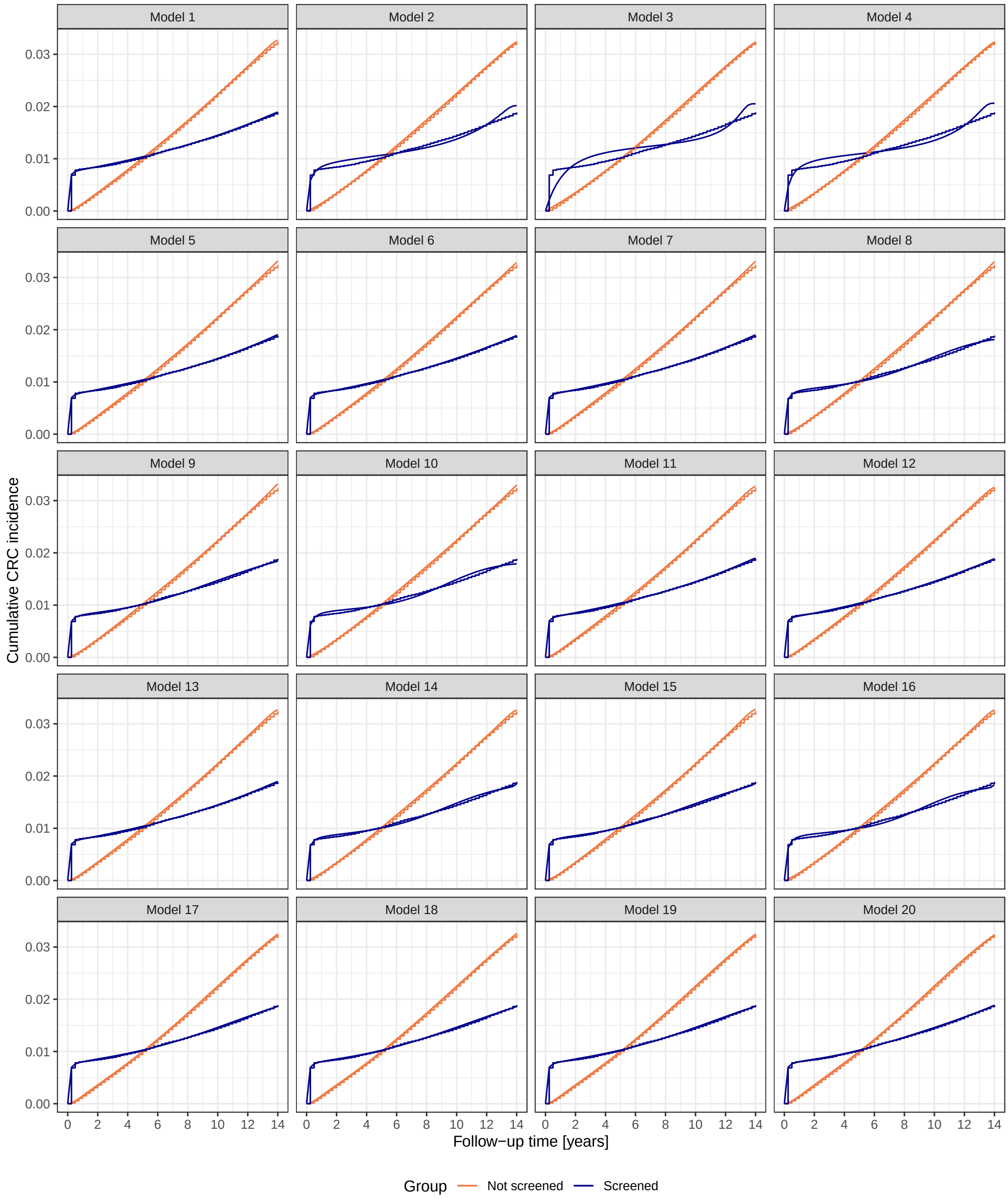


**Figure 4.3:** Comparison of non-parametric Kaplan-Meier curves with candidate parametric models for estimating CIFs. Step functions represent Kaplan-Meier curves while continuous line graphs represent the parametric estimate.

higher in both groups in García-Albéniz et al. [2017a], likely due to the older study population. The adjusted relative risk after eight years was 0.84, whereas the adjusted relative risk after eight years was 0.83 (95% CI: 0.78 - 0.88) when using the adjusted curves from Braitmaier et al. [2022b].

In 2022, after the publication of the initial findings, the results of the first RCT on the effectiveness of colonoscopy screening on CRC incidence (and mortality, although sample sizes were small) became available [Bretthauer et al., 2022]. While sample sizes were not sufficient to stratify analyses by site of CRC, the authors reported the effect of screening colonoscopy on overall CRC incidence. In the supplement to Braitmaier et al. [2022b], we provided results for the age group of 55 - 64, the same age group as was included in the RCT. Given that due to the assignment of individuals to the screening strategies in our emulated trial [Braitmaier et al., 2022b] there was no non-adherence at baseline, our results are more closely comparable to the per-protocol results published in Figure S3 in the supplement to Bretthauer et al. [2022] as compared to the intention-to-screen results published in the main paper, although contamination in the control arm during follow-up will likely have differed. There, the authors report an adjusted cumulative incidence at the end of the 10-year follow-up of 1.22 for the usual care (i.e. control) group and 0.84 for the screened group, resulting in an RR of 0.67. This is very close to the 11-year RR of 0.64 reported in Table S4 in the supplement to Braitmaier et al. [2022b]. While many differences in study design remain, this agreement in results appears to support the validity of our target trial emulation. However, further analyses were conducted to emulate the trial of [Bretthauer et al., 2022] more closely by using artificial censoring and IPCW to adjust for contamination of the control arm. These analyses are described below in section 4.5.

## 4.4 Sensitivity analyses

The assumptions underlying a causal interpretation of the results reported in Braitmaier et al. [2022b] were investigated as illustrated in the supplement to the published paper. Overlap plots of the propensity score for exposure at baseline were used to check for any indication for potential positivity violations. In a scenario without positivity violation and without confounding, the PS distributions of the exposure groups should overlap completely and be approximately identical. In a scenario without positivity violation, but with confounding by the observed covariates $X$, the PS distributions of the exposure groups should still cover the same value range, but the probability density functions will not be identical with more probability density towards 1 in the exposed group and more probability density towards 0 in the unexposed group. This scenario can be mitigated by appropriate adjustment for the covariates $X$. In a scenario with strong positivity violation, there should be visible non-overlap between the PS distributions, i.e. some or all of the PS distribution of one exposure group lies outside of the range covered by the PS distribution of the other group. Strong positivity violations cannot be mitigated by confounder adjustment, given that one would extrapolate beyond the data support. Instead, restricting the study population or changing the research question may be necessary.

Covariate balance after applying inverse probability weights was checked using the absolute standardized mean difference. This step is used to check if inverse weighting achieved satisfactory balance in observed covariates. For a continuous variable this measure is defined as

$$\text{ASMD} = \left| \frac{\bar{x}_{\text{treated}} - \bar{x}_{\text{untreated}}}{\sqrt{\frac{s^2_{\text{treated}} + s^2_{\text{untreated}}}{2}}} \right|, \tag{4.9}$$

while for a binary variable it is defined as

$$\text{ASMD} = \left| \frac{\bar{x}_{\text{treated}} - \bar{x}_{\text{untreated}}}{\sqrt{\frac{\bar{x}_{\text{treated}}(1-\bar{x}_{\text{treated}}) + \bar{x}_{\text{untreated}}(1-\bar{x}_{\text{untreated}})}{2}}} \right|. \tag{4.10}$$

Generally, if the absolute standardized mean difference after weighting is below 0.1, the respective covariate is considered sufficiently balanced [Austin, 2009]. These checks did not give reason for concern, as discussed in Braitmaier et al. [2022b].

However, the above checks are not sufficient to rule out all potential sources of bias. To identify any weaknesses impacting the validity of the main findings, sensitivity analyses were tailored to this study, acknowledging which aspects of study design or underlying data carry the largest risk.

### 4.4.1 Negative control outcome

The most common concern with observational data is that of confounding bias, given that no baseline randomization can be conducted. Several approaches exist to address the issue of unobserved confounding: Instrumental variable analyses circumvent the issue, but make strong assumptions and are only possible if an appropriate instrumental variable exists in the data [Greenland, 2000]. Quantitative bias analyses can be applied to investigate specific unobserved confounders when e.g. the strength of the association between confounder and outcome is known from the literature and the distribution of the confounder between exposure groups is varied across scenarios [Schneeweiss, 2006; Fox et al., 2022]. If, however, one is concerned about more than one variable, quantitative bias analysis is often too restrictive. In those settings, negative control analyses [Lipsitch et al., 2010] are valuable for detecting the presence of residual confounding. The rationale of negative control analyses was explained in depth in Section 2.7.1.

In the TTE on the effectiveness of screening colonoscopy [Braitmaier et al., 2022b], pancreatic cancer incidence was chosen as negative control outcome. While there are some differences in the sets of risk factors for the two types of cancer, there is also substantial overlap. Many factors contribute to these cancer entities and the following list is not exhaustive: The risk of both cancers is thought to increase with tobacco smoking and the extent of smoking, although different strengths of association have been reported in the literature for the two cancer entities with a stronger effect of current smoking on pancreatic than on colorectal cancer [Hannan et al., 2009; Lowenfels and Maisonneuve, 2005]. Similar associations with high alcohol intake have been reported for both cancer entities [McNabb et al., 2020; Wang et al., 2016]. Both pancreatic and colorectal cancer occur more often in individuals with type 2 diabetes [Lowenfels and Maisonneuve, 2005; Yu et al., 2022], which in turn is associated with obesity and sedentary lifestyle. Lifestyle factors are poorly reflected in health claims data, which makes pancreatic cancer a valuable negative control outcome candidate, given that screening colonoscopy cannot possibly have a causal effect on pancreatic cancer incidence.

The same statistical methods were used for the negative control outcome analysis as for the main analysis, including the same set of adjustment variables. Figure 4.4 shows the adjusted cumulative incidence functions over an eleven-year follow-up (the figure was adapted from Braitmaier et al. [2022b]). Confidence intervals were derived by bootstrapping. During the first seven years of follow-up, the cumulative incidence curves are nearly identical. After seven years, the curves diverge slightly, however, each curve is overlapped by the confidence interval of the other curve.

These results indicate that, under the assumption of U-comparibility explained in section 2.7.1, it is unlikely that there is major unmeasured confounding that could qualitatively change the results from the main analysis.

### 4.4.2 Treating competing events as censoring events

In the evaluation of the effectiveness of screening colonoscopy [Braitmaier et al., 2022b], the outcome of interest was CRC incidence. With this, death was a competing event. Brait-

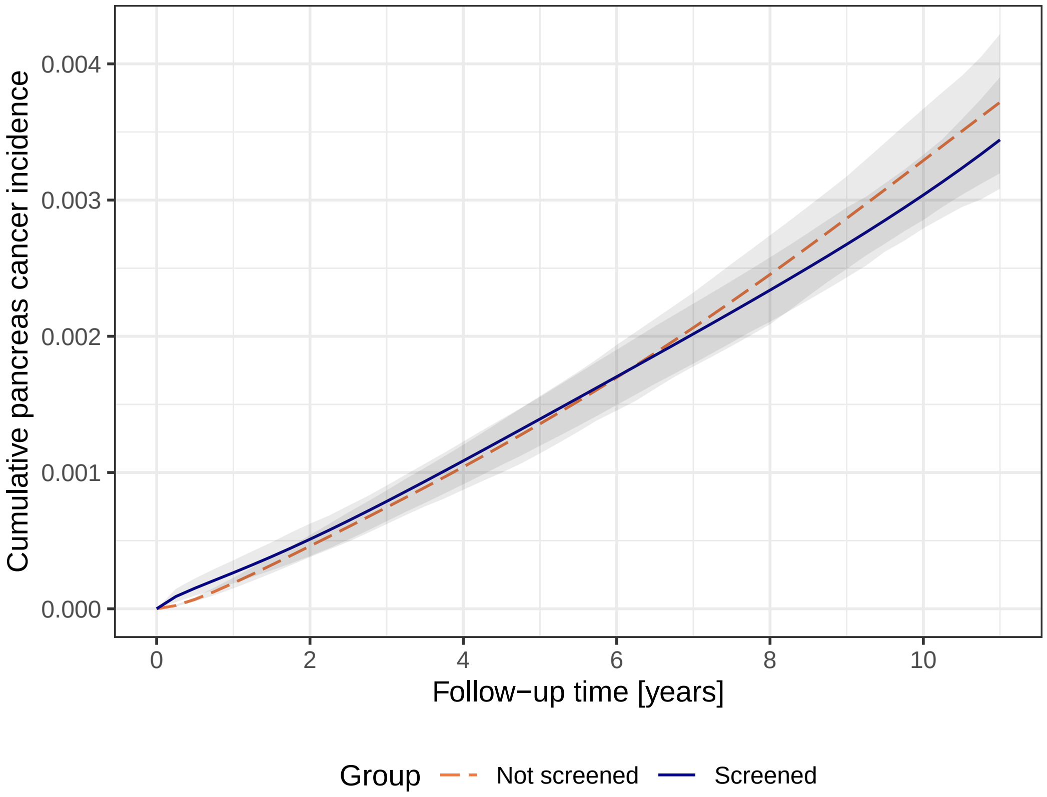


**Figure 4.4:** Adjusted cumulative incidence functions of the effect of screening colonoscopy on the negative control outcome of pancreatic cancer incidence (adapted from Braitmaier et al. [2022b])

maier et al. [2022b] reported the total effect of screening colonoscopy on site-specific and overall CRC incidence, also mediated by the competing event. In this approach, person-trials are not censored when experiencing a competing event. An alternative approach would be to estimate the controlled direct effect under elimination of competing events, i.e. censoring for competing events (see Young et al. [2020] for a discussion of total and controlled direct effect).

When death is the competing event, estimating the controlled direct effect is usually not very informative, given that it targets a hypothetical scenario under which the competing event is eliminated (i.e. in which no death occurs ever). However, censoring for competing events is often done in applied research without a sound causal justification. Therefore, a sensitivity analysis was carried out in which person-trials were censored at death. If the controlled direct effect were to differ substantially from the total effect, this could contribute to differences between Braitmaier et al. [2022b] and other published observational studies on screening colonoscopy.

While a comparison of total and direct effect was given in the supplement to Braitmaier et al. [2022b] for the effect on any CRC, Figure 4.5 shows a comparison by site. While the (baseline) adjusted risk estimates are slightly higher for the direct effect as compared to the total effect, the differences were small and did not substantially impact the results for any site.

A caveat to the results presented here is that further covariate adjustment would be needed for the controlled direct effect: First, censoring due to the competing event may introduce selection bias, which can be mitigated by adjusting for time-dependent covariates. Second, further assumptions regarding the adjustment set are needed for the controlled direct effect. Specifically, the adjustment set must also contain confounders between the competing event (death) and the event of interest (CRC incidence). The main objective of this sensitivity analysis, however, was to imitate the commonly-used approach to censor for death without further adjustment for time-dependent confounding.

Given that the effect estimates did not differ substantially between the two approaches, it is unlikely that the difference between Braitmaier et al. [2022b] and other published studies are due to a different approach regarding competing events.

### 4.4.3 Confounding between exposure and competing event

In Braitmaier et al. [2022b], covariates were selected so as to control for confounding between the exposure (screening colonoscopy) and the outcome of interest (CRC incid-

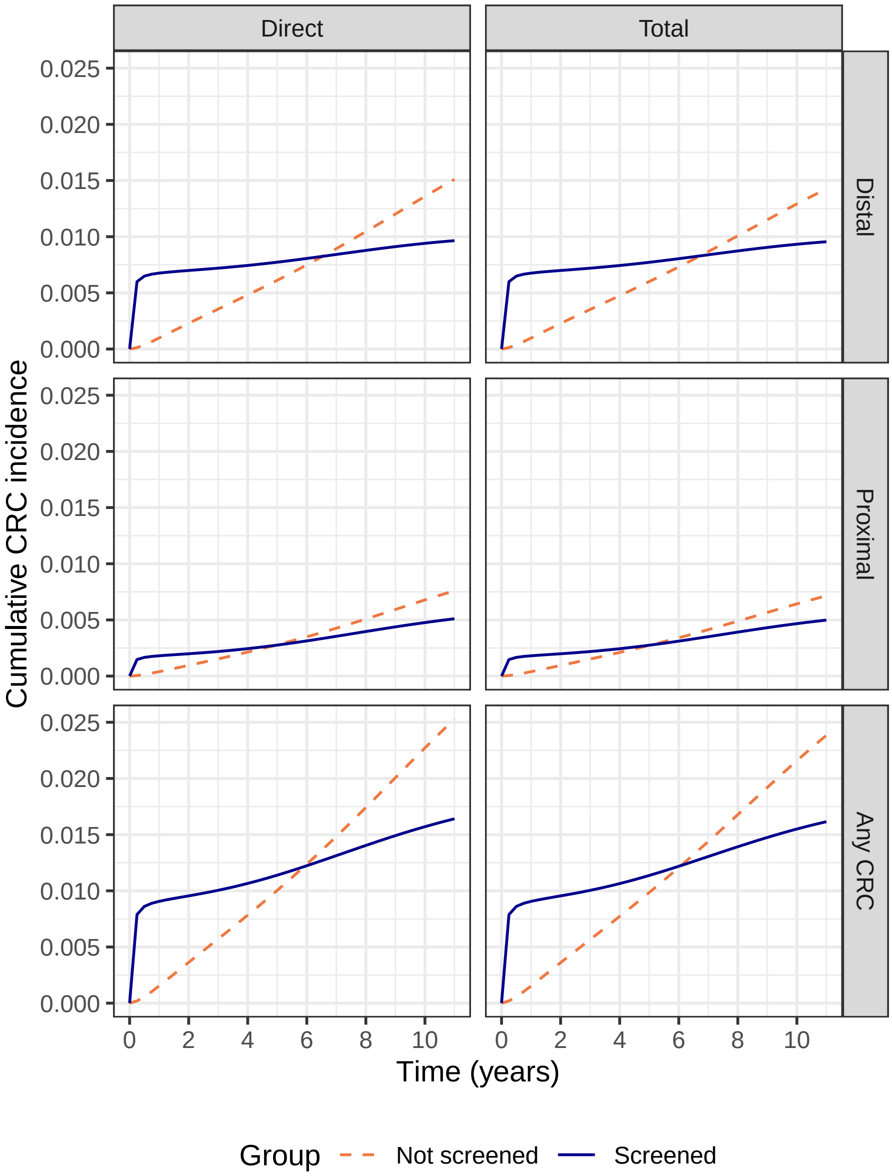


**Figure 4.5:** Comparison of total and controlled direct effect by site of CRC

ence). However, confounding between exposure and competing events may also lead to bias [Lesko and Lau, 2017]. It is best practice to report not only the adjusted effects of exposure on the outcome of interest, but also on the competing event [Latouche et al., 2013].

The problem of uncontrolled confounding in the example of screening colonoscopy and competing death is illustrated in Figure 4.6. Covariates $X$ were selected to include all variables that lead to confounding between exposure $A$ and outcome $Y$ if not controlled for. The thick directed edge from the competing event $D$ to the outcome of interest $Y$ indicates that $Y$ cannot happen, if it is precluded by $D$. The directed edge from $A$ to $D$ was omitted from the DAG, because screening colonoscopy affects overall mortality mainly through its effect on death from colorectal cancer, which only accounts for a negligible fraction of overall mortality. Adverse events of screening colonoscopy that lead to death, such as bleeding due to perforation of the colon, are not included in the discussion here because of their rarity. Laanani et al. [2019] found that perforation – which need not lead to death – occurred in 3.5 to 7.3 colonoscopies out of 10,000, with increasing rates at higher age (the age group considered here was comparably young) and decreasing rates with physician experience (physicians conducting screening colonoscopies in Germany are required to conduct at least 200 such procedures per year). Finally, node $U$ represents variables that lead to confounding between exposure $A$ and competing event $D$, if not adjusted for. No arrow was drawn from $U$ to $Y$, because covariates $X$ were selected based on subject-matter knowledge so as to include the most relevant predictors of $Y$. If $U$ was not present, no association between $A$ and $D$ should be apparent in the analysis.

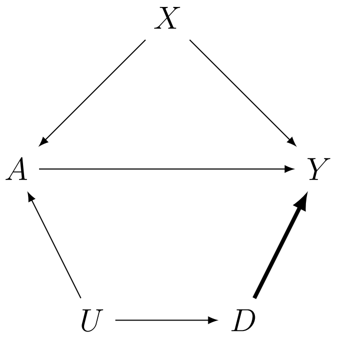


**Figure 4.6:** DAG of confounding between exposure $A$ and competing event $D$. The bold arrow from $D$ to $Y$ indicates that $D$ prevents $Y$ from happening.

In the current section, the adjusted, event-specific cumulative incidence curves for any death not preceded by a diagnosis of CRC are displayed for the two screening strategies of either attending screening colonoscopy during the baseline quarter or not (i.e. no sustained strategies). The same adjustment variables as in Braitmaier et al. [2022b] were

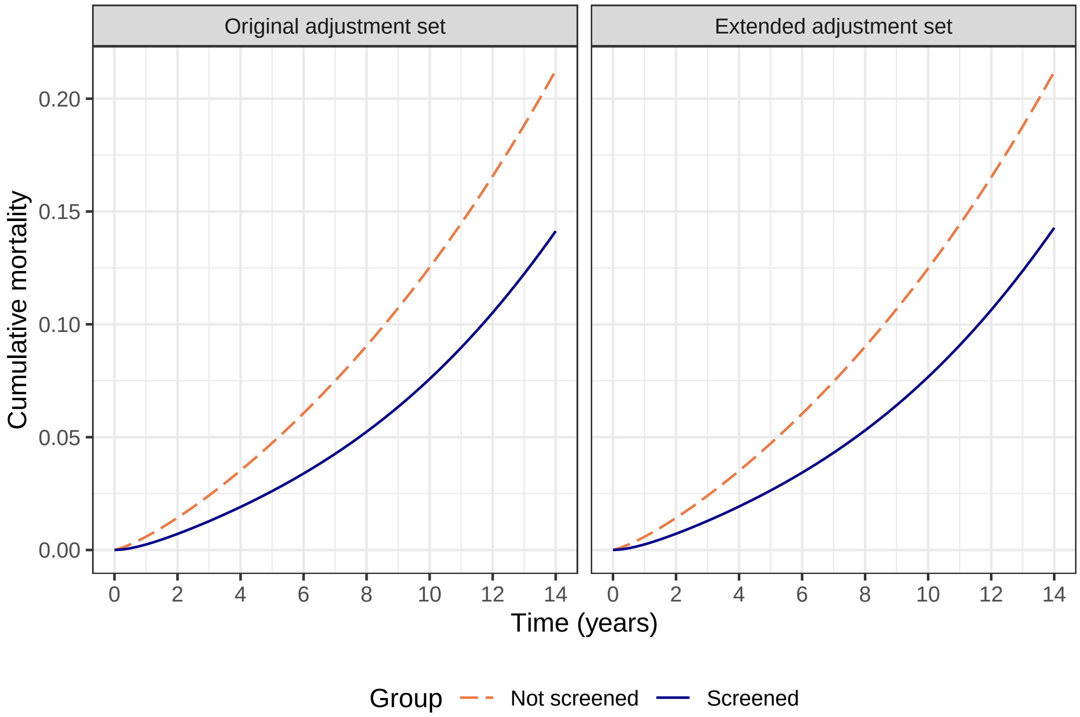


**Figure 4.7:** Adjusted, event-specific cumulative incidence curves for any death not preceded by CRC diagnosis

used. An extended adjustment set was also considered, including the following additional variables assumed to be relevant predictors of overall mortality: other cancer diagnoses, therapy with cytostatics, therapy with monoclonal antibodies, inpatient chemotherapy, radiotherapy, palliative care, antidepressant prescriptions, antipsychotic prescriptions, asthma, chronic obstructive pulmonary disease, coronary heart disease, dementia, drug abuse, chronic heart failure, hepatitis, treated hypertension, immunosuppressants, platelet aggregation inhibitors, lipid lowering drugs, liver disease, severe liver disease, acute myocardial infarction, hemiplegia, renal disease, stroke.

As is evident from Figure 4.7, the expectation of no association between exposure $A$ and competing event $D$ was not reflected by the adjusted, event-specific cumulative incidence curves. The mortality of the control group was substantially higher than that of the screened group throughout follow-up. The curves did not change notably when adjusted for further covariates. This result indicates the presence of bias, e.g. due to uncontrolled confounding between exposure to screening colonoscopy and overall mortality.

The result obtained here matches well with the published literature: García-Albéniz et al. [2017a] estimated the effect of screening colonoscopy on CRC incidence. After adjusting for confounding, they found a beneficial effect of screening on the 8-year cumulative incidence of CRC. However, they reported in a follow-up paper that the effect (or lack thereof) of screening colonoscopy on overall mortality was ”hopelessly confounded” [García-Albéniz et al., 2019], with an implausible reduction of overall mortality in the screened group.

When judging how much the observed confounding between exposure $A$ and competing event $D$ may have affected the estimates for the effect on the event of interest $Y$, results obtained in a simulation study by Lesko and Lau [2017] are helpful: The authors compared scenarios where the adjustment set included either only confounders of the exposure-outcome effect, or included confounders with both the outcome and the competing event. They found that omitting confounders for the effect on the competing event substantially biased results regarding the total effect, but not results regarding the controlled direct effect. Indeed, when one considers censoring a form of controlling for the occurrence of competing events, the backdoor path $A \leftarrow U \rightarrow D \rightarrow Y$ in Figure 4.6 would be blocked by $D$. With this in mind, total and controlled direct effect should substantially differ, if confounding between $A$ and $D$ was affecting the estimates for the effect of $A$ on $Y$. As discussed in Section 4.4.2, this was not the case in Braitmaier et al. [2022b]. Confounding of the effect of $A$ on $Y$ due to unobserved common causes of $A$ and $D$ is, therefore, unlikely to have played a role in Braitmaier et al. [2022b]. Nevertheless, uncontrolled residual confounding between $A$ and $D$ must be considered a limitation of the data source.

### 4.4.4 Covariate adjustment via g-formula instead of IPTW

IPTW was used in Braitmaier et al. [2022b] to adjust CIFs for confounding by observed covariates. In this approach, an exposure model is fitted to obtain propensity scores, which in turn are used to obtain adjustment weights. The outcome model itself does not include covariates, but is adjusted for confounding via the weights (see Section 2.5). This approach assumes that the exposure model is correctly specified.

An alternative approach is to include covariates in the outcome model instead and suitably standardize, which is known as the g-formula approach [Robins, 1986; Hernán and Robins, 2020], also called direct standardization. Here, no exposure model is required. Instead, covariates are included in the outcome model, which is fitted on the observed data, using observed exposure. The fitted model is then used to obtain predictions of the potential outcomes for each exposure level by modifying the original dataset so that all entries share the same exposure, potentially contrary to the factually observed exposure. Finally, marginal estimates are obtained by averaging over all observations. In settings with time-dependent

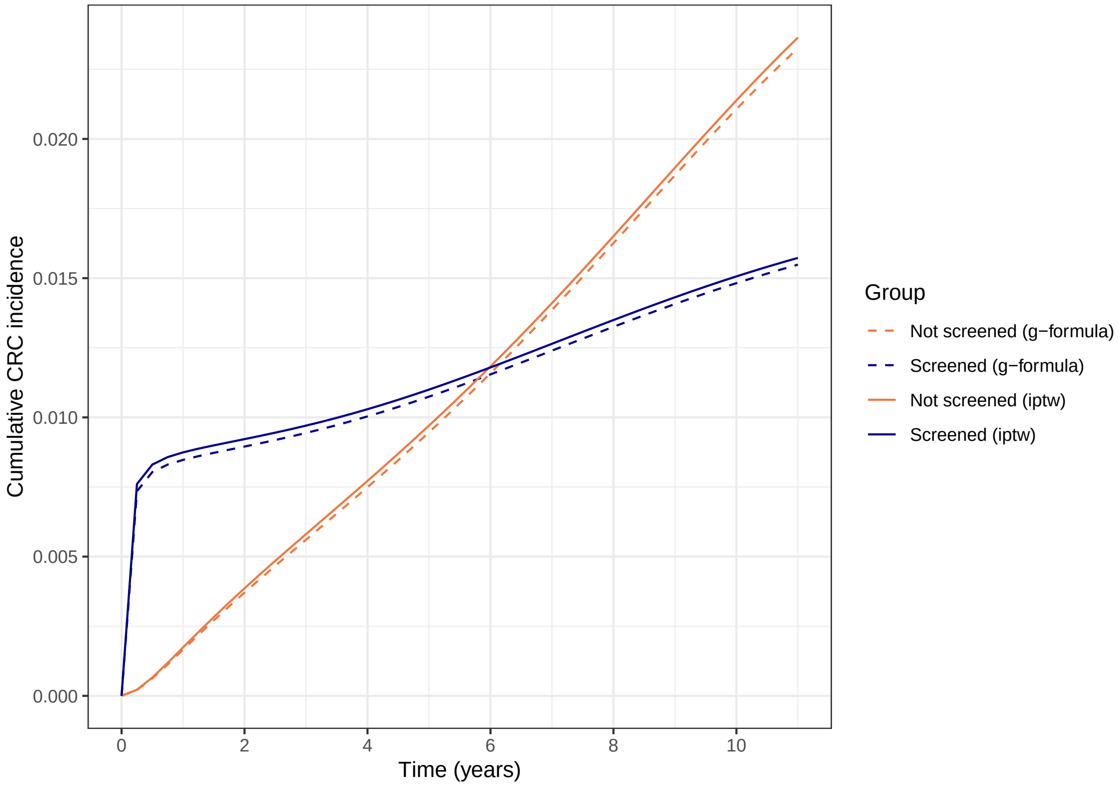


**Figure 4.8:** Comparison of IPTW and g-formula adjustment for baseline covariates

confounding, further modeling of covariates is needed (see e.g. Börnhorst et al. [2021]). A core assumption of this approach is that the outcome model is correctly specified.

A sensitivity analysis using the g-formula approach instead of IPTW was conducted for the results reported in Braitmaier et al. [2022b]. This analysis served two purposes: First, covariates might affect exposure differently than they affect the outcome. While not proving correctness of model specification, similar results from IPTW and g-formula methods may at least indicate that no strong model misspecification is present, unless one believes that both models are equally misspecified. This analysis, therefore, was a sensitivity analysis regarding model misspecification. Second, since the program code for the analysis was written from scratch, this sensitivity analysis served as a validation of the program code. The g-formula approach is an alternative to MSMs using IPTW [Robins et al., 2000]. Vastly differing results could therefore also indicate issues with the program code.

Here, the g-formula approach was as follows: First, a pooled logistic regression estimating the probability of not experiencing the outcome by time $k_r$ was fitted as

$$\mathbb{P}\left[Y_{j,k_r} = 0 | \bar{Y}_{j,k_r-1} = 0, Q_j, X_{j,t_r}\right] = \text{logit}^{-1}\left(\eta(q_j, k_r) + x_{j,t_r}\gamma\right). \tag{4.11}$$

In Equation 4.11, the same functional shape of time $\eta(q_j, k_r)$ was used as in the main analysis (Equation 4.2). However, Equation 4.11 also contained baseline covariates $x_{j,t_r}$ and did not use inverse weighting. The list of baseline covariates used was identical to the ones used in the inverse weighting approach and is given in Braitmaier et al. [2022b]. The model included main effects only, i.e. no transformations of or interactions between covariates were included in the model. As in the main analysis, a subdistribution approach was used, i.e. competing events were not treated as censoring events.

Once fitted, this pooled logistic regression was used to obtain predicted probabilities. For this, the original analysis dataset was copied twice, once setting $Q = 0$ and once setting $Q = 1$. These two modified datasets were used as input to the fitted model to obtain predicted probabilities

$$\hat{p}\left[Y^q_{j,k_r} = 0 | Y^q_{j,k_r-1} = 0, Q = q, X_{j,t_r}\right]. \tag{4.12}$$

Next, the cumulative, event-specific, person-trial-level risk conditional on baseline covariates was estimated by building the cumulative product over time as

$$\hat{p}\left[Y^q_{j,k_r} = 1 | Q, X_{j,t_r}\right] = 1 - \prod_{l=1}^{k_r} \hat{p}\left[Y^q_{j,l} = 0 | Y^q_{j,l-1} = 0, Q = q, X_{j,t_r}\right]. \tag{4.13}$$

Finally, the marginal, event-specific cumulative incidence function for strategy $Q = q$ was obtained by standardizing over the population as

$$\hat{p}\left[Y^q_{k_r} = 1\right] = \frac{1}{m} \sum_{j=1}^{m} \hat{p}\left[Y^q_{j,k_r} = 1 | Q = q, X_{j,t_r}\right]. \tag{4.14}$$

A comparison of the results obtained with both methods is given in Figure 4.8. As is evident from the figure, the two approaches yielded very similar results. This sensitivity analysis, therefore, did not find evidence for either model misspecification or issues with the program code.

### 4.4.5 Changing the adjustment set

Adjustment variables were selected a priori based on subject-matter expertise for the analyses described in Braitmaier et al. [2022b], i.e. no data-driven covariate selection strategy was used. Furthermore, adjustment variables were included in the exposure model and not

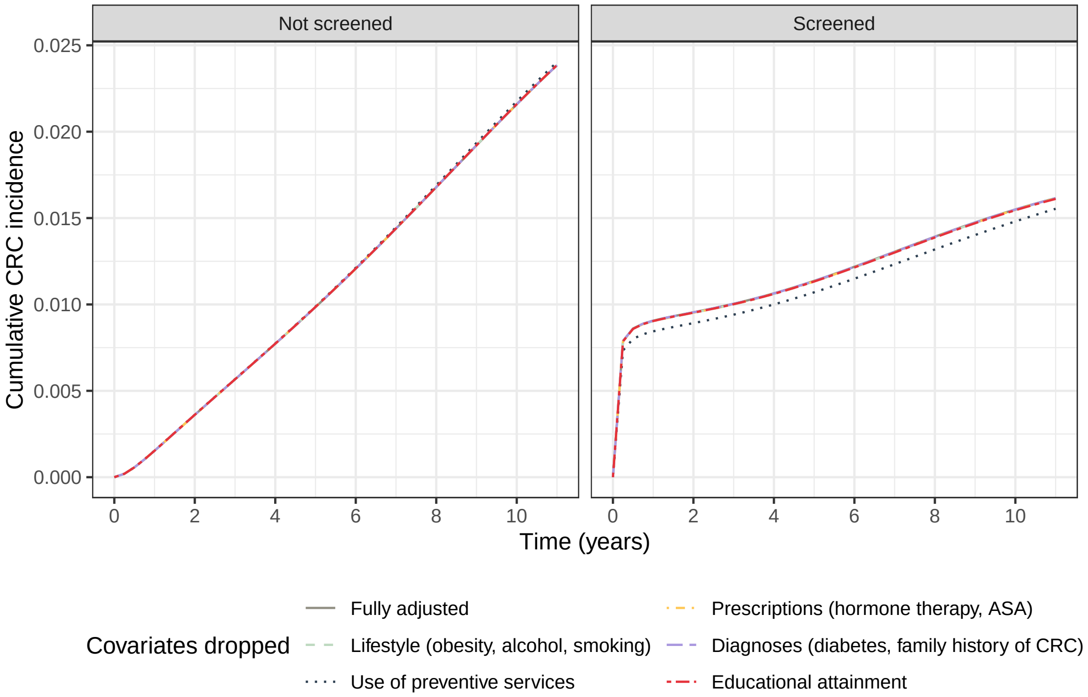


**Figure 4.9:** Sensitivity analysis regarding selection of covariate set; ASA: Acetylsalicylic acid, CRC: colorectal cancer

the outcome model, since IPTW was used for covariate adjustment. As such, no information regarding the association of individual covariates on the outcome were available to judge variable importance.

A sensitivity analysis was conducted to assess the magnitude of the influence that individual groups of covariates had on the result. In this analysis, (groups of) covariates were dropped from the adjustment set and the primary analysis was repeated. A covariate was assumed to have a large effect on the result, when the CIFs changed notably after dropping said covariate.

The primary results reported in Braitmaier et al. [2022b] were adjusted for the following baseline covariates (included as main effects in the propensity model): Age, sex, educational attainment (unknown or no degree, secondary degree, higher education), obesity, family history of CRC, menopausal hormone therapy, acetylsalicylic acid, diabetes, codes indicating alcohol abuse, codes indicating smoking and use of other preventive services before baseline (none, one, two or more). Five covariate sets were studied in this sensitivity analysis, dropping the following covariates, but keeping the others:

- Lifestyle factors (obesity, alcohol abuse, smoking)
- Use of other preventive services

- Prescriptions (menopausal hormone therapy, acetylsalicylic acid)
- Diagnoses (family history of CRC, diabetes)
- Educational attainment

The results of these sensitivity analyses are given in Figure 4.9. Most modified covariate sets yielded results that did not differ from the fully adjusted main effects model. The only exception was use of other preventive services in the three years before baseline. While other preventive measures, such as skin cancer screening or general health check-ups, are unlikely to affect the risk of developing CRC, they were assumed to be a proxy for health seeking behavior and health consciousness, but also general health status. Person-trials assigned to the screening strategy had previously undergone other preventive measures at a higher rate when compared to the non-screening strategy (see Table 1 in Braitmaier et al. [2022b]).

Assuming that participation in preventive measures is a proxy for health-seeking behavior and health consciousness and further assuming that health consciousness is associated with better health in general, an enrichment of the screening strategy with health conscious individuals would lead to a decreased risk of developing CRC, among other diseases. Indeed, when dropping use of preventive services from the adjustment set, the cumulative incidence of CRC dropped notably in the screening arm.

While use of preventive services was included as an adjustment variable in Braitmaier et al. [2022b], it was considered as a restriction criterion for subgroup analyses to increase internal validity and homogeneity across exposure groups in Braitmaier et al. [2022a] in the context of mammographic screening.

### 4.4.6 Non-truncated weights

Following the example of previous studies using target trial emulation (e.g. García-Albéniz et al. [2017a]), inverse weights used in the analyses described in Braitmaier et al. [2022b] were truncated at the 99th percentile. The goal of this truncation approach is to limit the influence of extreme observations. This, however, is a trade-off. If extremely large weights were present for a small set of individuals with unusual covariates, these few individuals would have a disproportionate influence on the adjusted effect estimates. Truncation, then, makes results more representative of the study population. Furthermore, truncating weights can decrease variance, leading to more efficient estimators. However, truncated weights may fail to remove confounding completely, thus leading to residual confounding [Goetghebeur et al., 2020].

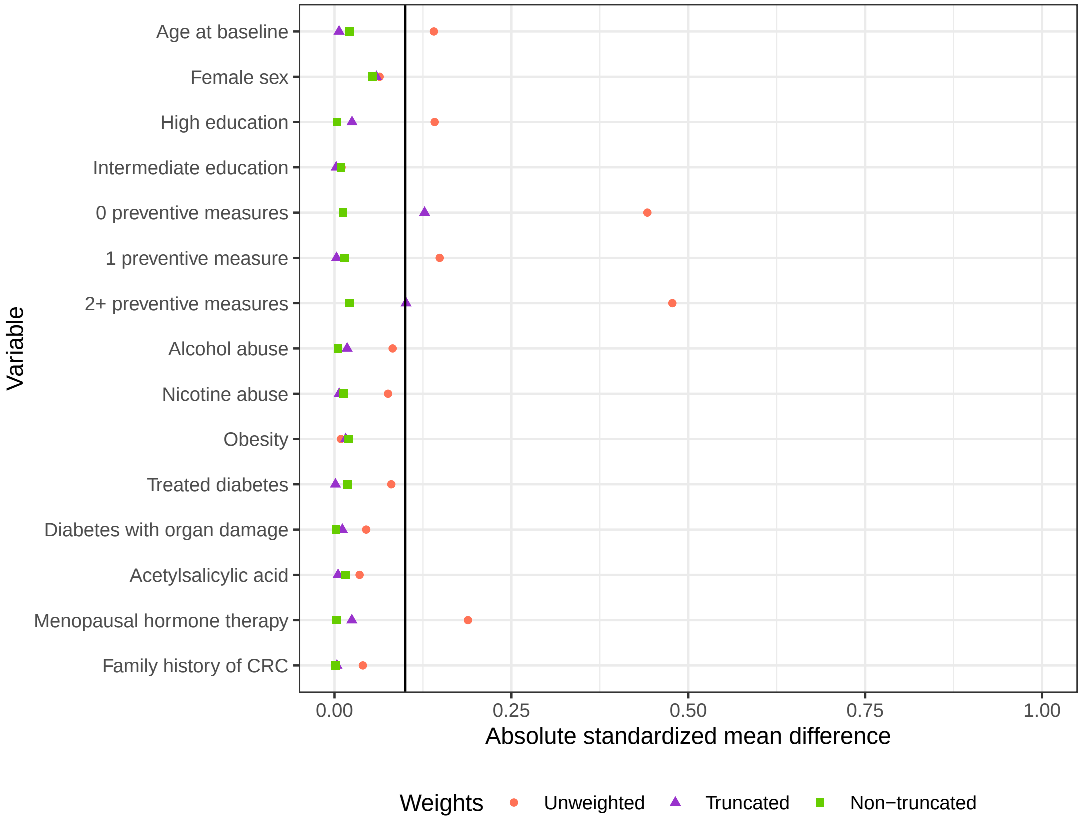


**Figure 4.10:** Covariate balance using truncated versus non-truncated inverse weights

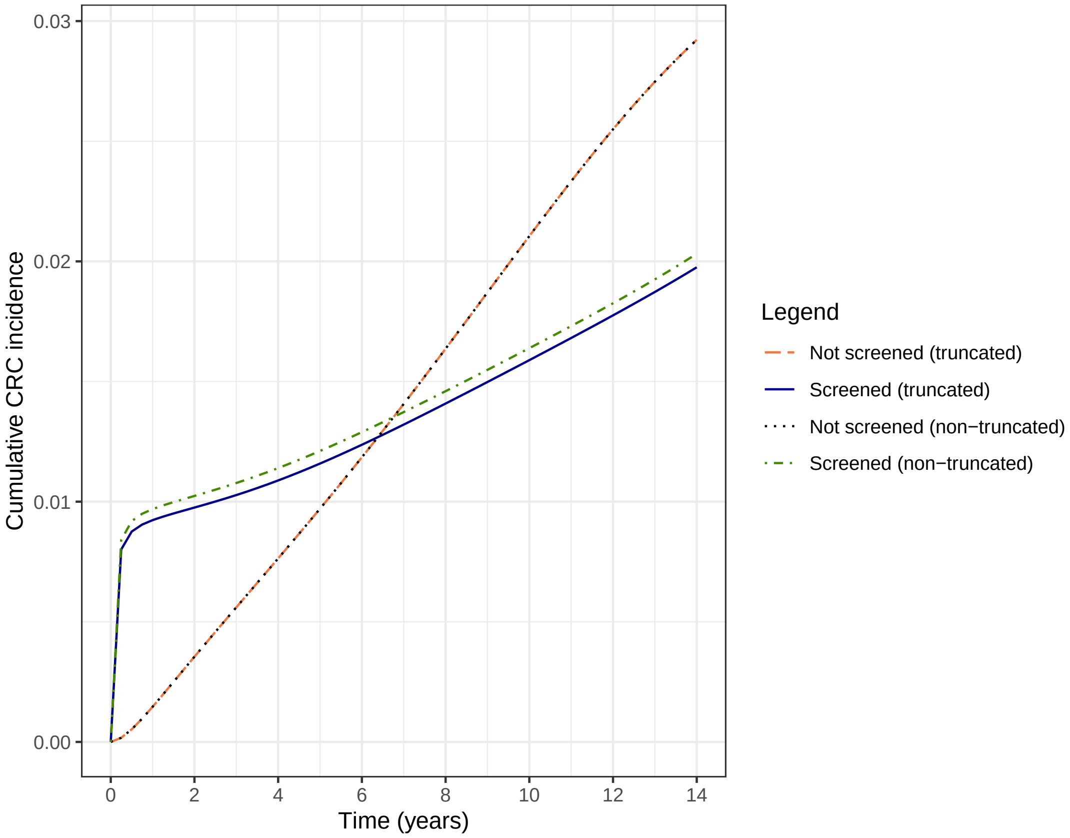


**Figure 4.11:** Adjusted cumulative CRC incidence curves using truncated versus non-truncated inverse probability weights to adjust for baseline confounding

The analysis regarding the causal effect of baseline screening colonoscopy on CRC risk during follow-up [Braitmaier et al., 2022b] was therefore repeated without truncating the inverse probability weights. Additional data years were available in this sensitivity analysis when compared to the initial publication described in Braitmaier et al. [2022b], resulting in slightly different results. Covariate balance before and after weighting (with truncated and non-truncated weights) is compared in Figure 4.10. Notably, most covariates were well balanced with both weights. However, the use of other preventive services before baseline could not be fully balanced with the truncated weights. With non-truncated weights on the other hand, also the use of other preventive services was well-balanced. This could indicate that the variable regarding use of preventive services identifies subgroups that almost never (namely those who do not undergo any preventive services) or almost always (namely those who also undergo several other preventive services) participate in colonoscopy screening.

Figure 4.11 shows the adjusted cumulative CRC incidence by screening strategy, once with truncated and once with non-truncated weights. While results are very similar and no clinically relevant changes result from using non-truncated weights, the cumulative incidence curve for the screening arm changes notably with a slightly higher cumulative incidence. Indeed, the relative risk reduction at the end of follow-up is slightly smaller when using non-truncated weights. The adjusted 14-year relative risk using truncated weights was 0.68 (risk reduction of 32%), whereas the relative risk using non-truncated weights was 0.70 (risk reduction of 30%).

### 4.4.7 Varying the number of bootstrap samples

As discussed above, standard parametric approaches for estimating confidence intervals are not valid in the emulated target trial setting described in Braitmaier et al. [2022b], since the same individual is potentially included in the dataset more than once. Instead, person-level bootstrapping as described in Section 4.1 was used. Given that the underlying statistical methods are computationally heavy and the analyzed datasets are large, obtaining bootstrap-based confidence intervals can become computationally prohibitive when using a large number of bootstrap samples. As a compromise between statistical accuracy and computational feasibility, 250 bootstrap samples were used in Braitmaier et al. [2022b], whereas 500 bootstrap samples are a more common choice in the literature. To investigate whether $B = 250$ bootstrap samples were sufficient to reliably estimate confidence intervals, the same procedure was repeated in a sensitivity analysis regarding the incidence curves for any CRC, but this time with $B = 500$ bootstrap samples. A comparison of the resulting confidence intervals is given in Figure 4.12.

As is evident in Figure 4.12, the bootstrap-based confidence intervals based on 250 versus 500 bootstrap samples do not differ notably. This result suggests that in the analysis reported in Braitmaier et al. [2022b] results would not have changed, had the number of bootstrap samples been larger. $B = 250$ was a sufficiently large number of bootstrap samples to reliably estimate 95% confidence intervals.

## 4.5 Update and per-protocol analysis

### 4.5.1 Rationale and methods

After the initial publication of Braitmaier et al. [2022b], the results of the first and only RCT comparing the effectiveness of screening colonoscopy at baseline versus no screening colonoscopy at baseline were published (see Bretthauer et al. [2022]). One major difference in the study designs of the emulated trial of Braitmaier et al. [2022b] and the NordICC trial [Bretthauer et al., 2022] was that the NordICC trial was conducted at a time when screening colonoscopy was not offered to the wider population in the coun-

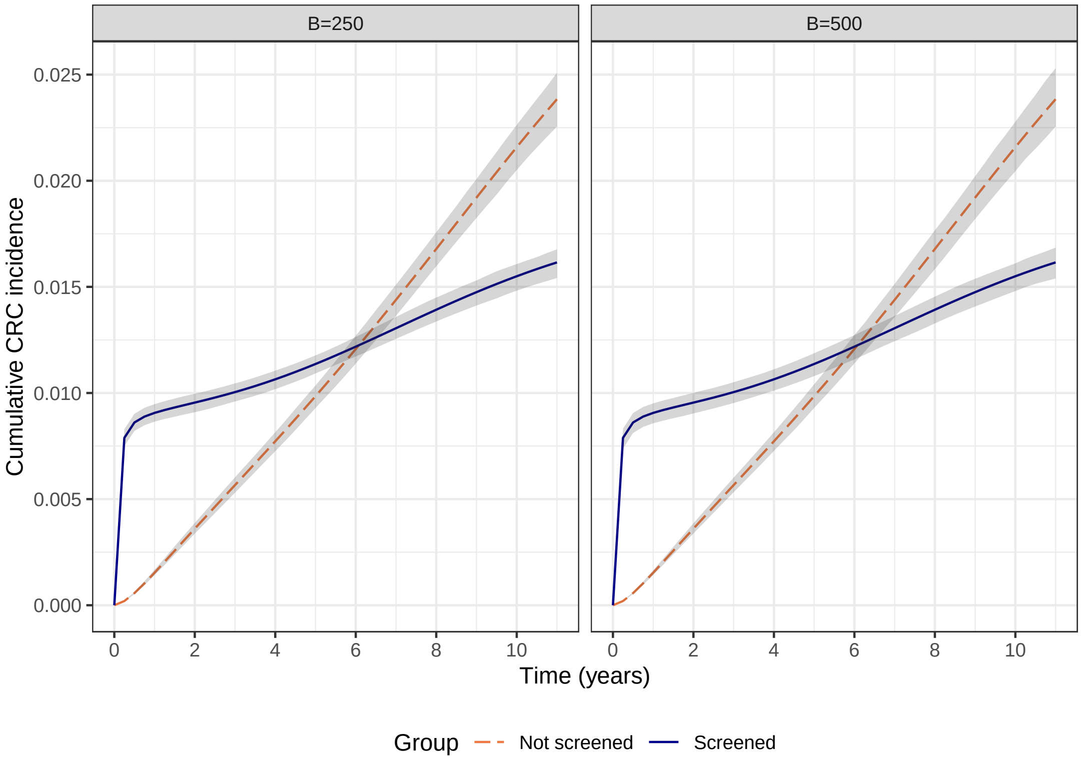


**Figure 4.12:** 95% confidence intervals based on 250 versus 500 bootstrap samples

tries involved in the study. This means that the control group did not feature a substantial contamination with screening colonoscopies conducted during follow-up. In Braitmaier et al. [2022b] on the other hand, screening colonoscopy was freely available to all eligible individuals. With the initial analysis scheme reported in Braitmaier et al. [2022b] where no restrictions regarding screening colonoscopy use during follow-up were made, this led to a contamination of the control arm.

A per-protocol analysis was added to make results more comparable. Additional data years had become available since the original publication [Braitmaier et al., 2022b] and the follow-up was extended. Furthermore, additional sequential trials were emulated until the end of 2013 (as opposed to 2011).

The censoring scheme for the per-protocol analysis was as follows: For person-trials assigned to the strategy with screening colonoscopy at baseline no artificial censoring was applied, since screening colonoscopy is a quasi point exposure with the option of repeat screening colonoscopy only once after ten years. For person-trials assigned to the control strategy, artificial censoring occurred at the end of the calendar quarter with the first screening colonoscopy. Screen-detected CRC was not counted as an outcome in this strategy, with screen-detected CRC being defined as a CRC diagnosis in the same calendar quarter as a screening colonoscopy or with a screening colonoscopy in the 180 days preceding the diagnosis. Importantly, even when controlling for baseline confounders via randomization or adjustment, artificial censoring may introduce bias, if time-dependent covariates affect both the probability of being censored and the probability of experiencing the outcome. Adjustment for time-dependent covariates becomes necessary.

Adjustment for baseline confounding and time-dependent confounding/informative censoring followed the approach described in [Robins et al., 2000]: Adjustment for baseline covariates was achieved as before by constructing IPTW weight contributions. Adjustment for time-dependent covariates was achieved by constructing time-dependent IPCW weight contributions using time-updated versions of the baseline covariates. The censoring model contained the same covariates as in the main analysis, namely main effects of number of preventive services (0, 1, 2 or more), acecylsalicylic acid, age, codes indicating alcohol abuse, family history of CRC, diabetes with pharmacological treatment, diabetes with organ damage, female sex, educational attainment, menopausal hormone therapy, obesity and smoking. Covariate balance could initially not be achieved for all covariates throughout follow-up, which led to the inclusion of the following interaction terms in the censoring model: menopausal hormone therapy with age categories (55 to 59, 60 to 64, 65 to 69, 70 to 74, 75 and older), menopausal hormone therapy with calendar year and menopausal hormone therapy with family history of CRC.

As described under section 4.1, baseline weight contributions were defined as

$$\widehat{iptw}_j = \frac{\hat{p}\left[Q_j = 1\right]}{\hat{p}\left[Q_j = 1 | X_{j,t_r}\right]} \tag{4.15}$$

for exposed person-trials and as

$$\widehat{iptw}_j = \frac{1 - \hat{p}\left[Q_j = 1\right]}{1 - \hat{p}\left[Q_j = 1 | X_{j,t_r}\right]} \tag{4.16}$$

for unexposed person-trials, with $X_{j,t_r}$ being the covariate vector of person-trial $j$ at the start of emulated trial $r$. Time-dependent weight contributions for artificial censoring were set to $\widehat{iptw}_{j,k_r} = 1$ for time $k_r = 1, ..., K_r$ and person-trials assigned to the active screening strategy $Q = 1$, since no artificial censoring was carried out in this strategy. Furthermore, $\widehat{iptw}_{j,1}$ was set to 1 for all person-trials $j$, because no artificial censoring was possible during the first time interval by design (exposure to screening colonoscopy in the first time interval was used for assignment to the exposure strategies $Q$).

For unexposed person-trials, the probability of not being artificially censored ($\text{Cens}_{j,k_r} = 0$) was estimated for each time point $k_r > 1$ using a pooled logistic model as described in Robins et al. [2000]:

$$\mathbb{P}\left[\text{Cens}_{j,k_r} = 0 | X_{j,k_r}\right] = \text{logit}^{-1}\left[\beta_0 + k_r\beta_1 + x_{j,k_r}\beta_4\right]. \tag{4.17}$$

Censoring model 4.17 also included the above-mentioned interaction terms relating to menopausal hormone therapy in the covariate vector $x_{j,k_r}$.

Another pooled logistic model was fitted to obtain the numerator of the weight contributions as

$$\mathbb{P}\left[\text{Cens}_{j,k_r} = 0\right] = \text{logit}^{-1}\left[\beta_0 + k_r\beta_1\right]. \tag{4.18}$$

The time-dependent, stabilized weight contribution for person-trials in the strategy with no screening colonoscopy at baseline for time point $k_r$ was then computed using the predicted probabilities from the fitted models as

$$\widehat{ipcw}_{j,k_r} = \frac{\prod_2^{k_r} \hat{p}\left[\text{Cens}_{j,k_r} = 0\right]}{\prod_2^{k_r} \hat{p}\left[\text{Cens}_{j,k_r} = 0 | X_{j,k_r}\right]}. \tag{4.19}$$

Final weights are then given as

$$\widehat{sw}_{j,k_r} = \widehat{iptw}_j * \widehat{ipcw}_{j,k_r}. \tag{4.20}$$

Stabilized weights are generally preferred over non-stabilized weights when assessing sustained treatments, because non-stabilized weights can grow very large, especially when many time points are considered, leading to unstable estimators [Hernán and Robins, 2020; Robins et al., 2000]. The stabilized weights were truncated at the 99th percentile of their distribution to avoid excessive influence of outliers.

Covariate balance of time-dependent confounders was assessed throughout follow-up by calculating the absolute standardized mean difference between screening strategies for each time point $k_r$ after applying the weights (see Figure 4.13). When balance could not be achieved for a confounder at baseline, sensitivity analyses were conducted within strata of the baseline covariate. When balance could not be achieved for a time-dependent covariate, sensitivity analyses were conducted by including the time-dependent covariate

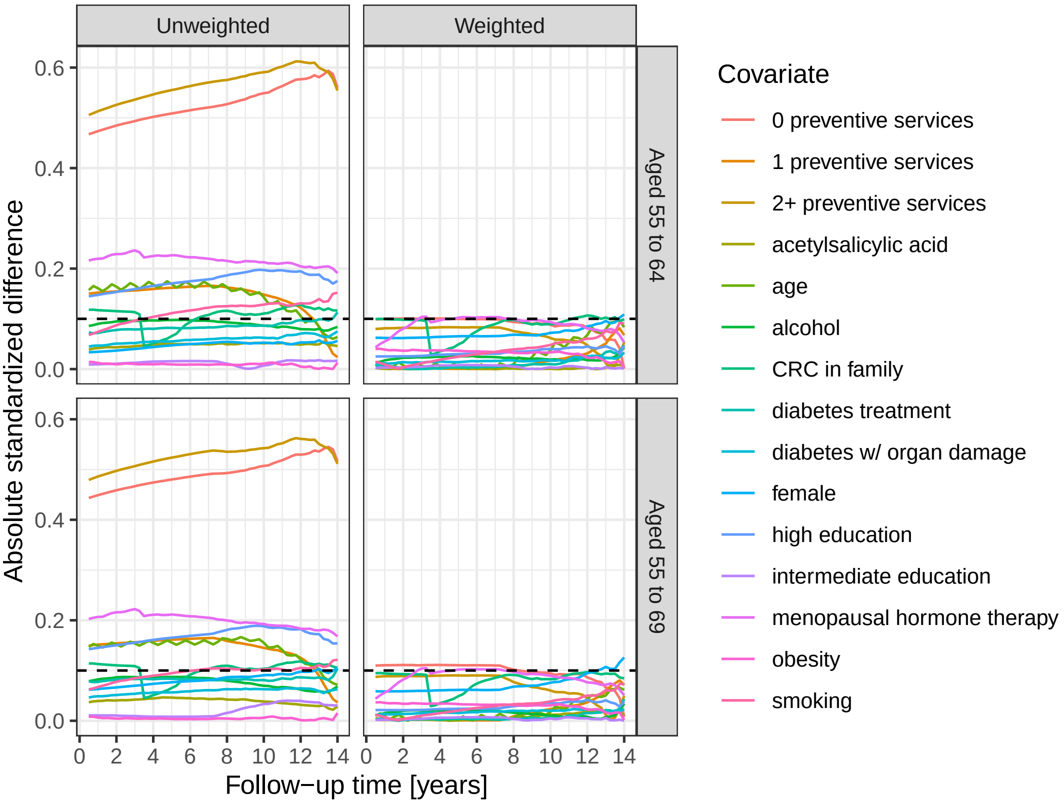


**Figure 4.13:** Covariate balance over follow-up in per-protocol analysis of screening colonoscopy. Dashed line represents the 0.1 threshold

that could not be balanced in the outcome model equation.

### 4.5.2 Results and discussion

Overall, 1,642,348 person-trials were included in the control strategy and 240,193 in the screening strategy when considering individuals aged 55 to 69 years old. In the control strategy, 18.7 % underwent screening colonoscopy at some point during the 14-year follow-up and were artificially censored in the per-protocol analysis. An additional analysis was conducted, restricting the population to the age group of 55 to 64, i.e. the age group assessed by [Bretthauer et al., 2022]. Here, 1,110,465 person-trials were included in the control group and 171,310 in the screening group. In the control strategy, 20.3 % underwent screening colonoscopy during follow-up and were artificially censored.

The 14-year adjusted RR among individuals aged 55 to 69 at baseline was 0.68 when assessing a point exposure control strategy without artificial censoring and 0.72 when assessing a sustained control strategy with artificial censoring and IPCW. Among individuals aged 55 to 64, the adjusted RR was 0.67 for the point exposure control strategy and 0.72 for the sustained control strategy. In both of these age groups, a point exposure

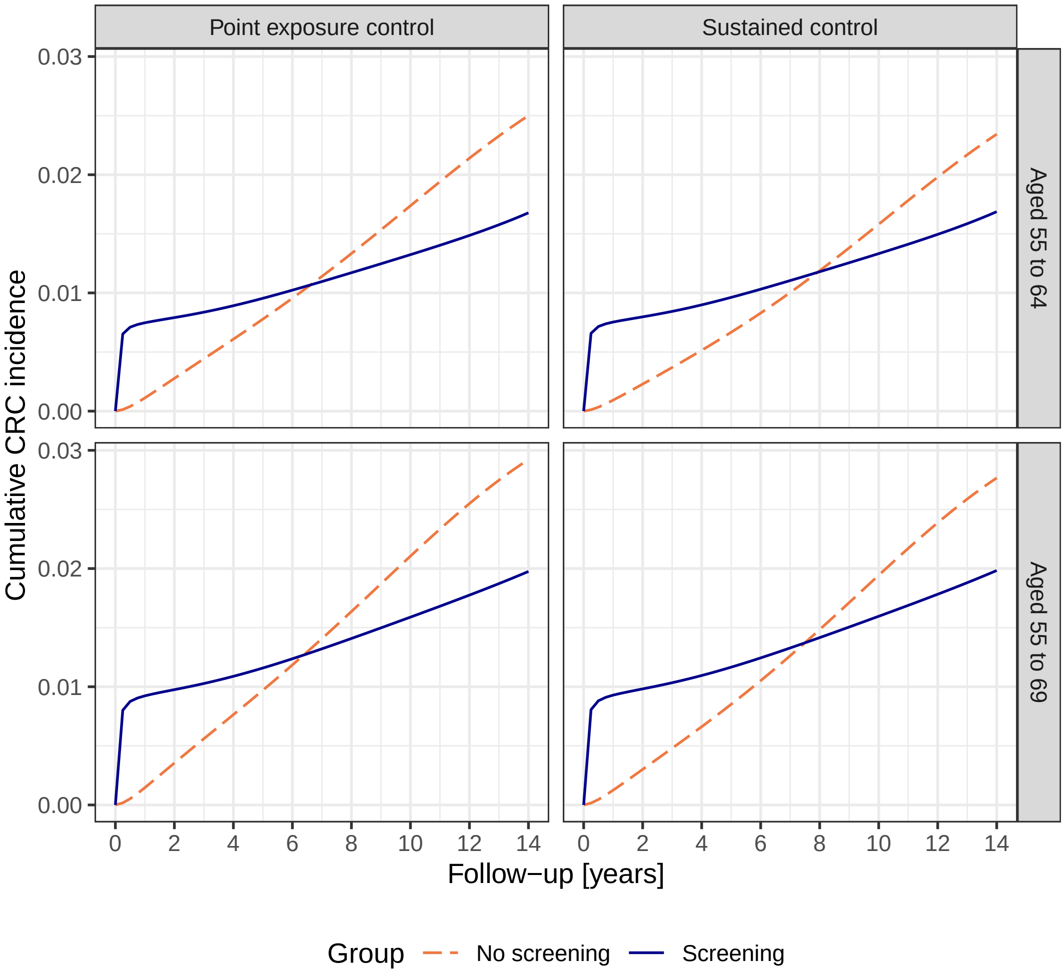


**Figure 4.14:** Per-protocol results for the effect of screening colonoscopy

control strategy led to a higher risk of receiving a CRC diagnosis than a sustained control strategy. The adjusted CIFs are given in Figure 4.14.

In both age groups, the effect estimates for a point exposure without artificial censoring indicated a stronger protective effect than the estimates regarding a sustained control strategy. The cumulative incidence throughout follow-up was higher for the point exposure control strategy as compared to the sustained control strategy. This is likely due to the initial increase in incidence due to the detection of cases at screening colonoscopies during follow-up, which are eliminated per design in the sustained control strategy.

The adjusted RR among individuals aged 55 to 64 was 0.84 when restricting the follow-up to 10 years. In Bretthauer et al. [2022], the length of available follow-up was 10 years, the study population was aged 55 to 64 at baseline and the control strategy was free of contamination by screening colonoscopy, because no screening colonoscopy was offered to the wider public in the countries included in the study (Poland, Norway, and Sweden). In their adjusted per-protocol analysis, the authors reported a RR of 0.69, i.e. the 10-year protective effect was stronger as indicated by the above results. This difference may be explained by a different background prevalence of CRC: The initial increase in cumulative incidence due to screen-detected CRC in the screened group was much smaller in Bretthauer et al. [2022] than in the results reported in Figure 4.14, indicating that the background prevalence at baseline was smaller in the population studied by Bretthauer et al. [2022]. With lower prevalence of CRC, the increase of the cumulative incidence due to screen-detected cancers is smaller and the long-term preventive effect of screening colonoscopy via removal of precursors is more prominent.

## 4.6 Bias due to non-alignment at time zero

While no RCT evidence is currently available regarding the site-specific (distal versus proximal colon) effectiveness of screening colonoscopy, several observational studies have reported results indicating that screening colonoscopy is much more effective in preventing CRC in the distal colon than it is in the proximal colon [Baxter et al., 2009, 2012; Brenner et al., 2011, 2014b; Doubeni et al., 2013; Guo et al., 2021; Kahi et al., 2018; Mulder et al., 2010; Nishihara et al., 2013]. This is in contrast to the findings reported in [Braitmaier et al., 2022b], where a TTE design was used and no clinically meaningful differences in effectiveness between site of the tumor were observed. In contrast to other study designs that do not prioritize causal interpretability, the TTE design of Braitmaier et al. [2022b] ensures alignment of key study design elements at time zero. This means that eligibility assessment uses only information from before time zero, exposure

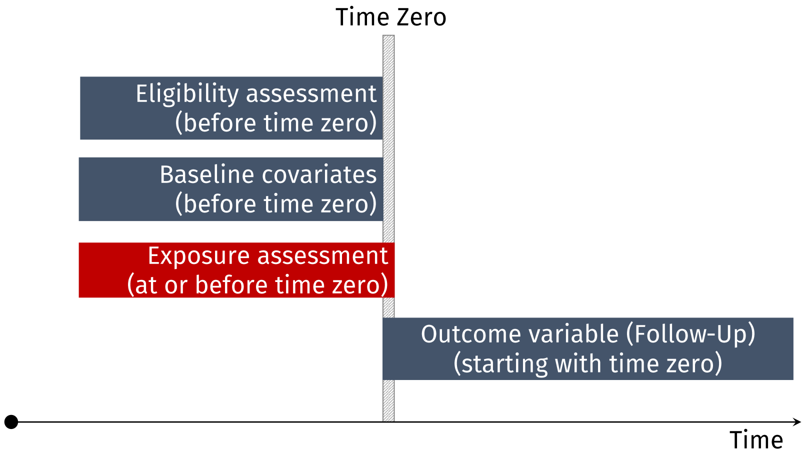


**Figure 4.15:** Misalignment at time zero for modified study on screening colonoscopy

assessment uses only information from time zero and outcome assessment uses only information starting with time zero, as is illustrated in Figure 3.1 in Chapter 3. The current section gives some methodological considerations regarding bias due to non-alignment at time zero in the context of screening colonoscopy and site-specific effectiveness, while empirical results are restricted to the corresponding publication (Braitmaier et al. [2024], see Chapter 7.6).

A study design as seen in previously published observational studies was used to investigate whether non-alignment at time zero could have caused bias that affected site-specific estimates differently. A hypothetical cohort design was used, corresponding to the setting of an observational cohort being recruited at a given point in time. At the baseline examination of such a cohort study, participants would be asked whether they had ever undergone screening colonoscopy and whether they ever received a CRC diagnosis. When interest lies in the effect of screening colonoscopy on CRC incidence, participants reporting past CRC diagnoses would be excluded from the analysis. The outcome of interest would then be any CRC diagnosis during follow-up. Such a study design was applied to the same data source used in [Braitmaier et al., 2022b], with some key design differences: Baseline was defined as a fixed time point (beginning of 2009). Exposure was defined as a coded screening colonoscopy during the baseline quarter or ever before. The resulting violation of alignment at time zero is illustrated in Figure 4.15. Importantly, the age structure at baseline still corresponds to a group in which distal CRC is much more common than proximal CRC, with proximal CRC becoming more relevant later during follow-up and at higher ages.

To illustrate such a study design, the following notation is introduced: Assume that time $t$ is split into three time windows, with $t = -1$ corresponding to the pre-baseline period,

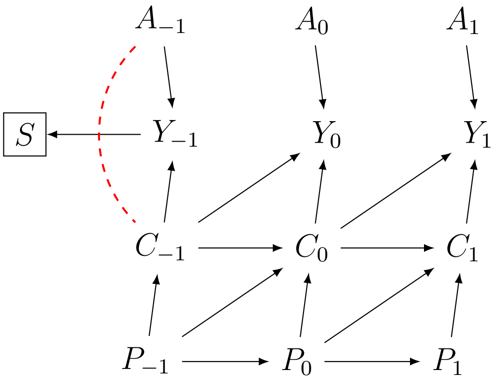


**Figure 4.16:** DAG of analysis of screening colonoscopy with time zero violation under a hypothetical scenario in which no directed path leads from past exposure $A_{-1}$ to subsequent outcomes $Y_0$ or $Y_1$

$t = 0$ corresponding to time zero and $t = 1$ corresponding to the post-baseline period. $A_t$ describes exposure to screening colonoscopy at time $t$, while $Y_t$ is a binary outcome indicator for time $t$, which is 1 if a CRC diagnosis occurred in time window $t$. The variable $S$ describes selection into the study cohort. Figure 4.16 gives a graphical representation of the causal relationships under such a design, with $P_t$ indicating the presence of cancer precursors at time $t$ and $C_t$ indicating the presence of undiagnosed/latent CRC. Since confounding is not pertinent to the discussion of violations of alignment at time zero, confounders $X$ are omitted from the graph. For illustration purposes, Figure 4.16 assumes a null-effect of exposure on subsequent outcomes, i.e. all arrows leading from exposure $A_k$ to the outcome at a later time, $Y_{t>k}$, are absent.

There is a violation of alignment at time zero in the hypothetical cohort design described above: Exposure definition uses information from time zero and before, instead of time zero alone. With this, exposure may precede exclusion criteria. Furthermore, exposure is ill-defined in that it does not correspond to a quantity that can be intervened upon, since past exposure cannot be changed. While a TTE as described in Braitmaier et al. [2022b] assesses the effect of $A_0$ on $\{Y_0, Y_1\}$, the flawed study design with time zero violation would instead attempt to assess the effect of $A_{\text{no alignment}} = \{A_{-1} = 1 \text{ or } A_0 = 1\}$ on $\{Y_0, Y_1\}$.

As shown in Figure 4.16, selection in the cohort design without alignment at time zero is based on $Y_{-1}$ as individuals with past CRC diagnosis are excluded. Given that exposure to screening colonoscopy will lead to CRC diagnosis when latent CRC is present, CRC diagnosis at time $-1$ is a collider on the path $A_{-1} \rightarrow Y_{-1} \leftarrow C_{-1}$. Controlling for a collider or its descendants will introduce a non-causal association between the two parent nodes, i.e. it will lead to collider bias [Pearl, 1995; Greenland, 2003]. With this, there are

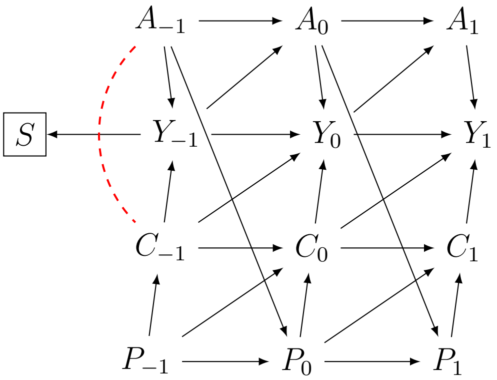


**Figure 4.17:** DAG of analysis of screening colonoscopy with time zero violation, with selection based on past outcome but not past exposure

now open non-causal paths from past exposure to later outcome. It is, therefore, evident from the DAG that the study design without alignment at time zero will yield biased results if $\mathbb{P}\left[Y_{-1}\right] > 0$.

To illustrate this further, consider the modified DAG in Figure 4.17: In this scenario, causal paths leading from exposure at time $k$ to subsequent outcomes $Y_{t>k}$ exist, due to the removal of precursor stages at the screening colonoscopy. Now, any analysis using the faulty study design will report an effect estimate that is a mixture of the true effect of exposure on the outcome and bias introduced by the study design.

To explain why the bias described above affects effect estimates of distal CRC more severely, it is important to consider the age structure under study. With the age at baseline being between 55 and 69, distal CRC is much more frequent than proximal CRC. Conceptually, if proximal CRC were to not occur at all before baseline, i.e. if Figure 4.17 were to reflect only proximal CRC and $\mathbb{P}\left[C_{-1} = 1\right] = \mathbb{P}\left[Y_{-1} = 1\right] = 0$, then no selection would take place and no association between $A_{-1}$ and $C_{-1}$ would be introduced. More generally, in the age group under study it is known that $\mathbb{P}\left[C_{-1}^{\text{distal}} = 1\right] > \mathbb{P}\left[C_{-1}^{\text{proximal}} = 1\right]$. The described collider bias will therefore be more severe for distal CRC.

When analyzing the same data source used in Braitmaier et al. [2022b], but with the study design without alignment at time zero, the results from previous observational studies indicating a stronger effect in the distal colon could be reproduced. More details on the study design and the empirical results are given in section 7.6.

## 4.7 Extending the set-up to more than two exposure strategies

So far, exposure was treated as a binary variable $A$, indicating either participation in screening colonoscopy at baseline or no participation. In an extension of the original study, exposure to screening colonoscopy was further subdivided into two categories depending on the screening physician's polyp detection record.

While adenoma detection rate is widely accepted as a quality marker of gastroenterologists conducting screening colonoscopy [Kaminski et al., 2017], this measure is not directly available in health claims data. Instead, polyp detection rate (PDR) was used to classify screening colonoscopies as high or low quality. PDR has been shown to be a close match to polyp detection rate [Schwarz et al., 2023].

With the above categorization, exposure strategies $Q \in \{0, 1, 2\}$ are then expressed by three levels: 0 = No screening colonoscopy at baseline, 1 = Low quality screening colonoscopy at baseline and 2 = High quality screening colonoscopy at baseline. Previously, a single logistic regression model was used for estimating propensity scores. Now, separate logistic models were fitted for each exposure strategy $q$, where a dummy exposure variable was defined as $A_q = 1$ if $Q = q$ and as $A_q = 0$ if $Q \neq q$. The same covariates were used for adjustment with identical parameterization, namely main effects modelling.

Based on these strategy-specific models, person-trial specific stabilized weights for strategy $q$ are given by

$$sw_{q,j} = \frac{\mathbb{P}\left[A_q = 1\right]}{\mathbb{P}\left[A_q = 1|X\right]} \tag{4.21}$$

for person-trials under exposure strategy $q$. Propensity models were fitted for all three exposure strategies separately. When using identical parameterizations, this approach is equivalent to multinomial regression. Weights were truncated at the 99th percentile of the combined weight distribution. Cumulative incidence functions were, as described in sections 4.1 and 4.2, estimated via pooled logistic regression models.

Covariate balance was assessed for each pairwise group comparison, using the absolute standardized mean difference. Again, a value of 0.1 or below was defined as sufficiently balanced.

The results and discussion are given in the publication [Schwarz et al., 2024], which is printed below in Section 7.5.

# CHAPTER 5

# Contributions to the field of mammography screening

Some 30 to 40 years ago, RCTs demonstrated a roughly 20 % reduction in breast cancer-related mortality due to mammography screening [Nelson et al., 2016]. However, treatment options have improved since then [Guarneri and Conte, 2004; Jansen et al., 2020]. Furthermore, no RCT was conducted in the German population. Effectiveness in the modern-day German population is, therefore, subject to debate.

An invitation-based mammography screening program was introduced in Germany starting in 2005 and reaching nation-wide coverage in 2009. German law requires that any medical screening tool that entails exposure to radiation must be safe and effective (see §84 of the German Radiation Protection Law). However, conducting an RCT where mammography screening is withheld from one study arm would be unethical, given that mammography is an established screening tool with some evidence supporting its efficacy. While mammography screening is known to have some harmful effects [Løberg et al., 2015], benefits are generally assumed to outweigh risks when assuming that previously reported reductions of breast cancer mortality by around 20 % can be relied upon [Lauby-Secretan et al., 2015; Marmot et al., 2013]. However, uncertainty regarding effectiveness in a modern day population persists [Biller-Andorno and Jüni, 2014]. One work package of the current thesis, therefore, consisted in developing an observational study design to evaluate the effectiveness of the German mammography screening program to reduce breast cancer-related mortality, as is described in Braitmaier et al. [2022a]. This effort was commissioned and funded by the Federal Office for Radiation Protection (see funding statement in Braitmaier et al. [2022a]).

Mammography screening is offered to women from the age of 50 to 69 in Germany (the age range was extended in 2024, so that women are eligible until the age of 75 [g-ba.de, 2024]). Eligible women are invited every two years, i.e. participation is sustained over time and a variety of screening trajectories are possible. Three strategies in particular will be compared: $Q = 0$: Never undergo screening, $Q = 1$: Undergo screening at least at baseline, possibly with further screening participation during follow-up, $Q = 2$: Undergo screening at baseline and thereafter at regular two-year intervals (plus a grace period of half a year), unless aged 70 or diagnosed with breast cancer. Only the pairwise comparisons with the never-screen strategy are of interest. The target trial protocol, its emulation and details regarding the statistical analysis are given in Braitmaier et al. [2022a]. In the following text, the issue of residual immortal time bias due to discretization of time will be discussed in the context of an extensive simulation study. The study design of the TTE itself is given in the corresponding publication (see Braitmaier et al. [2022a], which is printed in Section 7.4).

## 5.1 Simulation study: Discrete time, emulated target trials and residual immortal time bias

As discussed in Braitmaier et al. [2022a], some of the information in the health insurance claims database underlying the analysis is only available on a quarterly basis. As a consequence, sequential trials were emulated per calendar quarter and time was discretized to quarter years. When assignment to strategies is based on observed screening behavior during the first discrete time interval, i.e. when women who underwent screening in the first interval are assigned to the active screening strategies and all others to the comparator strategy, residual immortal time potentially remains within the first time interval; women who undergo screening at the end of the first quarter cannot possibly have died in the beginning of the quarter and, conversely, women who died at the beginning of the quarter will have had little chance to undergo screening and accumulate in the comparator strategy.

Avoiding self-inflicted biases is one of the main advantages of the TTE framework in comparison to other study designs in observational research [Hernán et al., 2016]. Nonetheless, residual immortal time may remain when granularity of discrete time is coarse, whereas an infinitesimally fine granularity of discrete time would approximately eliminate time-related biases. However, statistical methods such as pooled logistic regression, parametric g-formula and bootstrapping commonly used in the target trial literature in combination with the large data sources often employed, result in computationally heavy analyses and long run times. Coarser discretization of time could be useful in some settings to reduce the computational burden.

A simulation study was conducted to assess the impact of different granularity levels of discrete time in the setting of mammography screening and breast cancer related mortality. Values for random variables were taken from the published literature whenever possible. A hypothetical, population-based cohort study in which $n$ women were recruited simultaneously into the study was simulated. While real-life mammography screening differs in some aspects from this simulated study, simplifications were made whenever helpful to focus on the issue of residual immortal time bias. A study in which women included in the study were offered a once-only mammography screening during the first two years of the study period was simulated. While in reality, mammography screening is offered regularly every two years, this simplification was made to reduce complexity. The exact data generating mechanisms for all variables are described below. The general rationale behind some variables is given here: The sojourn period, i.e. the time between asymptomatic disease onset and development of symptoms, is the window of opportunity for the effect of mammography screening, which can only be effective by causing treatment at an earlier stage of the disease, thereby improving survival. In the simulation study, any potential effect could only occur if screening took place during this asymptomatic phase of the disease. Death due to breast cancer and the competing event of death due to other causes were observed during follow-up. This simulated cohort was then used to emulate target trials under various simulation scenarios and using varying granularity of discrete time. While the effect of screening on the outcome of interest was set to be null in most scenarios as to easily detect bias, the effect of screening on the competing event of death by other causes was set to be non-null in one scenario, corresponding to the possibility that mammography screening may cause death in some instances (e.g. complications after unnecessary treatment of overdiagnosed cases).

### 5.1.1 Causal structure of cohort data

The data generating mechanism of this simulation study is depicted in Figure 5.1. The DAG is a simplified version of the true causal mechanism, because it discards temporal effects. The binary exposure $A$ can only have an effect on the outcome $Y$, if it occurs during the time of asymptomatic disease $B_{asymp}$. Mammography screening does not affect survival directly, but it aids in the early detection of breast cancer and therefore leads to earlier treatment initiation, which in turn improves survival. If the disease has already progressed to the symptomatic stage $B_{symp}$, there is no effect of $A$ on $Y$ in this simulation, because the opportunity of earlier treatment initiation has passed. Similarly, treatment of overdiagnosed cases is only possible in cancers that did not present clinical symptoms

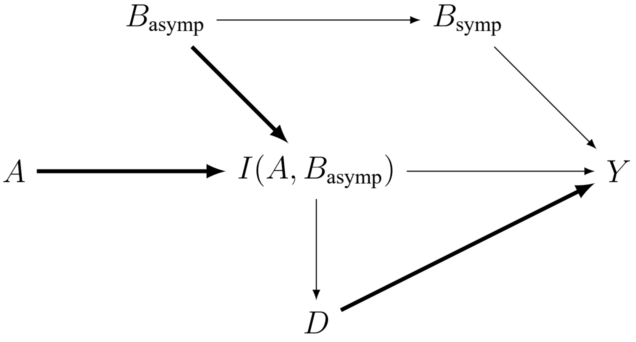


**Figure 5.1:** Simplified data generating mechanism of simulation study; thick arrows describe deterministic relationships

yet, with overdiagnosis being defined as the diagnosis of a disease which would not have become clinically apparent during the lifetime of the affected individual. Any effect of screening on the competing event death by causes other than breast cancer ($D$), e.g. mortality due to treatment complications of overdiagnosed cases, is only possible if screening takes place during the asymptomatic disease stage.

There is no confounding in the current simulation setup. This simplification was made as to focus on the effect of immortal time bias due to deaths in the first discrete time interval of each emulated trial.

### 5.1.2 Simulation of cohort data

The current simulation is complex in that it involves temporal aspects and feedback loops. The following sequence of steps was applied to simulate the cohort data underlying the emulated target trial:

1. Latent variables corresponding to a world free of mammography screening and latent screening participation
2. Realized disease and screening values depending on temporal order of terminal events (i.e. death from a world free of screening may prevent screening participation).
3. Latent variables in the presence of screening
4. Realized values of terminal events depending on temporal order
5. Additional scenario with early and sudden outcome events

More detail about the individual simulation steps is given in the following.

**Step 1: Latent variables in the absence of screening and latent screening participation**

The proportions for simulation variables are taken, as far as possible, from published research. Table 5.1 gives an overview of the simulated latent variables and references, when applicable.

**Step 2: Realized variables in the absence of screening and realized screening**

The realized variables in the absence of screening and the realized screening variable are determined based on the temporal ordering of terminal events. The realized variable on asymptomatic disease onset was defined as

$$B_{asymp} = \begin{cases} 0, & \text{if } T_{latD^{A=0}} < T_{latB_{asymp}} \\ latB_{asymp}, & \text{otherwise} \end{cases}$$

Similarly, the realized variable on symptom onset was defined as

$$B_{symp} = \begin{cases} 0, & \text{if } T_{latD^{A=0}} < T_{latB_{symp}} \\ latB_{symp}, & \text{otherwise} \end{cases}$$

The realized variable on screening participation was defined as

$$A = \begin{cases} 1, & \text{if } latA = 1 \text{ \& } T_{latD^{A=0}} \geq T_{latA} \text{ \& } T_{Y^{A=0}} \geq T_{latA} \\ 0, & \text{otherwise} \end{cases}$$

**Step 3: Latent variables in the presence of screening**

Non-breast cancer mortality caused by treatment was simulated as a latent variable as

$$latD_{overtreated} = \begin{cases} Bin(1, p_{lethal\ treatment}), & \text{if } B_{asymp} = 1 \text{ \& } T_{latA} \geq T_{latB_{asymp}} \text{ \&} \\ & T_{latA} < T_{latA_{symp}} \\ 0, & \text{otherwise} \end{cases}$$

The proportion of treated asymptomatic cases who died because of treatment is assumed to be zero in most simulation scenarios ($p_{lethal\ treatment} = 0$), but is set to a non-zero value (i.e. 1 %) in one simulation scenario to reflect some aspects of the ongoing debate surrounding harms of (over-)treatment (see e.g. Arrospide et al. [2015]; Baum [2013];

**Table 5.1:** Overview of latent variables underlying the simulated cohort

| Variable | Explanation | Reference |
|---|---|---|
| $latB_{asymp} \sim Bin(1, 0.13)$ | Latent asymptomatic disease (assuming lifetime risk of developing breast cancer of 13 %) | cancer.gov [2022] |
| $latB_{symp} \sim Bin(1, 0.78)$, if $latB_{asymp} = 1$ | Latent symptom onset (assuming that 22 % of breast cancers regress naturally) | Zahl et al. [2008] |
| $latD^{A=0} \sim Bin(1, 0.1)$ | Latent death by other causes in the absence of screening, assuming that 10 % of the study population would die from other causes during the study period | |
| $latY^{A=0} \sim Bin(1, 0.25)$, if $latB_{symp} = 1$ | Latent death due to breast cancer in the absence of screening, among individuals who developed symptomatic breast cancer and assuming lethality of 25 % | Narod et al. [2018] |
| $T_{latB_{asymp}} \sim \Gamma(\alpha = 1, \beta = \frac{1}{4}) - 1$ | Time to asymptomatic disease onset in years, assuming a mean of four years and then shifting the distribution to the left by one year, so that a portion of women enter the cohort with pre-clinical breast cancer present. | |
| $T_{Y^{A=0}} = \Gamma(\alpha = 1, \beta = \frac{1}{\text{sojourn time} + t_{lethal}})$ | Time to death due to breast cancer, where the duration from disease onset to symptom onset is defined by the sojourn time and assuming that death due to breast cancer on average occurs $t_{lethal}$ years after symptom onset. The sojourn time is assumed to be 7 years on average according to literature reports, but will be varied across scenarios. $t_{lethal}$ is assumed to be 5 years, but different values will be applied. | Weedon-Fekjær et al. [2005]; Narod et al. [2018] |
| $T_{latB_{symp}} = T_{Y^{A=0}} * z$ | Time to symptom onset. $z$ is a random number drawn from a truncated normal distribution bounded between 0 and 1 and with standard deviation 0.1 and a mean of $\frac{\text{sojourn time}}{\text{sojourn time} + t_{lethal}}$ | |
| $T_{latD^{A=0}} \sim \Gamma(\alpha = 1, \beta = \frac{1}{8})$ | Time to latent death by other causes in the absence of screening, assuming mean time of eight years | |
| $latA \sim Bin(1, 0.83)$ | Latent screening participation, assuming that 83 % of women would be willing to participate | Schmuker and Zok [2019] |
| $T_{latA} \sim Unif(0, 2)$ | Time to screening participation in years | |

Løberg et al. [2015]). Death due to complications from treatment of clinically irrelevant cancers is the most severe harm of mammography screening. Other major harmful effects, such as the impact of a cancer diagnosis and subsequent treatment on quality of life or psychological well-being, cannot be quantified in the context of this simulation.

Latent death due to causes other than breast cancer in the presence of screening was defined as

$$latD^{A=1} = \begin{cases} 1, & \text{if } latD^{A=0} = 1 \text{ or } latD_{overtreated} = 1 \\ 0, & \text{otherwise} \end{cases}$$

The time to this latent death due to other causes in the presence of screening was defined as

$$T_{latD^{A=1}} = \begin{cases} T_{latD^{A=0}}, & \text{if } latD^{A=0} = 1 \\ T_{latA} + \Gamma(\alpha = 1, \beta = 0.25), & \text{otherwise} \end{cases}$$

The event of interest, breast cancer mortality, in the presence of screening was modeled as

$$latY^{A=1} = \begin{cases} Bin(1, 1 - p_{screen\ effect}), & \text{if } latY^{A=0} = 1 \text{ \& } T_{latA} \geq T_{B_{asymp}} \text{ \&} \\ & T_{latA} < T_{B_{symp}} \\ 0, & \text{otherwise} \end{cases}$$

In the above definition, $p_{screen\ effect}$ describes the proportion among women whose screening takes place during the asymptomatic stage of the disease, who would have died from breast cancer in the absence of screening and whose death is avoided by screening. A null-effect was simulated by setting $p_{screen\ effect} = 0$ to easily identify bias. An additional scenario with a non-null treatment effect was carried out by setting $p_{screen\ effect} = 1$, i.e. assuming that all breast cancer deaths are preventable, if they are detected at the asymptomatic disease stage.

### Step 4: Realized values of terminal events

The final step in simulating the cohort data is to determine the realized values of all variables under the observed screening exposure. The realized competing event given the realized screening value was determined as

$$D = \begin{cases} 0, & \text{if } A = 0 \text{ \& } latD^{A=0} = 0 \\ 0, & \text{if } A = 0 \text{ \& } latD^{A=0} = 1 \text{ \& } T_{latD^{A=0}} \geq T_{latY^{A=0}} \\ 1, & \text{if } A = 0 \text{ \& } latD^{A=0} = 1 \text{ \& } T_{latD^{A=0}} < T_{latY^{A=0}} \\ 0, & \text{if } A = 1 \text{ \& } latD^{A=1} = 0 \\ 0, & \text{if } A = 1 \text{ \& } latD^{A=1} = 1 \text{ \& } T_{latD^{A=1}} \geq T_{latY^{A=1}} \\ 1, & \text{if } A = 1 \text{ \& } latD^{A=1} = 1 \text{ \& } T_{latD^{A=1}} < T_{latY^{A=1}} \end{cases}$$

Likewise, the realized event of interest given the realized screening value was determined as

$$Y = \begin{cases} 0, & \text{if } A = 0 \text{ \& } latY^{A=0} = 0 \\ 0, & \text{if } A = 0 \text{ \& } latY^{A=0} = 1 \text{ \& } T_{latY^{A=0}} \geq T_{latD^{A=0}} \\ 1, & \text{if } A = 0 \text{ \& } latY^{A=0} = 1 \text{ \& } T_{latY^{A=0}} < T_{latD^{A=0}} \\ 0, & \text{if } A = 1 \text{ \& } latY^{A=1} = 0 \\ 0, & \text{if } A = 1 \text{ \& } latY^{A=1} = 1 \text{ \& } T_{latY^{A=1}} \geq T_{latD^{A=1}} \\ 1, & \text{if } A = 1 \text{ \& } latY^{A=1} = 1 \text{ \& } T_{latY^{A=1}} < T_{latD^{A=1}} \end{cases}$$

**Step 5: Additional scenario with early and sudden outcome events**

After simulating cohort data as described above, an additional setting in which immortal time bias may play a particularly important role was simulated. For this, additional outcome events were simulated during the first two years, i.e. in the period during which screening is offered. In this scenario, additional early breast cancer deaths were simulated from a binomial distribution, independent of any other variables. The probability of early breast cancer death was set to $1.3\%$, i.e. at $10\%$ of the lifetime risk of breast cancer. If early breast cancer death occurred before screening exposure or any terminal event, these variables were reset to zero.

### 5.1.3 Emulating target trials based on the cohort data

Once person-level cohort data was simulated via the above-described process, target trials were emulated to examine the effect of undergoing screening ($A = 1$) in the first time interval after baseline on the time to death due to breast cancer. Analyses for both direct (censoring for competing death) and total (not censoring for competing death) effect were conducted. To obtain discrete time, a discretization function $d(t, l)$ was used, where $t$ is

**Table 5.2:** Overview of simulation scenarios

| Key | $p_{screen\ effect}$ | $p_{lethal\ treatment}$ | sojourn time | $t_{lethal}$ | early outcomes |
|---|---|---|---|---|---|
| S1 | 0 | 0 | 7.00 | 5.00 | no |
| S2 | 0 | 0.01 | 7.00 | 5.00 | no |
| S3 | 0 | 0 | 3.50 | 2.50 | no |
| S4 | 0 | 0 | 1.75 | 1.25 | no |
| S5 | 0 | 0 | 0.70 | 0.50 | no |
| S6 | 0 | 0 | 7.00 | 5.00 | yes |
| S7 | 0 | 0 | 3.50 | 5.00 | no |
| S8 | 0 | 0 | 1.75 | 5.00 | no |
| S9 | 0 | 0 | 0.70 | 5.00 | no |
| S10 | 0 | 0 | 7.00 | 2.50 | no |
| S11 | 0 | 0 | 7.00 | 1.25 | no |
| S12 | 0 | 0 | 7.00 | 0.50 | no |
| S13 | 1 | 0 | 7.00 | 5.00 | no |

continuous time and $l$ is the length of discrete time units in days. Emulated trials were conducted using discretization values $l = 7, 30, 91, 182, 365$. There was one emulated trial per discrete time unit in the first 2 years after cohort start, since screening was only offered during this time in the hypothetical study underlying the simulation. Individuals were assigned to the exposure strategy (i.e. screening at baseline), if they participated in screening during the first discrete time interval and to the control strategy (no screening at baseline) otherwise. Women who underwent screening or received a breast cancer diagnosis before baseline were not included in the respective emulated trial.

An intention-to-screen analysis would yield highly conservative effect estimates due to strong contamination of the control strategy, since the participation rate in mammographic screening is high and a large proportion in the control strategy undergo screening during follow-up. Therefore, person-trials in the control strategy were artificially censored when they participated in screening during follow-up, i.e. a per-protocol effect was estimated.

The effect of screening was assessed non-parametrically using event-specific cumulative incidence functions. The effect of interest was expressed as the relative risk at the end of follow-up.

### 5.1.4 Simulation scenarios

For each simulation scenario, a cohort of size n = 100,000 individuals was simulated in 500 simulation runs. Results were averaged. The scenarios are given in Table 5.2.

Scenarios S1 and S2 can be regarded as realistic based on background knowledge from

**Table 5.3:** Proportions of analysis variables in the simulated cohort, mean over all simulation runs

| Key | $A$ | $B_{asymp}$ | $B_{symp}$ | $D$ | $Y$ |
|---|---|---|---|---|---|
| S1 | 0.82 | 0.13 | 0.098 | 0.10 | 0.024 |
| S2 | 0.82 | 0.13 | 0.098 | 0.10 | 0.024 |
| S3 | 0.82 | 0.13 | 0.099 | 0.10 | 0.024 |
| S4 | 0.82 | 0.13 | 0.100 | 0.10 | 0.025 |
| S5 | 0.82 | 0.13 | 0.100 | 0.10 | 0.025 |
| S6 | 0.79 | 0.13 | 0.098 | 0.09 | 0.105 |
| S7 | 0.82 | 0.13 | 0.099 | 0.10 | 0.024 |
| S8 | 0.82 | 0.13 | 0.100 | 0.10 | 0.024 |
| S9 | 0.82 | 0.13 | 0.101 | 0.10 | 0.025 |
| S10 | 0.82 | 0.13 | 0.098 | 0.10 | 0.024 |
| S11 | 0.82 | 0.13 | 0.098 | 0.10 | 0.024 |
| S12 | 0.82 | 0.13 | 0.098 | 0.10 | 0.024 |
| S13 | 0.82 | 0.13 | 0.098 | 0.10 | 0.018 |

the published literature (excepting the null-effect $p_{screen\ effect} = 0$).

Scenarios S3 - S5 are increasingly unrealistic in that the time from asymptomatic disease onset to development of symptoms and finally breast cancer related death are much shorter than what is expected for the majority of cases. Scenario S6 is unrealistic in that it assumes a large number of early and sudden breast cancer related deaths. Scenarios S7 - S12 are modifications of scenarios S3 - S5 where either only sojourn time or only $T_{\text{lethal}}$ are decreased.

Scenario S13 can be regarded as realistic when assuming a strong protective effect. Screening mammography aims to detect breast cancer early so that treatment can be initiated at a stage where prognosis is favorable. The scenario is slightly exaggerated in that it assumes 100% of breast cancers detected at the asymptomatic stage can be treated successfully and do not result in a breast cancer related death, whereas the real 5-year survival rate of localized breast cancer is 99.3%, i.e. slightly below 100% [National Cancer Institute, 2024].

### 5.1.5 Results & discussion of simulation study

Realized proportions of the analysis variables in the simulated cohort underlying the emulated trials were checked and means were calculated across all simulation runs. The resulting mean proportions are given in Table 5.3. Results of the simulation are given as relative risks at the end of follow-up and are summarized (mean over all simulation runs) in Figure 5.2.

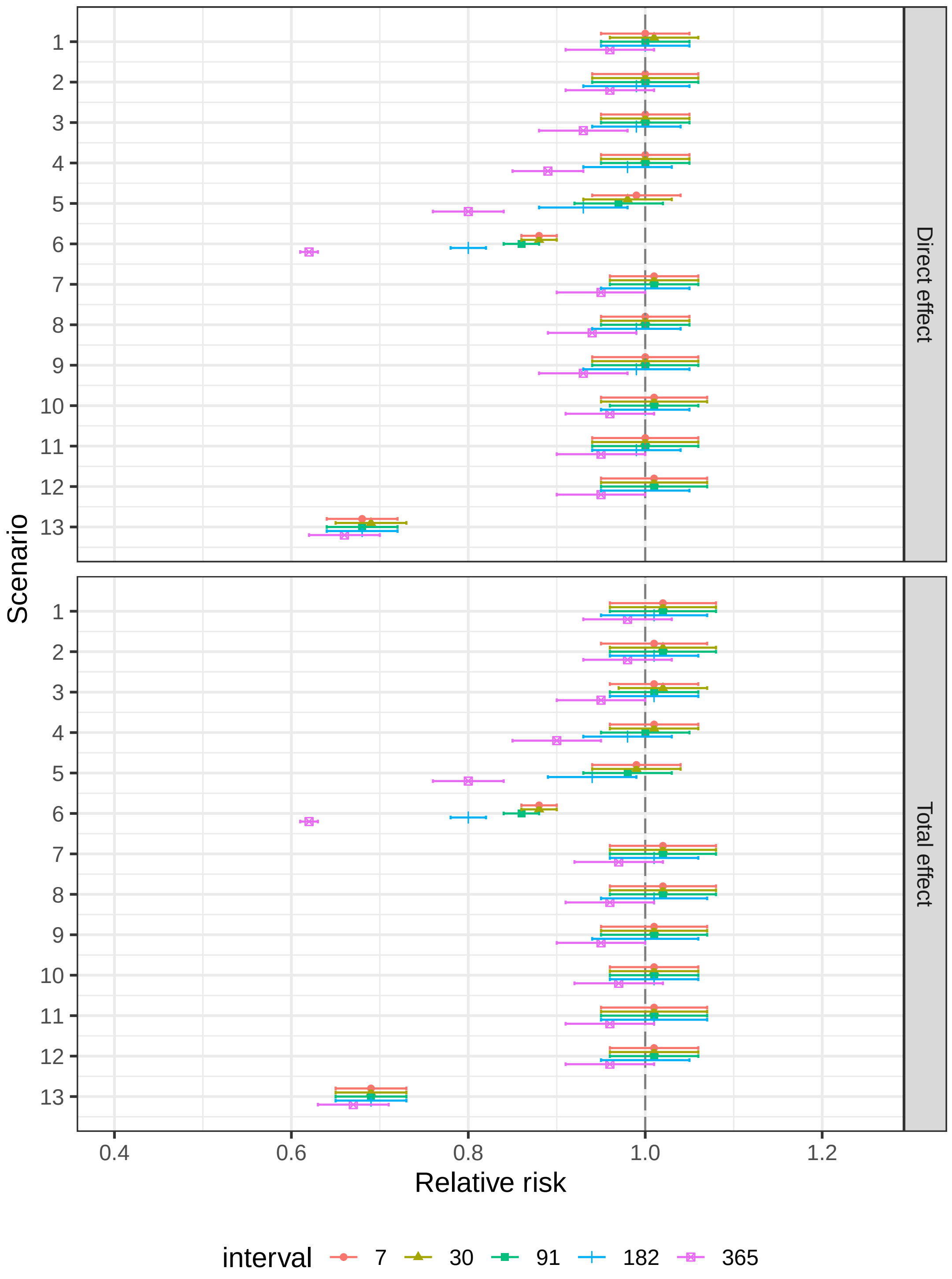


**Figure 5.2:** Results of simulation runs. Error bars represent ± standard error. The vertical, dashed line indicates the null-effect.

The presence of competing events complicates interpretation of the results slightly: Immortal time bias due to the competing event may falsely reduce the point estimate for the effect of screening on the competing event, but it may increase the estimate for the event of interest – women who underwent screening in the first time interval cannot have died before and are still at risk of breast cancer death, while women dying from other causes early cannot undergo screening and also cannot die from breast cancer later on. Both direct and total effect of screening mammography on breast cancer mortality were estimated to untangle this issue. As shown in Figure 5.2, results differed only slightly between direct and total effect with slightly larger relative risks for the total effect. The following discussion focuses on the controlled direct effect for simplicity, but extends to the total effect implicitly.

Scenarios S1 and S2 were the most realistic scenarios when assuming a null-effect, with S2 including a harmful effect of screening relating to increased breast cancer mortality due to overdiagnosed cases dying from complications of unnecessary treatment. In both scenarios, notable immortal time bias occurred only when discrete time intervals of length 365 days were used. At finer granularity, no substantial immortal time bias was visible. This indicates that under realistic settings, residual immortal time bias within the first discrete time interval after time zero does not pose a threat in this application, unless discrete time intervals become very long. The harmful effect of screening assumed in scenario S2 did not result in visibly increased RRs. This is likely due to the small effect size of $p_{lethal\ treatment} = 0.01$.

In scenarios S3 - S5 both sojourn time and time from symptom onset to breast cancer death were gradually and simultaneously decreased. In scenarios S7 - S9 only sojourn time was decreased, while in scenarios S10 - S12 only time from symptom onset to breast cancer death was decreased. In scenarios S3 - S5, immortal time bias increased with decreasing sojourn time and time to breast cancer death. Furthermore, immortal time bias became more severe with increasing granularity of discrete time. Especially in scenario S5, immortal time bias increased with every increase in granularity of discrete time. In scenarios S7 - S12, i.e. when only sojourn time or only time to death decreased, immortal time bias also increased for the coarsest level of discrete time, but to a lesser extent. For finer granularities of time, results did not stray far from the true null-effect. This indicates that substantial bias can arise when the time from exposure to outcome becomes short compared to the length of discrete time intervals.

Scenario S6 was an extreme case analysis in which a large amount of outcome events was observed early and suddenly. This scenario was susceptible to immortal time bias, which increased in severity as granularity of discrete time coarsened. This indicates that if a substantial portion of outcome events are observed during early follow-up, emulated target trials become susceptible to immortal time bias, particularly at coarser discretization of time and especially when these early outcomes are spontaneous. It is noteworthy that scenario S6 was unrealistic in that it simulated a large number of early breast cancer deaths that were not preceded by a symptomatic disease phase. A study assessing the effectiveness of mammography screening should include average risk women while excluding those who already show symptoms of breast diseases and are at a higher risk of breast cancer related mortality soon after baseline. High risk women are not the target population of mammographic screening, but should receive diagnostic or curative medical attention instead. While scenario S6 was an exaggeration to illustrate the potential for residual immortal time bias, it does refer to the real phenomenon that especially among young women (e.g. aged 40 or under), breast cancers are more often progressing much faster than in older women, often are diagnosed at a more advanced stage and, subsequently, tend to have worse survival outcomes [Assi et al., 2013]. However, even in those cases it is unlikely that breast cancer deaths are not preceded by a symptomatic phase, which could be used to restrict eligibility to the study. If, however, insufficient information were available in the data to properly check eligibility, prevalent cases could mistakenly be included in the cohort, leading to biased estimation as seen in scenario S6. This scenario illustrates the importance of applying strict exclusion criteria, so that women with either prevalent or past breast cancer or with symptoms of breast cancer are excluded from the study population.

Scenario S13 corresponds to scenario S1, with the difference that a non-null effect was simulated. This was a positive control to check that the simulation set-up worked as intended. Indeed, the results for scenario S13 indicate a protective effect of mammography screening on breast cancer mortality. This scenario seemed susceptible to immortal time bias only when the coarsest level of discrete time was used, as all other effect estimates were clearly aligned and only the estimate for $l = 365$ was lower.

# CHAPTER 6

# Discussion & outlook

In this thesis, target trial emulation (TTE) was used to assess the site-specific effectiveness of screening colonoscopy [Braitmaier et al., 2022b]. Substantive sensitivity analyses were conducted, tailored to the research question, data source and analysis methods used and strengthened confidence in the obtained results. For instance, unobserved confounding is a potential risk to observational studies. A suitable negative control outcome was identified in Braitmaier et al. [2022b] – see Section 4.4.1 for a discussion of pancreatic cancer as negative control outcome for screening colonosopy – and no evidence of strong unobserved confounding could be found. Chapter 4 gives an in-depth description of all sensitivity analyses conducted, including a discussion regarding the interpretation of each. Extensions to the original study design were implemented after the initial results were published – see e.g. Schwarz et al. [2024] for a set up to contrast high and low quality colonoscopy.

Furthermore, non-alignment at time zero was identified as a source of design-induced bias in site-specific effect estimates reported in previous observational studies [Braitmaier et al., 2024]. While bias due to non-alignment at time zero was discussed as a source of bias in previous work [García-Albéniz et al., 2017b], Braitmaier et al. [2024] was the first to investigate how this bias affects site-specific estimates regarding effectiveness of screening colonoscopy. Even though there was consensus in the published literature regarding a higher effectiveness of screening colonoscopy in the distal colon, Braitmaier et al. [2024] demonstrated that this difference in site-specific effectiveness is mostly an artifact resulting from selection bias.

Next, the German-language overview paper by Braitmaier and Didelez [2022] serves as a low threshold entry point to target trial emulation, discussing its strengths and limitations. It will hopefully increase the uptake of TTE in Germany.

As part of this thesis, a detailed study protocol for the effectiveness evaluation of the German mammography screening program was developed and published in a peer-reviewed journal [Braitmaier et al., 2022a]. As pre-registration and peer-review of the study design is not customary in observational research, this is a contribution to increase transparency and reproducibility in observational studies.

For a discussion of each of the individual studies, the reader is referred to the discussion sections in the respective papers attached in Chapter 7.

In addition to the study protocol, the present thesis contains methodological work in the context of the emulated target trial on screening mammography. In particular, a substantive simulation study was conducted to assess residual immortal time bias at varying granularities of discrete time. No major residual immortal time bias was found when using realistic settings related to screening mammography, breast cancer mortality and discretization of time as seen in GePaRD. Furthermore, the simulation study may serve as a guide for target trial emulation with other data sources: Using coarser granularities of discrete time considerably lightens the computational burden. The simulation results are informative as to how coarse discrete time intervals may be made before residual immortal time bias becomes problematic.

In summary, the work presented in this thesis combined recent methodological advances of causal inference with a rich data source containing information on many clinical factors. The publications in Chapter 7 represent the first use of target trial emulation to evaluate cancer screening programs in Germany, providing valuable information to inform patients and policy makers.

## 6.1 Future perspectives

Several research questions remain and merit future work. First, the analyses described in the study protocol by Braitmaier et al. [2022a] were underway when this thesis was written. The results will contribute to decision making by health authorities regarding the future of the German mammography screening program and will be published in peer-reviewed journals to inform the public.

Further methodological work relating to the analyses described here may be conducted in the future. For instance, bootstrapping in the context of target trial emulation is computationally demanding when using large health claims datasets and methods such as pooled logistic regression or the parametric g-formula. Bootstrap samples need to be drawn from the underlying population of i.i.d. individuals to repeat the entire emulation and analysis process several hundred times. Alternative approaches – such as extensions of the wild bootstrap to account for correlated data – could be explored, either analytically or in simulation studies, with the goal of decreasing computation time while ensuring nominal coverage.

In the context of screening for colorectal cancer, several research questions remain open. A natural extension of Braitmaier et al. [2022b] would be to assess the effectiveness of not only screening colonoscopy, but also the alternative screening test for fecal occult blood, which is offered in Germany starting at age 50 and is repeated every year until age 55 and every other year thereafter. Furthermore, the effectiveness of screening colonoscopy was only assessed in the age group of 55 to 69, but not in older individuals. Longer lookback needs to be available to appropriately exclude individuals exposed to screening colonoscopy in the ten years before time zero. This is less of a concern in younger individuals, because screening colonoscopy was only available to the average risk population starting at age 55. When more data years become available, sufficient lookback might be achieved to assess the effectiveness also in older individuals. The presence of competing events and confounding between exposure and competing death, however, further complicate the assessment of effectiveness in an elderly, partly frail population. Further restriction criteria regarding frailty and end-of-life may be explored to mitigate these issues.

## 6.2 Conclusion

The work presented in this thesis demonstrated that health claims data can be used to reliably estimate the effect of cancer screening programs on cancer incidence, if appropriate study designs and methods are used. At the same time, potential for substantial, self-inflicted bias was found in the context of site-specific effectiveness of colorectal cancer screening, if alignment at time zero was violated. A clear and concise definition of the estimand and the target protocol along with its emulation is, therefore, invaluable for convincing causal analyses with real world data.

# CHAPTER 7

# Publications

## 7.1 Author contributions

The following Table 7.1 gives an overview of the author contributions of Malte Braitmaier for each paper contributing to this thesis.

**Table 7.1:** Author contributions to each publication contributing to this thesis

| Term | Paper 1 | Paper 2 | Paper 3 | Paper 4 | Paper 5 |
|---|---|---|---|---|---|
| | Section 7.2 | Section 7.3 | Section 7.4 | Section 7.5 | Section 7.6 |
| Conceptualization | Lead | Equal | Lead | Equal | Lead |
| Methodology | Lead | N/A | Lead | Equal | Lead |
| Software | Lead | N/A | Lead | Lead | Lead |
| Formal analysis | Lead | N/A | Lead | Lead | Lead |
| Investigation | Equal | N/A | Equal | Equal | Equal |
| Data curation | Lead | N/A | Lead | Lead | Lead |
| Writing - original draft | Lead | Lead | Lead | Supporting | Lead |
| Writing - review & editing | Equal | Equal | Equal | Equal | Equal |
| Visualization | Lead | Lead | Lead | Lead | Lead |

## 7.2 Paper 1: Screening colonoscopy similarly prevented distal and proximal colorectal cancer: a prospective study among 55-69-year-olds

This paper was published under a CC-BY open access license in the Journal of Clinical Epidemiology. For details on how to cite the paper, refer to https://doi.org/10.1016/j.jclinepi.2022.05.024

Journal of Clinical Epidemiology 149 (2022) 118−126

Journal of Clinical Epidemiology

ORIGINAL ARTICLE

# Screening colonoscopy similarly prevented distal and proximal colorectal cancer: a prospective study among 55−69-year-olds

Malte Braitmaier[a], Sarina Schwarz[b], Bianca Kollhorst[a], Carlo Senore[c], Vanessa Didelez[a,d], Ulrike Haug[b,e,*]

[a]*Department of Biometry and Data Management, Leibniz Institute for Prevention Research and Epidemiology − BIPS, Bremen, Germany*
[b]*Department of Clinical Epidemiology, Leibniz Institute for Prevention Research and Epidemiology − BIPS, Bremen, Germany*
[c]*Epidemiology and Screening Unit − CPO, University Hospital Città della Salute e della Scienza, Turin, Italy*
[d]*Faculty of Mathematics and Computer Science, University of Bremen, Bremen, Germany*
[e]*Faculty of Human and Health Sciences, University of Bremen, Bremen, Germany*

Accepted 30 May 2022; Published online 6 June 2022

## Abstract

**Objectives:** We aimed to evaluate the effectiveness of screening colonoscopy in reducing incidence of distal vs. proximal colorectal cancer (CRC) in persons aged 55−69 years.

**Study Design and Setting:** Using observational data from a German claims database (German Pharmacoepidemiological Research Database), we emulated a target trial with two arms: Colonoscopy screening vs. no-screening at baseline. Adjusted cumulative incidence of total, distal, and proximal CRC over 11 years of follow-up was estimated in 55−69-year-olds at an average CRC risk and without colonoscopy, polypectomy, or fecal occult blood test before baseline.

**Results:** Overall, 307,158 persons were included (screening arm: 198,389 and control arm: 117,399). The adjusted 11-year risk of any CRC was 1.62% in the screening group and 2.38% in the no-screening group resulting in a relative risk of 0.68 (95% CI: 0.63−0.73). The relative risk was 0.67 for distal CRC (95% CI: 0.62−0.73) and 0.70 (95% CI: 0.63−0.79) for proximal CRC. The cumulative incidence curves of the groups crossed after 6.7 (distal CRC) and 5.0 years (proximal CRC).

**Conclusion:** Our results suggest that colonoscopy is effective in preventing distal and proximal CRC. Unlike previous studies not using a target trial approach, we found no relevant difference in the effectiveness by location.

*Keywords:* Screening colonoscopy; Colorectal neoplasms; Observational study; Effectiveness; Target trial emulation; Proximal

## 1. Introduction

Colorectal cancer (CRC) is among the most common cancers and leading causes of cancer death worldwide [1]. An intention-to-screen meta-analysis of randomized controlled trials (RCTs) on screening with flexible sigmoidoscopy showed a reduction of CRC incidence by 18% and of CRC mortality by 28% [2]. The ongoing Nordic-

Funding: BIPS intramural funding.

Competing interests: The authors have no relevant financial or nonfinancial interests to disclose.

Ethics approval: In Germany, the utilization of data from health insurances for scientific research is regulated by the Code of Social Law. All involved health insurance providers and the responsible authorities approved the use of the health claims data for this study. The Ethics Committee of the University of Bremen determined that studies based on claims data are exempt from an institutional review board review.

Data transparency statement: In Germany, use of personal data is protected by the Federal Data Protection Act and particularly the use of claims data for research is regulated by the Code of Social Law. Researchers must apply for a project-specific permit from the statutory health insurance providers which then need an approval from their governing authorities. The use of the data on which this publication is based was only allowed for BIPS employees within the framework of the specified project and limited to a predefined time span. Researchers who want to access the data on which this publication is based need to ask for a new approval by the statutory health insurance providers DAK-Gesundheit (service@dak.de), die Techniker (service@tk.de), hkk Krankenkasse (info@hkk.de), and AOK Bremen/Bremerhaven (info@hb.aok.de) which upon granting approval will ask their respective authorities for approval. Please contact gepard@leibniz-bips.de for help with this process.

* Corresponding author. Department of Clinical Epidemiology, Leibniz Institute for Prevention Research and Epidemiology − BIPS, Achterstraße 30, 28359 Bremen, Germany. Tel.: +49 421 21856862; fax: +49 421 21856821.

*E-mail address:* haug@leibniz-bips.de (U. Haug).

https://doi.org/10.1016/j.jclinepi.2022.05.024

**What is new?**

**Key findings**

- Using a target trial approach that avoids self-inflicted biases, we found no notable difference in effectiveness between distal and proximal colon among persons aged 55–69 years at baseline.

**What this adds to what is known?**

- Previous observational studies suggested a lower effectiveness of screening colonoscopy to prevent colorectal cancer in the proximal than distal colon, but these studies might be compromised by self-inflicted selection bias.

**What is the implication, what should change now?**

- Our findings support the use of colonoscopy for screening purposes in persons aged 55–69 years to prevent both distal and proximal colorectal cancer.

European Initiative on Colorectal Cancer study, the only RCT assessing the effectiveness of screening colonoscopy compared to no screening, will provide key insights into the overall effect of colonoscopy on CRC incidence and mortality [3]. Nonetheless, despite its large sample size, it is not powered to investigate differences in the effect as per tumor location.

Observational studies suggested a markedly lower effectiveness of screening colonoscopy in reducing CRC incidence in the proximal vs. distal part of the colorectum [4–6]. For example, a recent cohort study by Guo et al. suggested an incidence reduction of 64% for distal and 31% for proximal CRC. However, validity of existing observational studies on this topic is questionable due to possibly self-inflicted biases introduced by the analytical approach. García-Albéniz et al. demonstrated how effects of screening colonoscopy on CRC incidence are overestimated when treatment/exposure assignment is done before baseline, whereas eligibility is assessed at baseline [7]. This overestimate may differentially affect CRCs in the distal vs. proximal colon. An accurate assessment of the difference of colonoscopy in reducing incidence in the distal vs. proximal colon is important, particularly for estimating the risk-benefit ratio of screening colonoscopy compared to the less invasive screening sigmoidoscopy.

As it seems unlikely that any RCT will be powered to clarify this question, observational studies on screening colonoscopy remain important to complement existing evidence. These studies should exploit large databases with sufficiently long follow-up. Furthermore, the observational data must be analyzed in a way that facilitates causal conclusions and avoids self-inflicted biases. The emulation of target trials is well recognized in this regard and was successfully applied by García-Albéniz et al. to estimate the effectiveness of screening colonoscopy in Medicare beneficiaries aged 70 years or older [8].

Extending this approach, we aimed at evaluating the effectiveness of screening colonoscopy in reducing incidence of distal vs. proximal CRC in persons aged 55–69 years using a large German claims database.

## 2. Methods

We emulated target trials comparing the strategies "screening colonoscopy at baseline" vs. "no screening at baseline", both with access to further screening and diagnostic colonoscopies during follow-up. Supplement Table S1 contains a summary of our target trial protocol and its emulation.

### 2.1. Data source and study population

We used the German Pharmacoepidemiological Research Database (GePaRD) which comprises claims data from four statutory health insurance providers in Germany and covers about 20% of the German population [9]. We used data from 2004 to 2017. Information on utilization of screening colonoscopy, offered in Germany to persons aged 55 years or older since 2002, is distinguishable from diagnostic colonoscopy. Supplement 4 provides details on GePaRD and the identification and classification of CRCs in GePaRD.

To be eligible, persons had to be aged 55–69 years at baseline, that is, at the start of the respective trial and had to be continuously insured for at least 3 years before baseline. As detailed in Figure 2 and Supplement 1, further inclusion and exclusion criteria were applied to focus on an average-risk population, corresponding to ongoing colonoscopy trials and prior target trials on colonoscopy [2,3,8].

### 2.2. Treatment arms and follow-up

The first quarter of 2007 was the baseline quarter of the first trial. In this quarter, we assessed eligibility criteria for all persons. The persons meeting the eligibility criteria were assigned to the screening arm if they underwent colonoscopy screening in the baseline quarter or to the no-screening arm otherwise. As previously described [8], this procedure was repeated for all quarters from 2007 to 2011, yielding 20 successive trials. Persons could be enrolled in more than one trial (Fig. 1). In particular, our sample consisted of $n_{unique}$ persons, some of which were included in more than one emulated trial, so that the final sample size (including nonunique persons) was $n \geq n_{unique}$. To reduce computational time, we used a 5% random sample of those in the no-screening arm (drawn at person level), which still yielded a high number of persons in this arm.

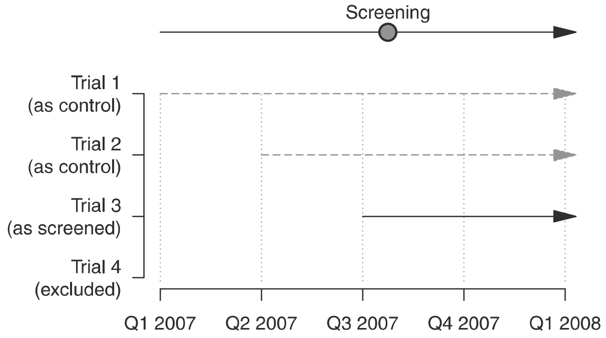


**Fig. 1.** Illustration of emulation of a series of target trials. The figure displays a hypothetical person who met all eligibility criteria at the start of 2007. This person was included in the emulated trial starting on January 1, 2007. One-quarter later, the same person was still eligible and was included in the emulated trial starting on April 1, 2007. The person was assigned to the control arm in both these trials, as no screening colonoscopy was observed during the respective baseline quarters. The person was still eligible for the emulated trial starting on July 1, 2007. However, the person was allocated to the screening arm as a screening colonoscopy was observed in the quarter following July 1, 2007. The person was not eligible for the trial starting on October 1, 2007 because a screening colonoscopy before this trial's baseline was observed. Data from all these trials were pooled and time since baseline (of the respective trial) was used in all time-to-event analyses.

Persons were followed up until the end of study period (end of 2017), end of continuous insurance coverage, death, or CRC diagnosis, whichever occurred first. The arms were defined as screening vs. no-screening in the respective baseline quarter regardless of screening behavior during the remaining follow-up. Persons were not censored from earlier trials once they changed strategy in subsequent trials. We chose this approach over imposing full adherence during follow-up by analysis because it avoids strong assumptions on time-varying confounding (details in Supplement 3).

### 2.3. Outcome and confounding variables

The outcome was the time until first diagnosis of CRC during follow-up. This was analyzed for any type of CRC and further analyzed separately for CRCs proximal and distal to the splenic flexure (no separate analysis for the category "both/unknown location" due to small numbers) (Supplement 4).

We adjusted for confounding baseline covariates using direct (e.g., age, gender, menopausal hormone therapy) or proxy information (e.g., use of preventive services) on relevant factors (Supplement 4).

### 2.4. Statistical analysis

We pooled persons across all emulated trials. The effect of interest was measured as contrast between cumulative incidence functions (CIF). We used pooled logistic regression to estimate a parametric version of the Aalen–Johansen estimator (details in Supplement 3). Adjustment for baseline confounding was achieved by inverse probability of treatment (i.e., propensity score) weighting. Covariate balance after weighting was examined using absolute standardized differences. Overlap of the propensity score distributions was assessed using histograms. Point wise, percentile-based 95% confidence intervals were obtained using a robust, person-level bootstrap with 250 iterations.

The above contrast of adjusted CIFs is known as total effect where death is not eliminated as competing event; in a sensitivity analysis, we also estimated the direct effect (i.e., censoring and thus hypothetically eliminating death), expecting no relevant difference between the two approaches in the age group of our study (details in Supplement 3).

Confounding variables were mostly operationalized as binary variables. Missing values for educational attainment were included as a distinct category. A negative control analysis with pancreatic cancer as outcome variable was conducted to assess residual confounding.

Supplement 3 contains a detailed description of the statistical methods. Data management and statistical analyses were done in SAS software version 9.4 (SAS Institute, North Carolina).

## 3. Results

Overall, 2,378,416 persons fulfilled all eligibility criteria. Of these, 198,389 persons were assigned to the screening colonoscopy arm. The random sample of controls assigned to the no-screening arm included 117,399 persons (1,247,913 nonunique persons, Fig. 2). Results reported below refer to nonunique persons and outcome events, that is, *n* always includes nonunique persons. Median follow-up was 8.3 years (interquartile range: 3.0).

About half of the study population was female with median age 60–62 years in both arms (Table 1). The proportion of persons with higher education was 20% in the screening and 15% in the control arm. The group differences in the prevalence of the further confounders were $\leq$3 percentage points, except for menopausal hormone therapy (23% among screened vs. 14% among controls) and uptake of at least one preventive service before baseline (85% among screened vs. 66% among controls). Covariate balance checks and propensity score overlap were satisfactory (Supplement Figures S4–S5).

We observed 2,540 incident CRCs in the screening and 21,973 in the control arm (Table 2). In men, the ratio of the number of distal to proximal CRCs was 2.7 in the screening (women: 1.5) and 2.5 in the control arm (women: 1.6). Figure 3 shows the adjusted CIF curves for any distal and proximal CRC. After the initial spike in cumulative CRC incidence in the screening arm (0.79%), the slope of the CIF curve remained lower than in the no-screening arm throughout follow-up. The CIF curves for any CRC of both arms crossed after 6 years. After 11 years, the adjusted risk

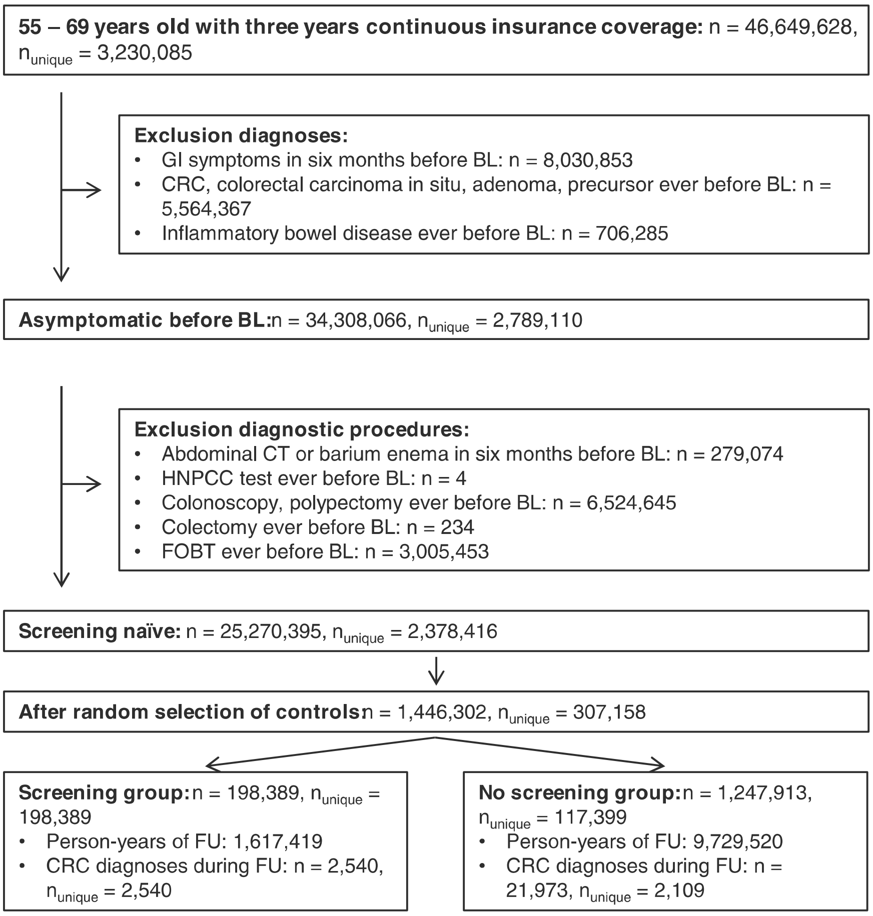

55 – 69 years old with three years continuous insurance coverage: n = 46,649,628, $n_{unique}$ = 3,230,085
Exclusion diagnoses:
• GI symptoms in six months before BL: n = 8,030,853
• CRC, colorectal carcinoma in situ, adenoma, precursor ever before BL: n = 5,564,367
• Inflammatory bowel disease ever before BL: n = 706,285
Asymptomatic before BL:n = 34,308,066, $n_{unique}$ = 2,789,110
Exclusion diagnostic procedures:
• Abdominal CT or barium enema in six months before BL: n = 279,074
• HNPCC test ever before BL: n = 4
• Colonoscopy, polypectomy ever before BL: n = 6,524,645
• Colectomy ever before BL: n = 234
• FOBT ever before BL: n = 3,005,453
Screening naïve: n = 25,270,395, $n_{unique}$ = 2,378,416
After random selection of controlsn = 1,446,302, $n_{unique}$ = 307,158
Screening group:n = 198,389, $n_{unique}$ = 198,389
• Person-years of FU: 1,617,419
• CRC diagnoses during FU: n = 2,540, $n_{unique}$ = 2,540
No screening group:n = 1,247,913, $n_{unique}$ = 117,399
• Person-years of FU: 9,729,520
• CRC diagnoses during FU: n = 21,973, $n_{unique}$ = 2,109

**Fig. 2.** Flow into study cohort of persons aged 55 to 69 years with at least 3 years continuous health insurance coverage prior to baseline (allowing for 15-day gaps in insurance coverage). GePaRD data from 2004 to 2017 were used, with emulated target trials in every calendar quarter from 2007 to 2011.

was 1.62% (1.54–1.68%) in the screening arm compared to 2.38% (2.26–2.51%) in the control arm (adjusted absolute risk difference: 0.77%, adjusted relative risk [aRR]: 0.68, Table 2). The overall pattern of the CIF curves for distal and proximal CRC was similar to any CRC. For proximal CRC, the curves crossed earlier (after 5.0 years) compared to distal CRC (after 6.7 years). After 11 years, the adjusted absolute risk difference for distal CRC was 0.47% and the aRR was 0.67. For proximal CRC, the adjusted absolute risk difference was 0.22% and the aRR was 0.70 (Table 2). Supplement 9 provides a comparison of adjusted and unadjusted CIF curves.

Supplement Table S2 provides characteristics of incident CRCs, by screening arm and site of CRC. It also shows that 4.2% of distal CRCs and 4.8% of proximal CRCs in the screening arm were included in at least one earlier emulated trial in the control group. Overall, 16.9% of controls were included in the screening arm of a later trial. Supplement 8 presents the results of additional analyses restricting to persons aged 55–64 years. Sensitivity analyses treating death as a censoring event, that is, estimating the direct instead of the total effect did not deviate substantially from the main results (Supplement 10). Supplement 6 displays the results of a negative control analysis using pancreatic cancer incidence as an outcome.

## 4. Discussion

This study including more than 300,000 persons aged 55–69 years is—to our knowledge—the largest observational study on the effectiveness of colonoscopy in preventing distal vs. proximal CRC. Unlike previous observational studies, our study did not show any substantial difference in effectiveness between proximal and distal CRC. The 11-year risk of CRC in the colonoscopy vs. control arm was reduced by 33% (confidence interval [CI]: 27–38%) for distal and by 30% (CI: 21–37%) for proximal CRC. The cumulative incidence curves of the screening and control arm crossed after 6.7 years follow-up for distal CRC and after 5.0 years for proximal CRC.

This is the first observational study on the effectiveness of screening colonoscopy in reducing distal vs. proximal CRC incidence using a target-trial emulation. The advantage of

**Table 1.** Baseline characteristics stratified by gender and treatment arm (screening colonoscopy arm vs. control arm). All numbers refer to nonunique persons

| | Male | | | | Female | | | | Total | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Screening (*N* = 99,101) | | No screening (*N* = 583,861) | | Screening (*N* = 99,288) | | No screening (*N* = 664,052) | | Screening (*N* = 198,389) | | No screening (*N* = 1,247,913) | |
| **Characteristic** | *n* | % | *n* | % | *n* | % | *n* | % | *n* | % | *n* | % |
| Age | | | | | | | | | | | | |
| Median (Q1–Q3) | 61 | 57–65 | 61 | 58–66 | 60 | 57–65 | 62 | 58–66 | 61 | 57–65 | 62 | 58–66 |
| Mean (SD) | 61.3 | 4.50 | 61.6 | 4.47 | 60.9 | 4.55 | 61.9 | 4.51 | 61.1 | 4.53 | 61.8 | 4.49 |
| Education | | | | | | | | | | | | |
| No degree/unknown | 51,793 | 52.3 | 343,765 | 58.9 | 66,309 | 66.8 | 495,934 | 74.7 | 118,102 | 59.5 | 839,699 | 67.3 |
| Basic or secondary degree | 19,876 | 20.1 | 117,450 | 20.1 | 20,391 | 20.5 | 109,463 | 16.5 | 40,267 | 20.3 | 226,913 | 18.2 |
| Higher education | 27,432 | 27.7 | 122,646 | 21.0 | 12,588 | 12.7 | 58,655 | 8.8 | 40,020 | 20.2 | 181,301 | 14.5 |
| Region | | | | | | | | | | | | |
| East Germany | 21,926 | 22.1 | 114,590 | 19.6 | 22,423 | 22.6 | 131,645 | 19.8 | 44,349 | 22.2 | 246,235 | 19.7 |
| West Germany | 77.175 | 77.9 | 469,271 | 80.4 | 76,865 | 77.4 | 532,407 | 80.2 | 154,040 | 77.6 | 1,001,678 | 80.3 |
| Codes indicating obesity[a] | 12,178 | 12.3 | 71,137 | 12.2 | 13,649 | 13.7 | 94,443 | 14.2 | 25,827 | 13.0 | 165,580 | 13.3 |
| Diabetes type 2 | 14,762 | 14.9 | 98,120 | 16.8 | 8,689 | 8.8 | 75,849 | 11.4 | 23,451 | 11.8 | 173,969 | 13.9 |
| Codes indicating a family history of CRC[a] | 91 | 0.1 | 145 | <0.05 | 409 | 0.4 | 851 | 0.1 | 500 | 0.3 | 996 | 0.1 |
| Menopausal hormone therapy | N.A. | | N.A. | | 22,759 | 22.9 | 95,439 | 14.4 | N.A. | | N.A. | |
| Use of acetylsalicylic acid | 4,743 | 4.8 | 31,008 | 5.3 | 1,527 | 1.5 | 13,164 | 2.0 | 6,270 | 3.2 | 44,172 | 3.5 |
| Codes for alcohol abuse[a] | 2,911 | 2.9 | 27,212 | 4.7 | 1,485 | 1.5 | 14,477 | 2.2 | 4,396 | 2.2 | 41,689 | 3.3 |
| Codes for heavy smoking[a] | 5,487 | 5.5 | 42,871 | 7.3 | 4,742 | 4.8 | 35,438 | 5.3 | 10,229 | 5.2 | 78,309 | 6.3 |
| Use of other preventive services during 3 years before baseline[b] | | | | | | | | | | | | |
| None | 23,109 | 23.3 | 258,419 | 44.3 | 5,888 | 5.9 | 162,228 | 24.4 | 28,997 | 14.6 | 420,647 | 33.7 |
| One or more | 75,992 | 76.7 | 325,442 | 55.7 | 93,400 | 94.1 | 501,824 | 75.6 | 169,392 | 85.4 | 827,266 | 66.3 |

Q1–Q3, interquartile range; SD, standard deviation.

[a] Only coded if there is a reimbursement of treatment or services due to these conditions, not coded for all persons with the respective condition.

[b] Used as a proxy variable for preventive behavior.

this approach lies in avoiding time-related and other biases that can be introduced by a poor study design, also called self-inflicted biases because they are avoidable [10]. Previous observational studies addressing this research question may have suffered from such biases. For example, a study by Guo et al. suggesting a 64% risk reduction for distal CRC but only a 31% risk reduction for proximal CRC inquired at baseline about past colonoscopies and—based on this information—assigned persons as exposed or unexposed to colonoscopy. Persons reporting at baseline CRC in the past were excluded [6]. Given that colonoscopy is typically used for CRC diagnosis, this exclusion criterion mainly affects the colonoscopy group, leading to an imbalance regarding prevalent CRCs yet undetected at baseline (i.e., less in the colonoscopy group). As a result, the cumulative CRC incidence in the colonoscopy group during follow-up is artificially lowered, leading to an overestimate of the preventive effect of colonoscopy as described by Garcia–Albeniz et al [7]. As the vast majority of CRCs diagnosed at an age when persons are typically included into screening studies are in the distal colon [11], whereas proximal CRC is more common at an older age, it seems plausible that the overestimation mainly concerned the preventive effect for distal rather than proximal CRC. Accordingly, also the difference in the effectiveness of colonoscopy by location was overestimated. It seems likely that the differential age distribution of distal and proximal CRC also introduced considerable bias in case-control studies and other types of cohort studies suggesting a substantially higher effectiveness of colonoscopy in the incidence or mortality of distal vs. proximal CRC [4,5,12].

In the interpretation of prior studies suggesting a substantially lower effectiveness of colonoscopy for proximal vs. distal CRC, a higher miss rate of colonoscopy or special biological properties of precursor lesions in the proximal colon were assumed to explain the findings. Particularly, sessile serrated lesions play a major role in this reasoning as they primarily occur in the proximal colon. They act as precursors to CRC developing via the serrated pathway characterized by the CpG methylator phenotype and microsatellite instability and are assumed to account for 25% of sporadic CRCs [13]. Some studies reported that the

**Table 2.** Number of incident CRC and adjusted effect estimates at 11 years of follow-up, stratified by site

| | | # Nonunique cases | | | 11-year absolute risk difference | | 11-year relative risk | |
|---|---|---|---|---|---|---|---|---|
| Site | Gender | Screening ($N$ = 198,389) | No screening ($N$ = 1,247,913) | NNS | % | [95% CI[a]] | | [95% CI[a]] |
| Distal | Male | 1,046 | 8,211 | | | | | |
| | Female | 521 | 5,004 | | | | | |
| | Total | 1,567 | 13,215 | 213 | 0.47 | [0.35; 0.57] | 0.67 | [0.62; 0.73] |
| Proximal | Male | 385 | 3,244 | | | | | |
| | Female | 350 | 3,215 | | | | | |
| | Total | 735 | 6,459 | 463 | 0.22 | [0.14; 0.29] | 0.70 | [0.63; 0.79] |
| Both distal and proximal or unknown site | Male | 133 | 1,153 | | | | | |
| | Female | 105 | 1,146 | | | | | |
| | Total | 238 | 2,299 | | | | | |
| Total | Male | 1,564 | 12,608 | | | | | |
| | Female | 976 | 9,365 | | | | | |
| | Total | 2,540 | 21,973 | 131 | 0.77 | [0.62; 0.91] | 0.68 | [0.63; 0.73] |

*Abbreviation:* NNS, number needed to screen, calculated as the inverse of the absolute risk reduction.

No effect estimates are given for both distal and proximal or unknown site and for gender-specific incidence, as there were too few cases for reliable estimation.

[a] Person-level percentile bootstrap confidence intervals based on 250 bootstrap samples.

detection rate for sessile serrated lesions varied between endoscopists and correlated with their adenoma detection rate [14,15]. In the real-world setting, variation in the detection of sessile serrated lesions might thus be relevant. Our findings, however, do not suggest a strong impact of this variability regarding potential differences in the effectiveness of colonoscopy by site as proposed previously. Colonoscopies in our study were performed in 2007 or later, that is, at a time of heightened attention toward the quality of colonoscopy but we think it is unlikely that this explains the large discrepancy with the results of prior studies on site-specific effectiveness of colonoscopy.

Sessile serrated lesions have also been associated with a higher risk of metachronous neoplasia compared to conventional adenomas [16–18]. Whether this could lead to lower effectiveness of colonoscopy in the proximal colon also depends on adherence to surveillance colonoscopy. We previously showed that among persons with prior snare polypectomy in Germany about 60% underwent at least one repeat colonoscopy within 5 years and about 80% within 10 years

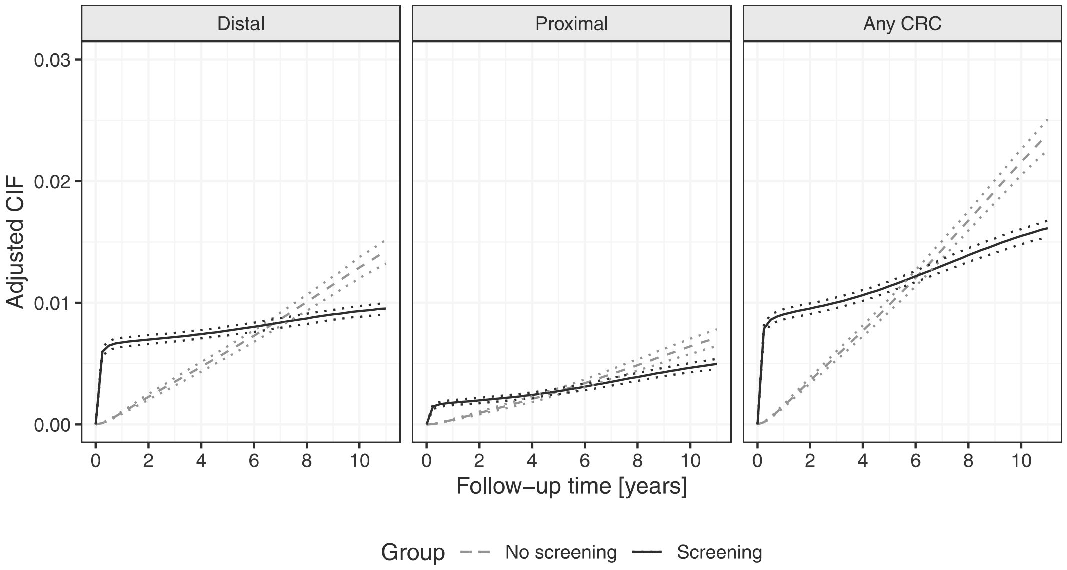


**Fig. 3.** Adjusted cumulative incidence functions showing 11 years of follow-up. Analyses were done by site of incident CRC. No separate analyses were done for incident CRCs of unknown location and simultaneous distal and proximal incident CRCs because too few events were observed.

[19]. The effectiveness of screening colonoscopy estimated in our study includes the potential effect attributable to these surveillance colonoscopies. Effectiveness might have been higher in case of perfect adherence to surveillance or lower in case of a poorer uptake.

In our study, the curves for proximal CRC crossed about 2 years earlier than for distal CRC. This may suggest that the time between transition from precancerous lesions or preclinical CRC to clinical CRC is, on average, shorter for proximal than for distal CRC. In view of the distinct molecular features of distal and proximal CRC [20], differences in the natural history by location seem plausible and could further be elucidated by the promising field of molecular pathological epidemiology [21]. Although direct evidence on adenoma dwell and sojourn time is hardly obtainable, analyses showing poorer survival for proximal than for distal CRC [22] and case reports on fast-growing sessile serrated lesions indirectly support a high progressive potential of neoplasia in the proximal colon [23,24]. Of note, our findings refer to persons aged 55–69 years at screening colonoscopy. Caution is needed when extrapolating the results to older ages, also because the natural history likely differs by age and the importance of the competing event death increases with age.

When comparing our results for distal CRC to RCT findings on screening with flexible sigmoidoscopy [2,25], one should note that no exact agreement was expected for several reasons. First, most RCTs included persons aged 55–64 years at baseline [2,26], whereas we included persons aged 55–69 years. Second, the intention-to-screen effect reported in these trials depends on adherence at baseline (varying 60–80%) and is thus not directly comparable to the effect estimate in our study where all persons in the screening arm underwent colonoscopy at baseline. Also, the per-protocol effects reported by RCTs are not directly comparable to our results, because in our study, persons in the control arm were not censored if they underwent screening colonoscopy later. This contamination equally affected distal and proximal CRC, so there was no differential effect (Supplement Table S2). As our research question focused on the difference in the effectiveness by location, we favored this approach over censoring nonscreened persons who were screened during follow-up, as it avoided further assumptions and we preferred the more conservative method. Had our aim been to assess the overall efficacy of screening colonoscopy, corresponding to the per-protocol effect of an RCT, this would have been inadequate, so we would have chosen another approach. Third, the effect of screening also depends on adherence to recommended surveillance intervals, which may be lower in a real-world setting compared to trials. In Germany, at least 40% of persons with polypectomy have been estimated to not adhere to recommended surveillance intervals [19]. Furthermore, the effect of screening depends on the background prevalence of diagnostic colonoscopy. In Germany, the 10-year prevalence of diagnostic colonoscopy among persons aged 55–69 years was about 22–26% in 2017 [27], that is, a relevant proportion of persons in the control arm may have had a diagnostic colonoscopy during follow-up. This concerns the control group in general, so it is not expected to bias the comparison of site-specific effectiveness of screening colonoscopy. Also, fecal occult blood testing may have occurred in the control arm during follow-up. However, it is not expected that the recommended fecal occult blood test during the study period—the guaiac test—had a relevant impact on CRC incidence, since RCT evidence on this test mainly showed an effect on CRC mortality rather than on incidence [28].

In the interpretation of our results, some limitations must be considered. First, although we used as much information as possible to control for confounding, claims data are suboptimal in this regard, especially with respect to lifestyle factors. As proxy information we mainly used conditions like obesity or diabetes and the use of other preventive services. However, as discussed by Garcia–Albéniz et al. [8] it is unlikely that residual confounding plays a major role here as adjustment for potential confounders had little impact on previous observational studies [29,30]. Furthermore, CRC incidence in the control group and in noncompliers was similar in RCTs [25,26,31]. There was also no indication of any noteworthy residual confounding in a negative control analysis (Supplement 6). Second, there are specific codes for screening colonoscopy in our database and we additionally used several exclusion criteria to focus on an asymptomatic average-risk population, that is, the target population of screening. Nonetheless, it is possible that symptoms were not coded in the database. We assume this did not play a major role in our study as the CRC detection rate observed in the screening arm at baseline is plausible and comparable to that reported in screening trials (0.6% in an analysis restricted to 55–64-year-olds compared to 0.5% in the Nordic-European Initiative on Colorectal Cancer trial including 55–64-year-olds).

In conclusion, the results of our observational study using an emulated target-trial approach suggest that colonoscopy is effective in preventing distal and proximal CRC. Unlike in previous studies not using a target-trial approach, there was no relevant difference in the effectiveness by location. The distinct temporal patterns of the cumulative incidence curves support hypotheses regarding differences in the natural history of distal vs. proximal CRC.

## Author contributions

M.B.: Conceptualization, data curation, formal analysis, investigation, methodology, software, visualization, writing—original draft, and writing—review and editing; S.S.: Conceptualization, investigation, and writing—review and editing; B.K.: Conceptualization, investigation, and writing—review and editing; C.S.: Investigation and writing—review and editing; V.D.: Conceptualization,

investigation, methodology, supervision, and writing−review and editing; U.H.: Conceptualization, funding acquisition, investigation, methodology, project administration, resources, supervision, writing−original draft, and writing−review and editing.

## Acknowledgments

We thank Anja Gabbert and Inga Schaffer for the programming of analysis datasets and descriptive analyses. Furthermore, we thank Simon Klau for his validation of statistical program code. We thank Heike Gerds for proofreading the manuscript.

## Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jclinepi.2022.05.024.

## References

[1] Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2021;71:209−49.

[2] Elmunzer BJ, Hayward RA, Schoenfeld PS, Saini SD, Deshpande A, Waljee AK. Effect of flexible sigmoidoscopy-based screening on incidence and mortality of colorectal cancer: a systematic review and meta-analysis of randomized controlled trials. PLoS Med 2012; 9:e1001352.

[3] Bretthauer M, Kaminski MF, Loberg M, Zauber AG, Regula J, Kuipers EJ, et al. Population-based colonoscopy screening for colorectal cancer: a randomized clinical trial. JAMA Intern Med 2016; 176:894−902.

[4] Nishihara R, Wu K, Lochhead P, Morikawa T, Liao X, Qian ZR, et al. Long-term colorectal-cancer incidence and mortality after lower endoscopy. N Engl J Med 2013;369:1095−105.

[5] Brenner H, Chang-Claude J, Seiler CM, Rickert A, Hoffmeister M. Protection from colorectal cancer after colonoscopy: a population-based, case-control study. Ann Intern Med 2011;154:22−30.

[6] Guo F, Chen C, Holleczek B, Schottker B, Hoffmeister M, Brenner H. Strong reduction of colorectal cancer incidence and mortality after screening colonoscopy: prospective cohort study from Germany. Am J Gastroenterol 2021;116:967−75.

[7] Garcia-Albeniz X, Hsu J, Hernan MA. The value of explicitly emulating a target trial when using real world evidence: an application to colorectal cancer screening. Eur J Epidemiol 2017;32: 495−500.

[8] Garcia-Albeniz X, Hsu J, Bretthauer M, Hernan MA. Effectiveness of screening colonoscopy to prevent colorectal cancer among medicare beneficiaries aged 70 to 79 years: a prospective observational study. Ann Intern Med 2017;166:18−26.

[9] Haug U, Schink T. German pharmacoepidemiological research database (GePaRD). In: Sturkenboom M, Schink T, editors. Databases for pharmacoepidemiological research. Cham, Switzerland: Springer Nature Switzerland AG; 2021:119−24.

[10] Hernan MA, Sauer BC, Hernandez-Diaz S, Platt R, Shrier I. Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses. J Clin Epidemiol 2016; 79:70−5.

[11] Meza R, Jeon J, Renehan AG, Luebeck EG. Colorectal cancer incidence trends in the United States and United Kingdom: evidence of right- to left-sided biological gradients with implications for screening. Cancer Res 2010;70:5419−29.

[12] Baxter NN, Goldwasser MA, Paszat LF, Saskin R, Urbach DR, Rabeneck L. Association of colonoscopy and death from colorectal cancer. Ann Intern Med 2009;150:1−8.

[13] Crockett SD, Nagtegaal ID. Terminology, molecular features, epidemiology, and management of serrated colorectal neoplasia. Gastroenterology 2019;157:949−966.e4.

[14] Meester RGS, van Herk MMAGC, Lansdorp-Vogelaar I, Ladabaum U. Prevalence and clinical features of sessile serrated polyps: a systematic review. Gastroenterology 2020;159:105.

[15] Kahi CJ, Hewett DG, Norton DL, Eckert GJ, Rex DK. Prevalence and variable detection of proximal colon serrated polyps during screening colonoscopy. Clin Gastroenterol Hepatol 2011;9:42−6.

[16] Macaron C, Vu HT, Lopez R, Pai RK, Burke CA. Risk of metachronous polyps in individuals with serrated polyps. Dis Colon Rectum 2015;58:762−8.

[17] Lu FI, van Niekerk de W, Owen D, Tha SP, Turbin DA, Webber DL. Longitudinal outcome study of sessile serrated adenomas of the colorectum: an increased risk for subsequent right-sided colorectal carcinoma. Am J Surg Pathol 2010;34:927−34.

[18] Schreiner MA, Weiss DG, Lieberman DA. Proximal and large hyperplastic and nondysplastic serrated polyps detected by colonoscopy are associated with neoplasia. Gastroenterology 2010;139:1497−502.

[19] Schwarz S, Schafer W, Horenkamp-Sonntag D, Liebentraut J, Haug U. Follow-up of 3 million persons undergoing colonoscopy in Germany: utilization of repeat colonoscopies and polypectomies within 10 years. Clin Transl Gastroenterol 2020;12:e00279.

[20] Baran B, Mert Ozupek N, Yerli Tetik N, Acar E, Bekcioglu O, Baskin Y. Difference between left-sided and right-sided colorectal cancer: a focused review of literature. Gastroenterol Res 2018;11: 264−73.

[21] Hughes LAE, Simons CCJM, van den Brandt PA, van Engeland M, Weijenberg MP. Lifestyle, diet, and colorectal cancer risk according to (epi) genetic instability: current evidence and future directions of molecular pathological epidemiology. Curr Colorectal Cancer Rep 2017;13:455−69.

[22] Yahagi M, Okabayashi K, Hasegawa H, Tsuruta M, Kitagawa Y. The worse prognosis of right-sided compared with left-sided colon cancers: a systematic review and meta-analysis. J Gastrointest Surg 2016;20:648−55.

[23] Amemori S, Yamano HO, Tanaka Y, Yoshikawa K, Matsushita HO, Takagi R, et al. Sessile serrated adenoma/polyp showed rapid malignant transformation in the final 13 months. Dig Endosc 2020;32: 979−83.

[24] Oono Y, Fu K, Nakamura H, Iriguchi Y, Yamamura A, Tomino Y, et al. Progression of a sessile serrated adenoma to an early invasive cancer within 8 months. Dig Dis Sci 2009;54:906−9.

[25] Holme O, Loberg M, Kalager M, Bretthauer M, Hernan MA, Aas E, et al. Long-term effectiveness of sigmoidoscopy screening on colorectal cancer incidence and mortality in women and men: a randomized trial. Ann Intern Med 2018;168:775−82.

[26] Segnan N, Armaroli P, Bonelli L, Risio M, Sciallero S, Zappa M, et al. Once-only sigmoidoscopy in colorectal cancer screening: follow-up findings of the Italian Randomized Controlled Trial–SCORE. J Natl Cancer Inst 2011;103:1310−22.

[27] Hornschuch M, Schwarz S, Haug U. 10-year prevalence of diagnostic and screening colonoscopy use in Germany: a claims data analysis. Eur J Cancer Prev 2022. https://doi.org/10.1097/CEJ.0000000000000736.

[28] Hewitson P, Glasziou P, Irwig L, Towler B, Watson E. Screening for colorectal cancer using the faecal occult blood test, Hemoccult. Cochrane Database Syst Rev 2007;2007:CD001216.

[29] Kahi CJ, Myers LJ, Slaven JE, Haggstrom D, Pohl H, Robertson DJ, et al. Lower endoscopy reduces colorectal cancer incidence in older individuals. Gastroenterology 2014;146:718−725.e3.

[30] Brenner H, Chang-Claude J, Jansen L, Knebel P, Stock C, Hoffmeister M. Reduced risk of colorectal cancer up to 10 years after screening, surveillance, or diagnostic colonoscopy. Gastroenterology 2014;146:709−17.

[31] Atkin WS, Edwards R, Kralj-Hans I, Wooldrage K, Hart AR, Northover JM, et al. Once-only flexible sigmoidoscopy screening in prevention of colorectal cancer: a multicentre randomised controlled trial. Lancet 2010;375:1624−33.

Supplementary Information for the manuscript

# Screening colonoscopy similarly prevented distal and proximal colorectal cancer; A prospective study among 55-69-year-olds

Malte Braitmaier, Sarina Schwarz, Bianca Kollhorst, Carlo Senore, Vanessa Didelez, Ulrike Haug

Corresponding author: Ulrike Haug, Leibniz Institute for Prevention Research and Epidemiology – BIPS; Department of Clinical Epidemiology; haug@leibniz-bips.de

## Contents

# Supplement 1: Study protocol of target trial and its emulation

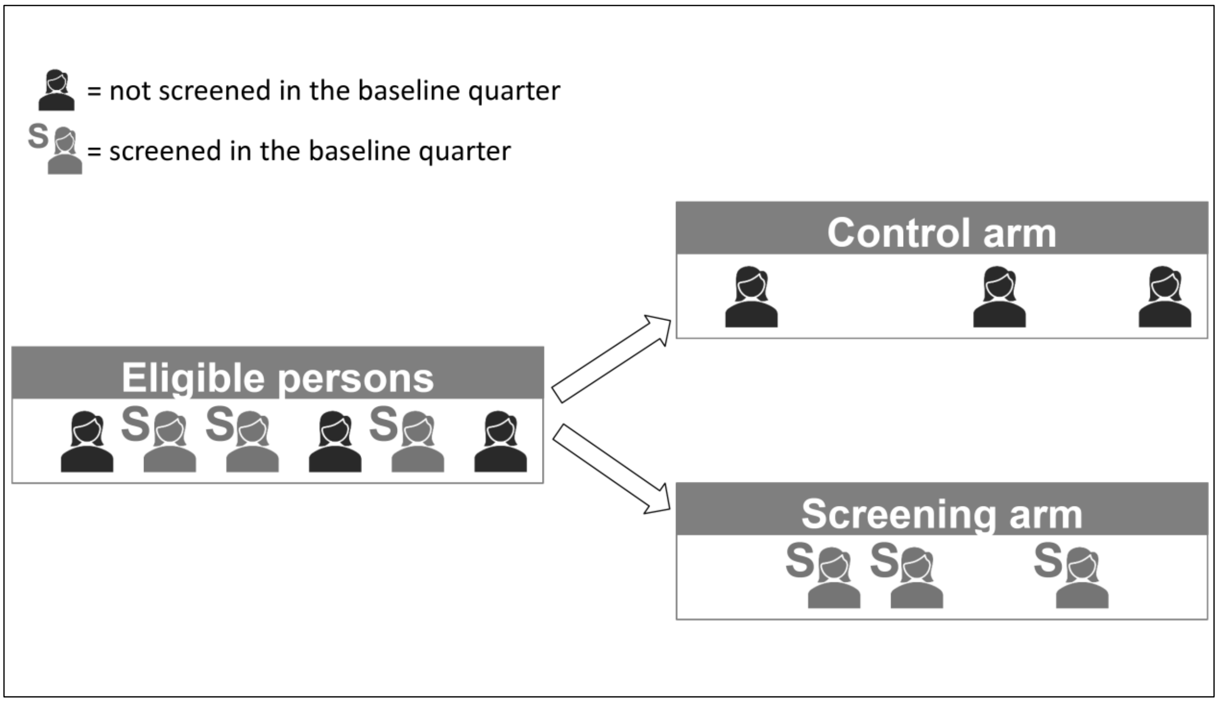
**Table S1**: Specification of the study protocol of a target trial and its emulation using observational health insurance claims data from GePaRD.

| ***Component*** | ***Target trial*** | ***Emulated trial*** |
|---|---|---|
| Aim | To estimate the effect of screening colonoscopies on the 11-year risk of any CRC, distal CRC and proximal colon cancer in the German population aged 55 to 69. | Same |
| Eligibility | To be eligible, persons must<br>• be aged 55 to 69 at baseline.<br>• be without gastrointestinal symptoms in the last 6 months before baseline.<br>• have no history of CRC (including carcinomas in situ), adenomas and precursors.<br>• be colonoscopy naïve.<br>• have no diagnosis of HNPCC.<br>• have no history of inflammatory bowel disease. | To be eligible, persons must<br>• be aged 55 to 69 in the year of the trial.<br>• have no coded gastrointestinal symptoms in the last six months before baseline, as well as no coded abdominal computed tomography and barium enema.<br>• have no coded CRCs, colorectal carcinomas in situ, adenomas, and precursors ever before baseline.<br>• have no coded colonoscopy, polypectomy, colectomy, or fecal occult blood test (FOBT) ever before baseline.<br>• be continuously insured ever before baseline.<br>• have no coded HNPCC test ever before baseline.<br>• have no diagnosis codes of inflammatory bowel disease ever before baseline. |
| Treatment strategies | • Screening colonoscopy at baseline.<br>• No screening colonoscopy at baseline.<br>Access to surveillance, further CRC screening or diagnostic colonoscopy during follow-up under both strategies. | Same |

| | | |
|---|---|---|
| Treatment assignment | Randomized assignment | Non-random, patients who received a screening colonoscopy in the baseline quarter are assigned to the screening group, and otherwise to the comparison group. Randomization will be emulated via adjustment for the following baseline variables:   age, sex, education, obesity, CRC in family history, use of menopausal hormone therapy, use of (low-dose) acetylsalicylic acid, diabetes mellitus type 2, alcohol, smoking, use of preventive services other than colonoscopy screening during three years before baseline (0, 1, ≥2). |
| Follow-up (FU) | • Start: Treatment assignment.<br>• End: CRC diagnosis, death, loss to FU, or 31 December 2017 (end of study), whichever occurs first. | Same except start is quarter of treatment assignment with a new trial starting every quarter from 2007 to 2011; 20 emulated trials in total. |
| Outcome | CRC diagnosis within 11 years after baseline. | CRC diagnosis within 11 years as from quarter of screening colonoscopy: |
| Causal contrast | Effect of receiving screening colonoscopy at baseline | Effect of receiving a screening colonoscopy vs not at baseline, regardless of screening utilization after baseline |
| Statistical analysis | Total effect measured as contrast of cumulative incidence functions over the whole follow-up (i.e. not eliminating death as competing event). | Same with additional adjustment for baseline confounding by inverse probability of treatment weighting. |

# Supplement 2: Illustration of target trial emulation

**Figure S1:** Illustration of treatment assignment. The persons displayed in the figure depict a hypothetical trial with a fixed date as baseline (e.g. January 1, 2007). Eligible persons who received a screening colonoscopy in the baseline quarter are assigned to the screening arm, whereas eligible persons without a screening colonoscopy are assigned to the control arm.


= not screened in the baseline quarter
S = screened in the baseline quarter
Control arm
Eligible persons
S S S
Screening arm
S S S

# Supplement 3: Statistical modeling of cumulative incidence curves

Contrasts between cumulative incidence curves (CIF) were used as statistical measure of the effect of interest. CIFs were estimated using flexible pooled logistic regression models (D'Agostino et al. 1990) and adjusted for baseline covariates via inverse probability of treatment weighting (IPTW). Note that sequential trials start at different time points and time $t$ is follow-up time (i.e., time from start of the respective trial). Let time be discrete in quarterly intervals and let $t \geq 0$ be a time point. Furthermore, let $Y_t \in \{0,1\}$ be an indicator variable for a CRC diagnosis at time $t$, let $A \in \{0,1\}$ be an indicator variable for the assigned screening strategy. The screening strategy $A$ was not time-varying since we assessed the effect of receiving colonoscopy screening at baseline, regardless of subsequent screening behavior during follow-up. In the following, the overbar notation ($\bar{Z}_t$) is used to denote a variable's history up to time $t$. In the following, person-specific subscripts are mostly suppressed. However, assume that our sample comprised $n$ entries from $m$ unique persons, some of which were included in more than one trial (i.e. $n \geq m$).

Then, the discrete-time hazard of a CRC diagnosis was modelled as

$$\begin{aligned}&\operatorname{logit}\{\mathbb{P}(Y_{t+1} = 1 | \bar{Y}_t = 0, A)\}\\ &\qquad = \beta_1 t + \beta_2 t^2 + \beta_3 \sqrt{t} + \beta_4 \log(t) + \beta_5 tA + \beta_6 t^2 A + \beta_7 \sqrt{t} A + \beta_8 \log(t)\, A.\end{aligned}$$

In the above equation, $\beta$ are the coefficients of the pooled logistic model. The transformations of time were selected so that the unadjusted parametric model returned the same results as a non-parametric Aalen-Johansen analysis (assessed visually, Aalen & Johansen 1978). The above hazard is then transformed and weighted to obtain the adjusted CIF for CRC (Hernán & Robins 2020). Covariate balance after weighting was examined using absolute standardized differences (Stuart et al. 2013). The contrast of CIFs for screened and unscreened can be interpreted as total causal effect of screening on CRC incidence, where 'total' means that the competing event of death is not eliminated. We preferred this approach as it is meaningful in a real-world setting and avoids additional assumptions regarding no unobserved confounding between death (i.e., the competing event) and colorectal cancer incidence (i.e., the outcome event). The alternative of estimating the direct effect by treating the competing event of death as a censoring event, which is often the default analysis, was carried out in a sensitivity analysis (see Supplement 10: below). This direct causal effect corresponds to the question of what would have happened in a hypothetical setting where death as competing event was eliminated had all individuals in the sample received screening versus had all individuals not received screening. As illustrated below in Supplement 10:, this did not substantially change the results. For details on total and direct effects, see Young et al. (2020).

Adjustment for baseline confounding was achieved via stabilized inverse probability of treatment weighting. For this, PS were calculated via logistic regression, i.e., the probability of undergoing screening in the baseline quarter was calculated conditional on age at baseline, sex, educational attainment, CRC in family history, obesity, use of acetylsalicylic acid, menopausal hormone therapy, type 2 diabetes, alcohol dependence, nicotine dependence and use of preventive services during three years before baseline (zero, one, at least two). Weights were truncated at the 99th percentile to avoid instable estimation due to extreme values. The CIFs were estimated as the product over time of estimated outcome probabilities based on the above discrete-time hazards. Our approach corresponds to the total effect estimated via inverse probability of treatment weighted estimators using subdistribution hazards as described in Young et al. (2020).

Our sample consisted of $m$ unique persons, some of which were included in more than one emulated trial, so that the final sample size (including non-unique persons) was $n \geq m$. Confidence intervals were computed using robust, person-level bootstrapping. For this, a bootstrap sample was obtained by sampling (with replacement) $m$ observations from the list of unique persons. The process of emulating target trials and including some persons in more than one emulated trial was then repeated for this bootstrap sample and the above, adjusted standardized CIFs were computed. This process was repeated for $B = 250$ bootstrap iterations and pointwise, percentile-based 95 % confidence intervals were derived from the resulting $B$ bootstrap estimates.

We chose the approach of not imposing full adherence over follow-up as it avoids the need for much stronger assumptions concerning fully measured time-varying confounding. Furthermore, we preferred the more conservative estimate resulting from individuals in the control strategy undergoing screening later during follow-up. We point out that this approach is not directly comparable to either intention-to-treat or per-protocol effects from RCTs. Although some studies that emulated target trials refer to this approach as an intention-to-treat effect, there is no non-adherence at baseline (which would occur in an intention-to-treat analysis in an RCT). At the same time, we do not censor individuals during follow-up, when they stop adhering to the assigned screening strategy, which would be required for a per-protocol effect. However, our focus in this study was to compare site-specific effects of colonoscopy and not to estimate effects that are directly comparable to RCTs. We therefore chose the approach that avoided additional, strong assumptions to ensure the highest possible validity of results regarding our main research question.

# Supplement 4: Data source, study population and identification/ classification of CRC cases in GePaRD

We used the German Pharmacoepidemiological Research Database (GePaRD) which is based on claims data from four statutory health insurance providers in Germany and currently includes information on approximately 25 million persons who have been insured with one of the participating providers since 2004 or later. Details about GePaRD have been described elsewhere (Pigeot & Ahrens 2008, Haug & Schink 2021). In addition to demographic data, GePaRD contains information on drug dispensations as well as outpatient (i.e., from general practitioners and specialists) and inpatient services and diagnoses. Per data year, there is information on approximately 20% of the general population, and all geographical regions of Germany are represented. For this study, we used data from 2004 to 2017.

In GePaRD information on utilization of screening colonoscopy, which has been offered in Germany since 2002 to persons aged 55 or older, is available including the date of the procedure. Screening colonoscopy can be distinguished from diagnostic colonoscopy as there are different reimbursement codes for these procedures.

Age, sex, educational attainment, codes indicating a family history of CRC, codes indicating obesity, codes indicating type 2 diabetes, codes indicating severe alcohol abuse, codes indicating severe nicotine dependence, use of low-dose acetylsalicylic acid, use of menopausal hormone therapy, and use of preventive services (none, one, or at least two during three years before baseline) were assessed as baseline covariates. The latter served as a proxy variable for a preventive behavior. Diagnosis codes and prescriptions relevant for the ascertainment of baseline covariates were considered in the three years before baseline, except for codes regarding family history which were considered any time before baseline. Codes used to derive analysis variables are available upon request.

CRC diagnoses in GePaRD are coded according to the German modification of the International Classification of Diseases, 10th revision (ICD-10-GM). We considered inpatient diagnosis codes of CRC, which are considered to have a high validity. To avoid misclassification, patients with only outpatient diagnosis codes of CRC were classified as CRC cases if additional criteria such as codes for diagnostic procedures and surveillance were met. Roughly 98% of CRC cases had an inpatient CRC diagnosis. Regarding classification of location into proximal and distal to the splenic flexure, we used the information as provided by the ICD code (proximal: C18.0-C18.4; distal: C18.5-C18.7, C19, C20). CRCs with unclear information on location (C18.8 and C18.9) or with two or more codes providing discordant information regarding proximal vs. distal location were classified into the category “both / unknown”. A more detailed classification of tumor location would have resulted in more missing values given that information on the exact location was more often discordant. Stage at diagnosis was roughly estimated based on ICD codes indicating lymph node involvement or distant metastases as previously described (Oppelt et al. 2019). Additionally, we considered codes for cancer treatment typically used in more advanced stages. Based on this information, CRCs were classified into the categories “advanced” and “non-advanced”. All codes used are available on request.

# Supplement 5: Characterization of incident CRC cases

**Table S2**: Characterization of incident CRC cases

| Variable | Screening | | | | | | No screening | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Distal | | Proximal | | Both/Unknown | | Distal | | Proximal | | Both/Unknown | |
| | (N=1,567) | | (N=735) | | (N=238) | | (N=13,215) | | (N=6,459) | | (N=2,299) | |
| Age at diagnosis | | | | | | | | | | | | |
| Mean (SD[&]) | 64.2 | (4.97) | 66.2 | (5.47) | 65.8 | (5.09) | 67.6 | (4.83) | 68.1 | (4.82) | 66.9 | (5.18) |
| Median (Q1; Q3[$]) | 65 | (60; 68) | 66 | (62; 70) | 66 | (62; 69) | 68 | (64; 71) | 68 | (65; 72) | 67 | (63; 70) |
| Number of colonoscopies before diagnosis, % | | | | | | | | | | | | |
| 0[§] | 1,027 | 65.5 | 249 | 33.9 | 73 | 30.7 | 12,026 | 91.0 | 5,549 | 85.9 | 1,990 | 86.6 |
| 1 | 420 | 26.8 | 326 | 44.4 | 120 | 50.4 | 760 | 5.8 | 617 | 9.6 | 244 | 10.6 |
| 2 or more | 120 | 7.7 | 160 | 21.8 | 45 | 18.9 | 429 | 3.2 | 293 | 4.5 | 65 | 2.8 |
| Stage* at diagnosis of CRC, % | | | | | | | | | | | | |
| Non-advanced | 1,109 | 70.8 | 495 | 67.3 | 188 | 79.0 | 6,777 | 51.3 | 3,564 | 55.2 | 1,373 | 59.7 |
| Advanced | 458 | 29.2 | 240 | 32.7 | 50 | 21.0 | 6,438 | 48.7 | 2,895 | 44.8 | 926 | 40.3 |
| CRCs among persons in the screening arm who were also included as controls** in at least one previous trial, %. | 66 | 4.2 | 35 | 4.8 | 8 | 3.4 | N.A. | N.A. | N.A. | N.A. | N.A. | N.A. |

&: standard deviation

$: 1st and 3rd quartiles

§: Colonoscopies from the quarter of diagnosis were not counted, i.e. cancers detected at the baseline screening could be detected without prior colonoscopies according to this definition. Further, colonoscopies before baseline were an exclusion criterion, i.e., all colonoscopies counted for this table occurred earliest in the baseline quarter and latest in the quarter before diagnosis.

*: Stage was defined as described above.

**: Overall, 16.9% of controls were also included in the screening arm of a later trial.

# Supplement 6: Negative control analysis

A negative control analysis was conducted to assess the possibility of residual unmeasured confounding. The analysis was carried out as described for the effect of screening colonoscopy on overall CRC incidence, but CRC was replaced by pancreatic cancer as outcome variable. If no residual unmeasured confounding was present, one would expect no association as there is no mechanism by which screening colonoscopy could affect the risk of pancreatic cancer. However, the usual assumptions and limitations of negative control analyses must be kept in mind (see Lipsitch et al. 2010).

Figure S2 indicates no difference in cumulative incidence during the first seven years of follow-up with the possibility of only some small amount of residual confounding towards the end of follow-up. However, the difference is very small and the confidence intervals still allow this difference to be due to chance.

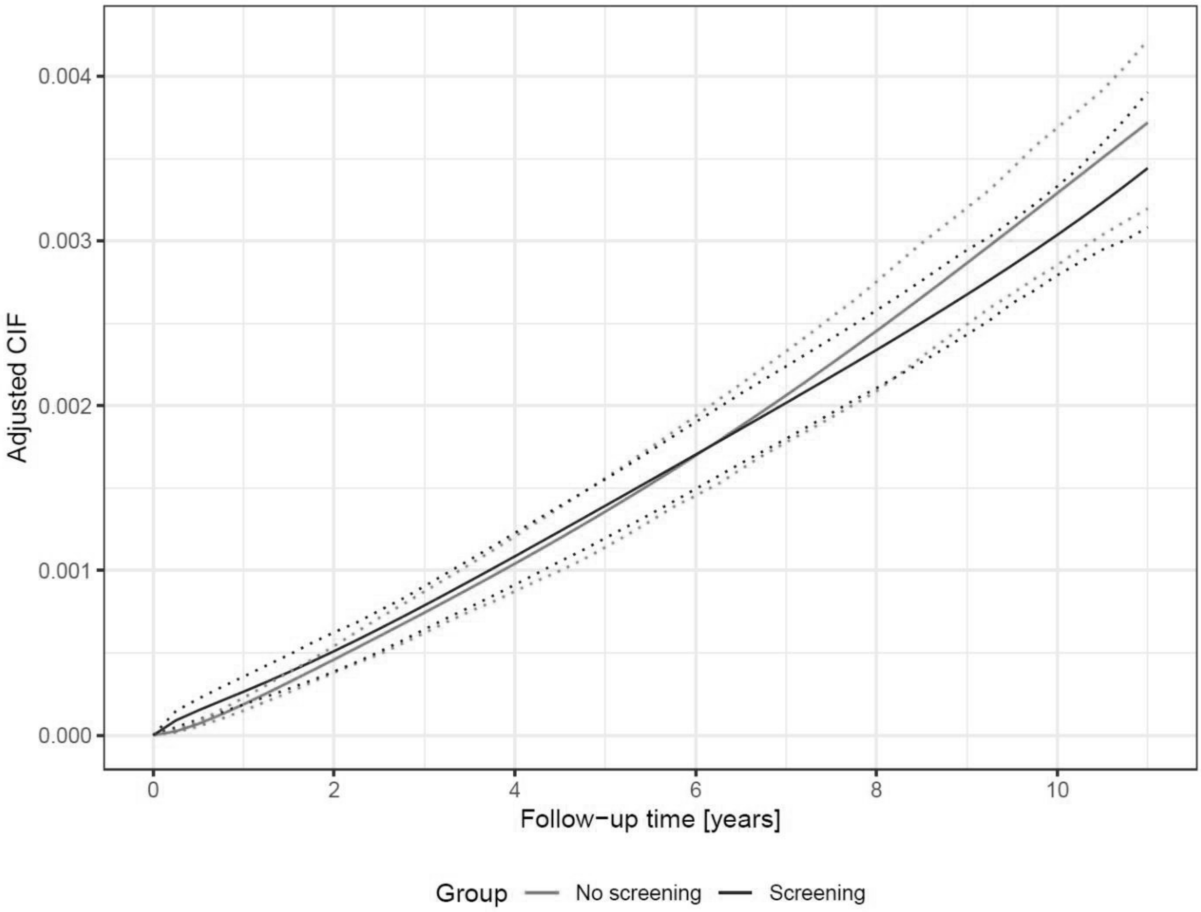


**Figure S2**: Parametric, adjusted cumulative incidence functions for incidence of pancreatic cancer in total study population, aged 55 to 69. Dashed curves represent 95% confidence intervals. The eleven-year relative risk was 0.93 (CI: 0.78-1.10)

# Supplement 7: Results of model checks

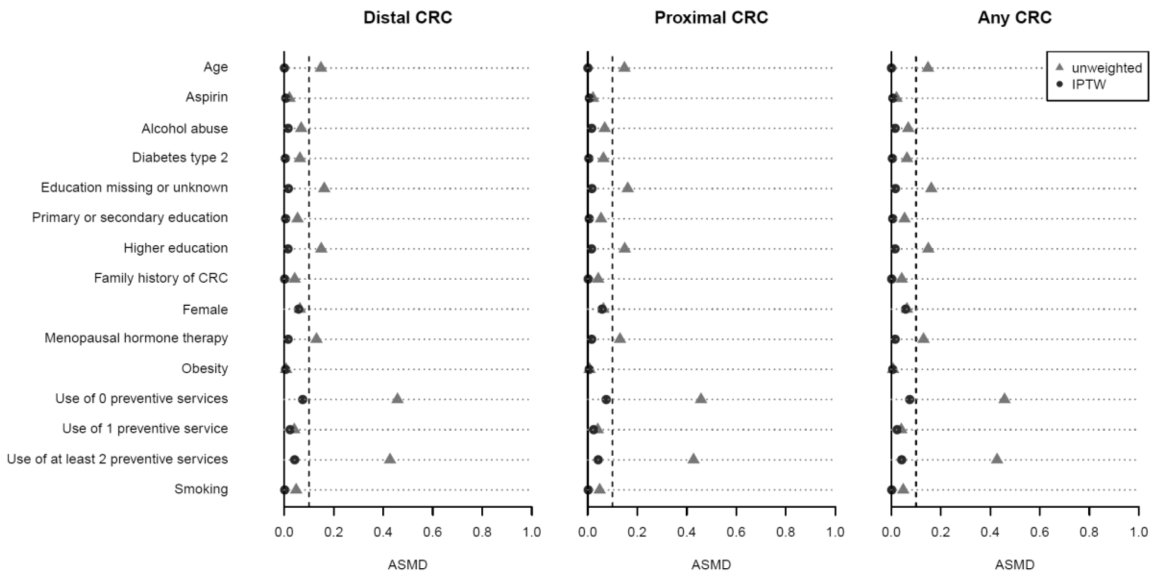


**Figure S3**: Results of covariate balance checks after IPT weighting. Covariates are plotted over absolute standardized mean differences (ASMD) before and after weighting. The vertical dashed line indicates the threshold of 0.1 commonly used to define covariate balance.

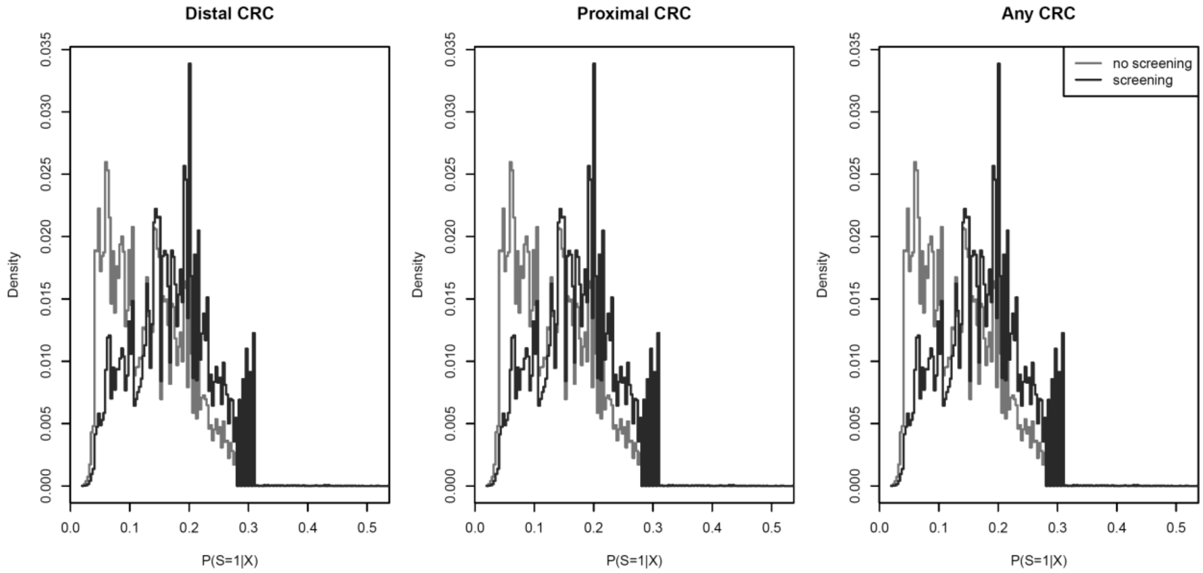


**Figure S4**: Distributions of conditional probability to receive screening (S) at baseline, given covariates X.

## Supplement 8: Results of sensitivity analyses restricted to persons aged 55 to 64 at baseline

**Table S3**: Baseline characteristics and covariates in the subpopulation aged 55 to 64. All numbers refer to non-unique persons.

| | male | | | | female | | | | total | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | screening<br>(N = 69,332) | | no screening<br>(N = 396,724) | | screening<br>(N = 71,980) | | no screening<br>(N = 435,538) | | screening<br>(N = 141,312) | | no screening<br>(N = 832,262) | |
| | n | % | n | % | n | % | n | % | n | % | n | % |
| Age | | | | | | | | | | | | |
| Median (Q1-Q3) | 59 | 56-61 | 59 | 57-62 | 58 | 56-61 | 59 | 57-62 | 58 | 56-61 | 59 | 57-62 |
| Mean (SD) | 58.9 | 2.94 | 59.1 | 2.88 | 58.6 | 2.93 | 59.1 | 2.88 | 58.7 | 2.94 | 59.1 | 2.88 |
| Education | | | | | | | | | | | | |
| no degree/unknown | 26,868 | 38.8 | 183,074 | 46.1 | 40,296 | 56.0 | 277,324 | 63.7 | 67,164 | 47.5 | 460,398 | 55.3 |
| basic or secondary degree | 17,424 | 25.1 | 102,072 | 25.7 | 19,553 | 27.2 | 102,730 | 23.6 | 36,977 | 26.2 | 204,802 | 24.6 |
| higher education | 25,040 | 36.1 | 111,578 | 28.1 | 12,131 | 16.9 | 55,484 | 12.7 | 37,171 | 26.3 | 167,062 | 20.1 |
| Region | | | | | | | | | | | | |
| East Germany | 15,011 | 21.7 | 77,362 | 19.5 | 16,257 | 22.6 | 84,055 | 19.3 | 31,268 | 22.1 | 161,417 | 19.4 |
| West Germany | 54,321 | 78.3 | 319,362 | 80.5 | 55,723 | 77.4 | 351,483 | 80.7 | 110,044 | 77.9 | 670,845 | 80.6 |
| Codes indicating obesity* | 8,628 | 12.4 | 48,181 | 12.1 | 9,832 | 13.7 | 60,831 | 14.0 | 18,460 | 13.1 | 109,012 | 13.1 |
| Diabetes type 2 | 9,192 | 13.3 | 59,822 | 15.1 | 5,470 | 7.6 | 42,598 | 9.8 | 14,662 | 10.4 | 102,420 | 12.3 |
| Codes indicating a family history of CRC | 73 | 0.1 | 98 | <.05 | 307 | 0.4 | 606 | 0.1 | 380 | 0.3 | 704 | 0.1 |
| Menopausal hormone therapy | N.A. | | N.A. | | 17,165 | 23.8 | 65,460 | 15.0 | N.A. | | N.A. | |
| Use of acetylsalicylic acid | 2,641 | 3.8 | 17,243 | 4.3 | 815 | 1.1 | 6,345 | 1.5 | 3,456 | 2.4 | 23,588 | 2.8 |
| Codes for alcohol abuse* | 2,119 | 3.1 | 19,561 | 4.9 | 1,115 | 1.5 | 9,715 | 2.2 | 3,234 | 2.3 | 29,276 | 3.5 |
| Codes for heavy smoking* | 4,145 | 6.0 | 31,499 | 7.9 | 3,823 | 5.3 | 26,277 | 6.0 | 7,968 | 5.6 | 57,776 | 6.9 |
| Use of other preventive services during 3 years before baseline** | | | | | | | | | | | | |
| None | 16,743 | 24.1 | 184,538 | 46.5 | 3,766 | 5.2 | 105,322 | 24.2 | 20,509 | 14.5 | 289,860 | 34.8 |
| One or more | 52,589 | 75.9 | 212,186 | 53.5 | 68,214 | 94.8 | 330,216 | 75.8 | 120,803 | 85.5 | 542,402 | 65.2 |

Q1-Q3: interquartile range; SD: standard deviation

* Only coded if there is a reimbursement of treatment or services due to these conditions, so not coded for all persons with the respective condition
** Used as a proxy variable for preventive behavior

**Table S4**: Number of incident CRC and effect estimates at 11 years of follow-up in the age group 55 to 64, stratified by site and sex.

| site | sex | # (non-unique) cases | | NNS | 11-year absolute risk difference | | 11-year relative risk | |
|---|---|---|---|---|---|---|---|---|
| | | screening (N = 165,464) | no screening (N = 832,262) | | % | [95% CI[a]] | | [95% CI[a]] |
| distal | male | 607 | 4,768 | | | | | |
| | female | 294 | 2,651 | | | | | |
| | total | 901 | 7,419 | 224 | 0.45 | [0.33; 0.60] | 0.64 | [0.56; 0.71] |
| proximal | male | 201 | 1,746 | | | | | |
| | female | 204 | 1,626 | | | | | |
| | total | 405 | 3,372 | 506 | 0.20 | [0.11; 0.30] | 0.66 | [0.54; 0.78] |
| both distal and proximal or unknown site | male | 74 | 683 | | | | | |
| | female | 72 | 697 | | | | | |
| | total | 146 | 1,380 | | | | | |
| total | male | 882 | 7,197 | | | | | |
| | female | 570 | 4,974 | | | | | |
| | total | 1,452 | 12,171 | 137 | 0.73 | [0.57; 0.92] | 0.64 | [0.58; 0.70] |

No effect estimates are given for both distal and proximal or unknown site and sex-specific incidence, as there were too few cases for reliable estimation.

[a]: Person-level percentile bootstrap confidence intervals based on 250 bootstrap samples.

NNS = number needed to screen

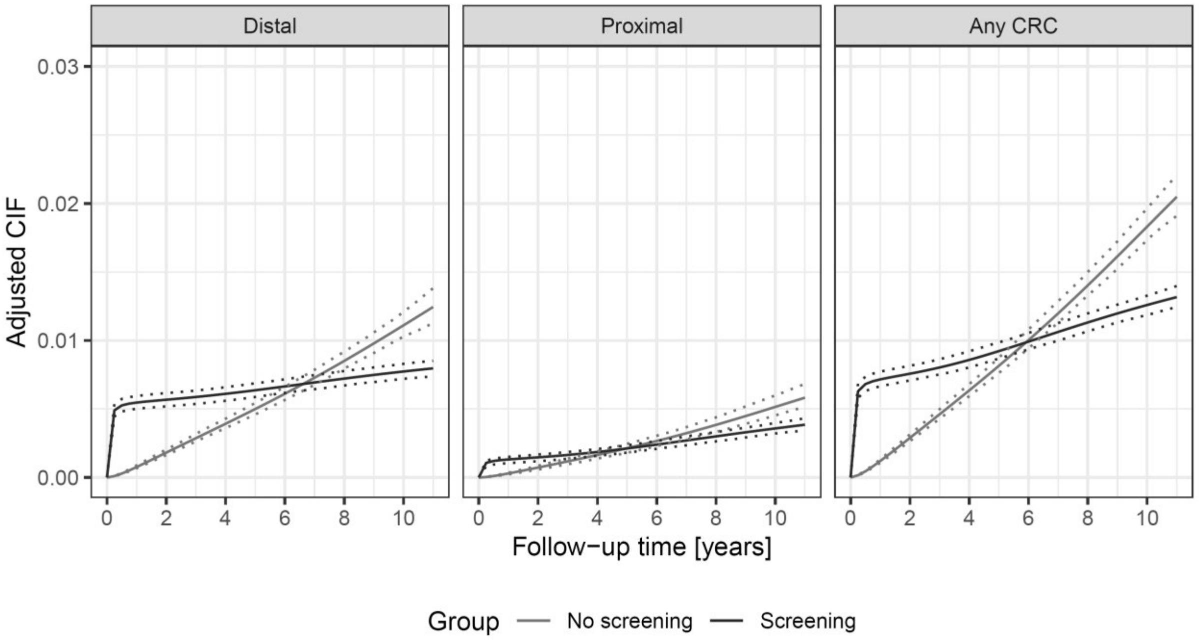


**Figure S5**: Adjusted cumulative incidence functions for the age group of 55 to 64 years showing eleven years of follow-up. Analyses were stratified by site of incident CRC. No separate analyses were done for incident CRC's of unknown location and simultaneous distal and proximal incident CRC's, since too few events were observed.

# Supplement 9: Results of unadjusted model

The below Figure S6 displays both the adjusted and the unadjusted cumulative incidence function for the incidence of any CRC in the total study population aged 55 to 69 years. The covariate adjustment led to a smaller effect size (the unadjusted eleven-year RR was 0.61), which was to be expected, given that non-screened persons tend to be less healthy and more prone to CRC than persons who opt for voluntary screening (healthy screenee bias).

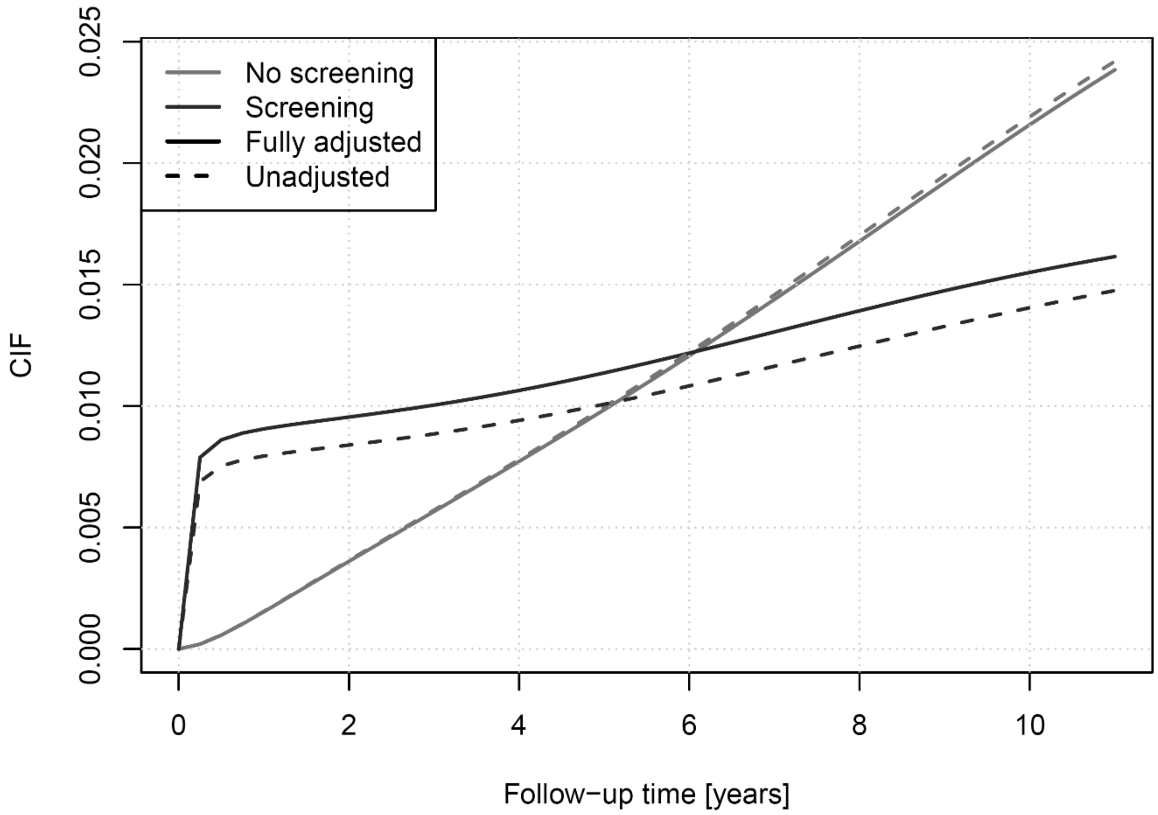


**Figure S6**: Cumulative incidence functions (CIF) of screened and non-screened persons. Solid lines indicate covariate adjusted CIFs and dashed lines indicate unadjusted CIFs.

# Supplement 10: Comparison of total and direct effects

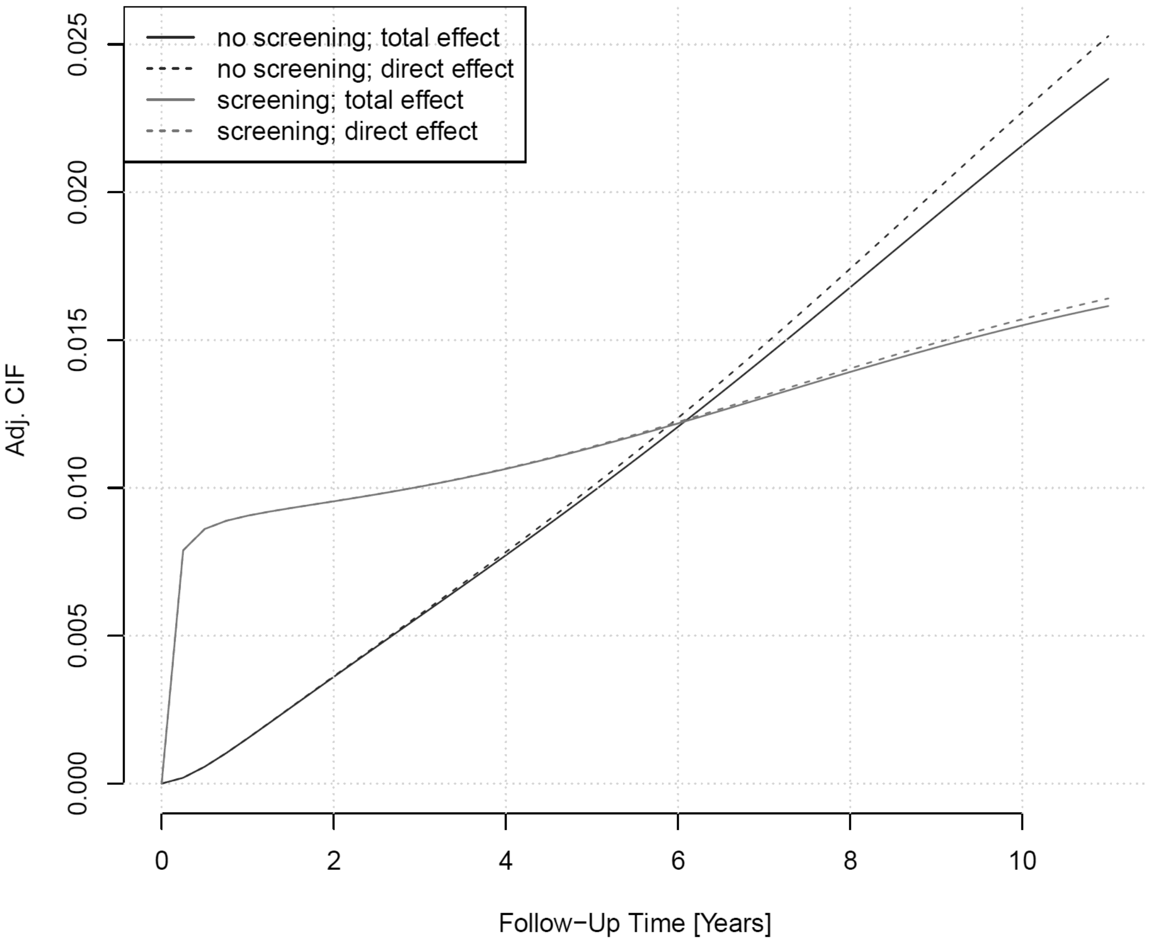


**Figure S7**: Parametric, adjusted cumulative incidence functions for incidence of any CRC in total study population, aged 55 to 69. Dashed curves represent the direct effect (i.e. under a hypothetical scenario of eliminating death as competing event) and solid lines represent the total effect as reported in the paper (i.e. allowing death as competing event).

# References

Aalen OO, Johansen S (1978) An empirical transition matrix for non-homogeneous Markov chains base on censored observations; Scandinavian Journal of Statistics; 5(3): 141 – 150

D'Agostino RB, Lee ML, Belanger AJ, Cupples LA, Anderson K, Kannel WB (1990) Relation of pooled logistic regression to time dependent Cox regression analysis: The Framingham heart study; Statistics in Medicine; 9: 1501 – 1515

Haug U, Schink T (2021) German Pharmacoepidemiological Research Database (GePaRD). In: Sturkenboom M, Schink T (editors): Databases for Pharmacoepidemiological Research; Springer Series on Epidemiology and Public Health; p. 119 – 124

Hernán MA, Robins JM (2020) Causal inference – What if; Boca Raton; Chapman & Hall/CRC

Lipsitch M, Tchetgen Tchetgen E, Cohen T (2010) Negative controls - A tool for detecting confounding and bias in observational studies; Epidemiology; 21: 383 – 388

Oppelt KA, Luttmann K, Kraywinkel K, Haug U (2019) Incidence of advanced colorectal cancer in Germany: comparing claims data and cancer registry data; BMC Medical Research Methodology; 19(1): 1 – 9

Pigeot I, Ahrens W (2008) Establishment of a pharmacoepidemiological database in Germany: methodological potential, scientific value and practical limitations; Pharmacoepidemiology and Drug Safety; 17(3): 215 – 223

Stuart EA, Lee BK, Leacy FP (2013) Prognostic score-based balance measures can be a useful diagnostic for propensity score methods in comparative effectiveness research; J Clin Epidemiol; 66(8 Suppl): S84 – S90

Young JG, Stensrud MJ, Tchetgen Tchetgen EJ, Hernán MA (2020) A causal framework for classical statistical estimands in failure-time settings with competing events; Statistics in Medicine; 39(8): 1199 – 1236

## 7.3 Paper 2: [Emulation of target trials using real-world data: a general principle to address the challenges of observational data]

This German-language overview paper was published under a CC-BY 4.0 open access license in the Journal ”Prävention und Gesundheitsförderung”. For details on how to cite the paper, refer to
https://doi.org/10.1007/s11553-022-00967-9

Präv Gesundheitsf
https://doi.org/10.1007/s11553-022-00967-9
Eingegangen: 1. April 2022
Angenommen: 6. Juli 2022

Malte Braitmaier[1] · Vanessa Didelez[1,2]
[1] Abteilung für Biometrie und EDV, Leibniz Institut für Präventionsforschung und Epidemiologie – BIPS, Bremen, Deutschland
[2] Fakultät für Mathematik und Informatik, Universität Bremen, Bremen, Deutschland

# Emulierung von „target trials“ mit Real-world-Daten

## Ein allgemeines Prinzip, um den Herausforderungen von Beobachtungsdaten zu begegnen

**Beobachtungsdaten, wie z. B. Abrechnungsdaten von Krankenkassen oder Daten von Patientenregistern, bieten eine reichhaltige Grundlage zur Beantwortung medizinischer Fragen. Während die fehlende Randomisierung oft als Schwäche genannt wird, findet es weniger Beachtung, dass Verzerrungen in der Analyse auch und v. a. durch ein unangemessenes Studiendesign bedingt sein könnten. Der Target-trial-Ansatz dient dazu, ein geeignetes Studiendesign und Auswertungskonzept zu erstellen, das den Prinzipien und dem Vorgehen einer randomisierten kontrollierten Studie („randomized controlled trial“, RCT) so ähnlich wie möglich ist und unnötige Verzerrungen vermeidet.**

## Motivation

Die „real world data“ (RWD), etwa in Form von Sekundärdaten [14, 26] wie Register-, Krankenkassendaten oder elektronischen Patientenakten, stellen eine sehr reichhaltige Informationsquelle für die medizinische und (pharmako)epidemiologische Forschung dar. Mit solchen Daten lassen sich z. B. Fragen über seltene oder späte Nebenwirkungen untersuchen oder sie können zur besseren Quantifizierung von Effekten bekannter schädlicher Substanzen herangezogen werden. Außerdem können sie wertvolle Erkenntnisse über vulnerable Personengruppen, wie etwa Schwangere, liefern. Sekundärdaten sind zudem oft die einzige Quelle, wenn es um das reale Versorgungsgeschehen geht (z. B. den Gebrauch von Medikamenten oder den Einsatz von Behandlungen). Nicht zuletzt spielt die Analyse von RWD im Vorlauf, als Information und Ergänzung für zukünftige klinische Studien eine wichtige Rolle [8].

Viele Typen von RWD werden routinemäßig und nicht zur Beantwortung bestimmter Forschungsfragen gesammelt, was im Vergleich zu RCT eine spezielle Herausforderung an die statistische Analyse und Interpretation darstellt. Hier wird oft auf die fehlende Randomisierung hingewiesen, wodurch es zu Verzerrung durch Confounding kommen kann [9]. Dies ist aber nur einer von vielen Unterschieden und vielleicht nicht einmal der wichtigste; andere Verzerrungsquellen sollten berücksichtigt und bestmöglich vermieden werden. Aus historischen Beispielen ist bekannt, dass eine naive statistische Analyse, die den vielen anderen Unterschieden zwischen Beobachtungsdaten und RCT ungenügend bzw. unangemessen Rechnung trägt, irreführende Ergebnisse liefern kann, dies aber durch ein verbessertes Studiendesign vermieden werden kann [5, 17]. Zu diesen Unterschieden zählt v. a. der bei Sekundärdaten fehlende, eindeutig ausgewiesene Startpunkt bzw. Null-Zeitpunkt („time zero“, ◘ **Abb. 1**), was in einer naiven Analyse zu Verzerrungen führen kann, z. B. aufgrund von sog. Selektionseffekten [12].

In dieser Arbeit wollen wir aufzeigen, wie durch ein geeignetes systematisches Vorgehen eine aussagekräftige und valide Analyse von Beobachtungsdaten gewährleistet werden kann. Speziell wollen wir die *Emulierung eines „target trials“* („target trial emulation“, TTE) darlegen [20]: Dies bezeichnet ein Vorgehen, das sich an einem expliziten Protokoll für eine hypothetische randomisierte Studie – dem „target trial“ – orientiert, die für die Forschungsfrage ideal wäre. Dabei werden Ein-/Ausschlusskriterien sowie zu vergleichende Behandlungsstrategien festlegt, und die Analyse von Sekundärdaten so nah wie möglich daran angelehnt (emuliert). Die Vorteile und Stärken der TTE liegen darin, dass zum einen eine klare Fragestellung am Anfang stehen muss [7] und zum anderen theoretisch sowie empirisch gezeigt werden kann, dass vermeidbare Verzerrungen durch die Anlehnung an ein „target trial“ auch tatsächlich vermieden werden können [2, 20].

## Herausforderungen bei Beobachtungsdaten: Fragestellung und Startzeitpunkt

Bei der Analyse von Beobachtungsdaten liegt es nahe, Fragestellungen zu formulieren, die den Effekt (oder Nebenwir-

Published online: 29 July 2022

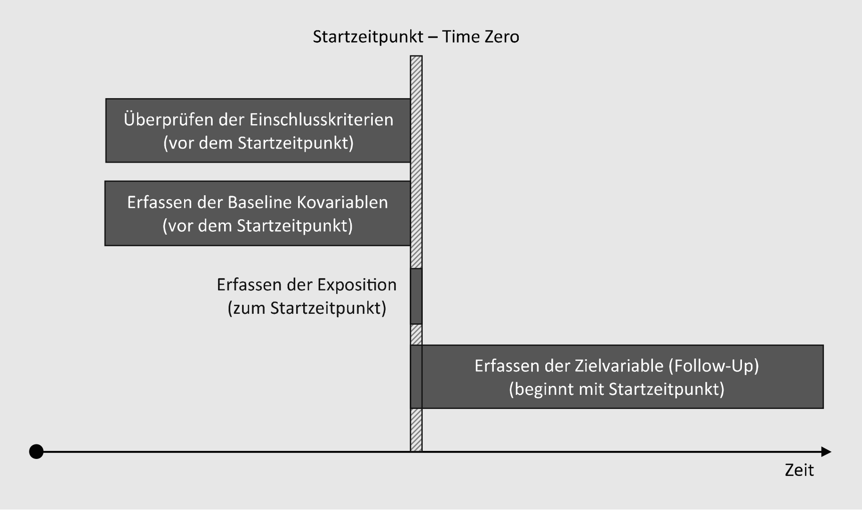

Startzeitpunkt – Time Zero
Überprüfen der Einschlusskriterien (vor dem Startzeitpunkt)
Erfassen der Baseline Kovariablen (vor dem Startzeitpunkt)
Erfassen der Exposition (zum Startzeitpunkt)
Erfassen der Zielvariable (Follow-Up) (beginnt mit Startzeitpunkt)
Zeit

**Abb. 1** ▲ Zentrale, zeitliche Elemente des Studiendesigns müssen korrekt am Startzeitpunkt der Studie ausgerichtet sein

kung) einer Behandlung/Therapie betreffen; als konkretes Beispiel betrachten wir den Effekt von Screeningkoloskopien auf die Darmkrebsinzidenz [4]. Dies allein ist noch keine eindeutige Fragestellung [16]: Möchte man etwa den Unterschied in der Gesamtpopulation schätzen, wenn der Gesamtpopulation gar keine Screeningkoloskopie angeboten würde, also niemand sie in Anspruch nehmen könnte? Oder möchte man für eine individuelle Person im Alter von 55 Jahren wissen, was der erwartete Nutzen davon wäre, sich sofort, erst in 5 Jahren oder jetzt und in 10 Jahren, anstatt niemals einer Koloskopie zu unterziehen? Dies sind unterschiedliche Fragen, die jeweils andere Studiendesigns und Herangehensweisen erfordern. Eine klare Fragestellung an den Anfang zu stellen, scheint eine Binsenweisheit, ist aber gerade bei Beobachtungsdaten aufgrund vieler verschiedener Optionen und zeitlicher Aspekte nicht selbstverständlich.

Eine Herausforderung bei RWD stellt, wie schon erwähnt, die Festlegung des Startzeitpunktes dar, der nicht wie bei einem RCT automatisch gegeben ist; dies kann u. a. zu „immortal time bias“ führen [30]. Wenn eine Person sich erst einige Zeit nach dem impliziten oder expliziten Startzeitpunkt der Behandlung unterzieht, trotzdem aber von Anfang an dem Behandlungsarm zugewiesen wird, muss diese Person per Definition des Behandlungsarms bis dahin überlebt haben. Es werden also Personen, die schon länger überlebt haben müssen, in den Behandlungsarm selektiert. Wenn die Kontrollgruppe keine Behandlung bekommt, wirkt dieser Selektionseffekt hier nicht, d.h. es kommt zu einer Anreicherung von Personen im Kontrollarm, die kurze Zeit nach dem Startzeitpunkt versterben. Der Behandlungseffekt auf das Überleben wird daher überschätzt [31]. Neben dem „immortal time bias“ können auch viele andere Selektionseffekte durch Unklarheit über den Startpunkt oder anderer Nichtbeachtung zeitlicher Effekte entstehen, so wie beispielsweise der „prevalent user bias“ [27]. Eine korrekte Analyse setzt voraus, dass die Überprüfung der Einschlusskriterien zum Startzeitpunkt abgeschlossen ist – es darf nicht „in die Zukunft geschaut“ werden (◘ **Abb. 1**). Es folgt dann unmittelbar die Zuteilung zu den Behandlungsarmen, für deren Definition keine Informationen von vor dem Startzeitpunkt verwendet werden darf. Gleichzeitig beginnt der Follow-up, sodass unmittelbar nach der Zuteilung zu den Behandlungsarmen mit der Erfassung der Zielvariable in den Behandlungsarmen begonnen wird [12]. Diese zeitliche Strukturierung einer Beobachtungsstudie ist in vielen Fällen nicht trivial, beispielsweise wenn der Kontrollarm keine Behandlung erhalten soll: Wann wäre der Startzeitpunkt für den Kontrollarm, der keine Behandlung erhält?

## Prinzipien der Target-trial-Emulierung

Die grundlegende Idee des Target-trial-Ansatzes ist, sich die Stärken des Studiendesigns von RCT zu eigen zu machen. Es gibt eine klare Fragestellung, eine realistische und relevante Intervention und eine zeitlich sinnvolle Anordnung, was die Überprüfung der Einschlusskriterien, die Behandlungszuweisung und den Beginn des Follow-up betrifft. Der erste Schritt einer TTE besteht darin, die Forschungsfrage zu formalisieren und zwar in Form von konkreten Behandlungsstrategien, die es zu vergleichen gilt. Die jeweiligen Behandlungsstrategien bilden dann den Ausgangspunkt für das Studienprotokoll des idealen, hypothetischen RCT – dem „target trial“. Dieser Arbeitsschritt hilft automatisch dabei, dass die Analyse so konkret und praktisch relevant wie möglich wird und eine klare Interpretation erlaubt. Nach der Festlegung des „target trials“ wird eine Beobachtungsstudie, die Emulierung, aufgesetzt; die hierfür relevanten Punkte sind in ◘ **Tab. 1** beschrieben.

Wie oben beschrieben erfolgt die Zuweisung von Personen zu den Behandlungsarmen dann anhand der zum Startzeitpunkt beobachteten Exposition. Wir betrachten wieder als Beispiel die Effektivität von Screeningkoloskopien hinsichtlich der Senkung der Darmkrebsinzidenz [4]. Im „target trial“ wird eine Person dem Behandlungsarm zugewiesen, wenn sie sich in dem Zeitfenster, das als „time zero“ festgelegt wurde, einer Screeningkoloskopie unterzieht und dem Kontrollarm, wenn sie sich in diesem Zeitfenster keiner Screeningkoloskopie unterzieht. Da durch das Studienprotokoll die Screeningstrategien klar definiert werden, ist auch unmittelbar klar, was die kausale Forschungsfrage ist: Führt eine Screeningkoloskopie zum Startzeitpunkt im Vergleich zu keiner durchgeführten Screeningkoloskopie zu einer Verringerung der Darmkrebsinzidenz im Follow-up [4]? Man beachte, dass im Kontrollarm spätere Koloskopien stattfinden können – der kausale Effekt entspricht in diesem Fall also in etwa einem Intention-to-screen-Effekt mit kompletter Adhärenz zum Start-

Präv Gesundheitsf https://doi.org/10.1007/s11553-022-00967-9

M. Braitmaier · V. Didelez

## Emulierung von „target trials“ mit Real-world-Daten. Ein allgemeines Prinzip, um den Herausforderungen von Beobachtungsdaten zu begegnen

### Zusammenfassung

**Hintergrund.** Die „real world data“ (RWD), z. B. Krankenkassendaten, bieten reichhaltige Informationen zu gesundheitsrelevanten Faktoren und können die Basis für Studien zur Arzneimittelsicherheit, Wirksamkeit medizinischer Interventionen u. v. m. darstellen. Ein besonderer Vorteil ist die je nach Datenquelle größere Verallgemeinerbarkeit, wenn z. B. Informationen zu bestimmten Subgruppen der Population vorliegen und ein Volunteer-Bias ausgeschlossen werden kann. Gerade in Fällen, in denen randomisierte kontrollierte Studien („randomized controlled trials“, RCT) nicht durchgeführt werden können, sind Beobachtungsstudien basierend auf RWD eine wichtige Informationsquelle. Die valide Analyse von RWD stellt allerdings einige Herausforderung dar, wobei insbesondere mögliche Verzerrungen, die durch ein sorgfältiges Studiendesign vermeidbar wären, Beachtung finden sollen. Hier setzt das Prinzip der Target-trial-Emulierung (TTE) an.

**Ziel der Arbeit.** In diesem Artikel soll aufgezeigt werden, wie die TTE den Herausforderungen bei der Analyse von RWD begegnet.

**Material und Methoden.** Die TTE wird allgemein verständlich vorgestellt. Prinzipien, Vorteile, Annahmen und spezifische statistische Aspekte werden anhand relevanter Literatur und praktischer Beispiele erläutert.

**Ergebnisse.** Damit die Analyse von RWD valide, kausal interpretierbare Ergebnisse liefern kann, müssen einige Bedingungen erfüllt sein. Neben einem ausreichenden Informationsgehalt der Daten sind auch eine klare Fragestellung und ein geeignetes Studiendesign, das u. a. Selektionseffekte vermeidet, von zentraler Bedeutung. Das Target-trial-Prinzip besteht darin, dass zunächst das Auswertungskonzept für einen RCT erarbeitet wird, welches in einem zweiten Schritt mit Beobachtungsdaten „emuliert“ wird. Somit liefert die TTE quasi eine Anleitung, um die Fragestellung zu definieren und ein geeignetes Studiendesign zu entwerfen. TTE kann mit unterschiedlichen statistischen Methoden kombiniert werden, wobei statistische Effizienz durch sequenzielle Trials und das sog. Klonen gewonnen werden kann.

**Schlussfolgerung.** Die TTE ist ein allgemeines und übergreifendes Prinzip, das zentralen Herausforderungen bei der Analyse von Beobachtungsdaten, also auch RWD, systematisch begegnet.

**Schlüsselwörter**
Selektionsbias · Beobachtungsstudien · Kausales Schlussfolgern · Gesundheitsdaten · Confounding

## Emulation of target trials using real-world data. A general principle to address the challenges of observational data

### Abstract

**Background.** Real world data (RWD), e.g., claims data, are a rich source of information regarding health-related factors and can be the basis for observational studies examining the safety of medicines or effectiveness of medical interventions, among others. A special feature of these studies is the high generalizability, due to the fact that, potentially, information on various subgroups of a population is available in the data source. Also, these studies contribute important answers to relevant health questions in cases where randomized controlled trials (RCT) cannot be conducted. Valid analyses of RWD, however, pose several challenges regarding possible biases. Here, we focus on target trial emulation (TTE) as a guide to avoid potential biases by a careful study design.

**Objectives.** This article will showcase how TTE can address certain challenges of RWD studies.

**Materials and methods.** TTE is introduced in an understandable way, using relevant literature and practical examples to explain principles, advantages, assumptions, and specific statistical aspects.

**Results.** Several conditions must be met in order for observational studies to yield valid causal inferences. Besides sufficiently informative data, a clear research question and a suitable study design must be chosen to avoid, for instance, selection effects. The core idea of the target trial principle is to first set out the analysis plan for an RCT that would answer the research question and, in a second step, emulate it using observational data. TTE, therefore, serves as a guide to defining both the research question and study design. Various statistical methods can be incorporated in TTE and statistical efficiency can be gained by sequential trials and so-called cloning of study participants.

**Conclusions.** TTE is a general and comprehensive approach to address the central challenges posed by the analysis of observational data, including RWD.

**Keywords**
Selection bias · Observational studies · Causal inference · Electronic health records · Confounding

zeitpunkt Baseline, d. h. dem Effekt der Zuweisung zu den Behandlungsarmen ohne Spezifizierung für das weitere Follow-up. Im Gegensatz zu diesem kausal interpretierbaren Studienansatz haben mehrere Beobachtungsstudien zur selben Fragestellung in der Vergangenheit ein Studiendesign gewählt, das Selektionsbias nicht ausschließen kann. Ein hypothetisches Beispiel ist das Erfassen der Screeningkoloskopie durch einen Fragebogen beim Startzeitpunkt. Personen, die in der Vergangenheit eine Screeningkoloskopie hatten, werden als exponiert gezählt. Gleichzeitig werden Personen mit prävalentem Darmkrebs ausgeschlossen. Da Koloskopien aber verwendet werden, um Darmkrebs zu diagnostizieren, kommt es unter Exponierten allein durch diese Definitionen von Exposition und Ausschlusskriteri-

**Tab. 1** Komponenten des Studienprotokolls eines „target trials" angelehnt an Hernán und Robins [20]

| Komponente | Beschreibung und Emulierung |
|---|---|
| Ziel der Studie | Definition der Forschungsfrage |
| Einschlusskriterien | In der emulierten Studie können zusätzliche Einschlusskriterien nötig werden, wie beispielsweise, dass eine gewisse Zeitspanne vor dem Startzeitpunkt in den Daten abgebildet sein muss oder dass Informationen zu Geschlecht oder Alter nicht fehlen dürfen |
| Behandlungsstrategien | Möglichst präzise Festlegung der Behandlungsstrategien. Handelt es sich z. B. um eine einmalige Behandlung, oder wird die Behandlung über eine gewisse Zeit hinweg erhalten? Gibt es bestimmte Ereignisse, die eine Änderung der Behandlung erfordern (z. B. Anpassung der Dosis ab einem bestimmten Laborwert)? Nur wenn die Strategien bis ins Detail definiert wurden, können sie valide mit den Beobachtungsdaten emuliert werden |
| Behandlungszuweisung | In einem RCT würde die Behandlungszuweisung durch eine Randomisierung vorgenommen werden. Es ist an dieser Stelle nicht wichtig, den genauen Randomisierungsprozess zu beschreiben (z. B. Blockrandomisierung o. ä.). Vielmehr muss im nächsten Schritt definiert werden, wie die Behandlungszuweisung in der Beobachtungsstudie vorgenommen wird (z. B. Behandlungsbeginn innerhalb der ersten Woche nach Startzeitpunkt). Es wird dann für die Beobachtungsstudie auch definiert, welche Kovariablen für die Adjustierung berücksichtigt werden müssen, um die Randomisierung zu emulieren |
| Follow-up | Genaue Festlegung: wann beginnt/endet der Follow-up? |
| Zielvariable | Kann die Zielvariable in den Beobachtungsdaten verlässlich abgebildet werden? |
| Kontrast | Wie werden die Behandlungsstrategien miteinander verglichen? Wie wird mit Nicht-Adhärenz umgegangen („intention-to-treat" ohne Adjustierung vs. Per-Protokoll mit Adjustierung für Nicht-Adhärenz)?<br>Können die Kontraste in der Emulierung abgebildet werden? Beispielsweise entspricht ein Intention-to-treat-Effekt der Emulierung eher einem Treatment-initiation-Effekt eines RCT |
| Statistische Analyse | Es muss ein besonderes Augenmerk auf Nicht-Adhärenz im Follow-up gelegt werden. Soll ein Per-protocol-Effekt emuliert werden, müssen Studienteilnehmer:innen künstlich zensiert werden, sobald sie gegen die zugewiesene Behandlungsstrategie verstoßen. Da diese Zensierung einen Selektionsbias verursachen kann, muss außerdem festgelegt werden, für welche Variablen im Follow-up adjustiert werden muss |

*RCT* „randomized controlled trial"

um zu einer verringerten Anzahl an Fällen und somit zu einer Überschätzung des Effekts der Koloskopie auf die Darmkrebsinzidenz. Dies entspricht dem „prevalent user bias". Durch die TTE werden solche Selektionseffekte vermieden, da sowohl die Überprüfung der Einschlusskriterien, als auch die Erfassung der Exposition und der Start des Follow-up zum selben Zeitpunkt stattfinden (◘ **Abb. 1**; [12]).

Die ◘ **Tab. 2** stellt einen Vergleich der möglichen Limitationen von RCT und Beobachtungsstudien mit und ohne TTE dar. Als Beispiele für Verzerrungen, die durch eine zeitliche Trennung zentraler Elemente des Studiendesigns entstehen, sind hier „prevalent user bias" und „immortal time bias" aufgeführt. Prinzipiell kann es aber immer zu Verzerrungen kommen, wenn die in ◘ **Abb. 1** aufgeführte Ausrichtung am Startzeitpunkt verletzt wird. Die Limitationen von Beobachtungsstudien aus ◘ **Tab. 2** können durch ein sorgfältiges Studiendesign auch ohne explizite Definition eines „target trials" umgangen werden, wenn man sich der jeweiligen Fehlerquellen bewusst ist. Der Vorteil des TTE-Ansatzes ist aber, dass diese Verzerrungen gar nicht erst entstehen können und damit auch versteckte bzw. weniger offensichtliche Fehlerquellen vermieden werden.

Wir möchten an dieser Stelle auch darauf verweisen, dass Randomisierung und RWD nicht unvereinbar sind (siehe z. B. [13]). Da RWD aber hauptsächlich für (nicht-randomisierte) Beobachtungsstudien verwendet werden, gehen wir in diesem Artikel nicht näher auf diese Schnittstelle ein.

## Sequenzielle Trials zur Effizienzsteigerung

Ein Unterschied zwischen dem „target trial" und der tatsächlich durchgeführten Beobachtungsstudie ist, dass im „target trial" unmittelbar klar ist, was der Startzeitpunkt ist. Personen werden rekrutiert und zum Startzeitpunkt zufällig einem Behandlungsarm zugewiesen. Auch in der entsprechenden Beobachtungsstudie könnte ein Startzeitpunkt an einem bestimmten Datum angesetzt werden. In diesem Fall hätten Personen nur zu diesem Zeitpunkt die Möglichkeit, in die Studie eingeschlossen zu werden. Wenn allerdings eine longitudinale Datenbank vorliegt, gibt es ein aus statistischer Sicht effizienteres Vorgehen: Es werden mehrere „target trials" hintereinander emuliert, beispielsweise einmal pro Quartal in einer gegebenen Zeitspanne. In der Beispielstudie zur Screeningkoloskopie wurde zu Beginn jedes Quartals von 2007 bis 2011 ein „target trial" emuliert, wobei alle Personen, die zum Zeitpunkt des jeweiligen Startzeitpunktes in der Datenquelle abgebildet sind, für den Trial berücksichtigt wurden [4]. Wenn also eine Person bei mehreren emulierten Trials in den Daten vorkommt und die Einschlusskriterien erfüllt, kann diese Person in mehreren Trials eingeschlossen werden. Da die Daten all dieser emulierten, sequenziellen Trials gemeinsam ausgewertet werden, kommt diese Person dann u. U. mehrmals im Analysedatensatz vor. Man spricht in diesem Zusammenhang auch von Klonen.

Dieses sequentielle Studiendesign kann als Form eines longitudinalen Matchings angesehen werden: Zu jedem Zeitpunkt, an dem ein emulierter Trial beginnt, werden exponierte und nicht-exponierte Personen daraufhin gematcht, dass beide zu diesem Zeitpunkt die gleichen Einschlusskriterien erfüllen [33]. Wie in anderen, in der Epidemiologie verbreiteten Matching-Ansätzen kann dann dieselbe Person mehrmals im Kontrollarm oder auch zunächst im Kontrollarm und später als exponierte Person in die Studie eingehen [32]. Dies muss beim Schätzen von Konfidenzintervallen berücksichtigt werden,

**Tab. 2** Übersicht von Limitationen unterschiedlicher Ansätze

| Risiko für Verzerrung durch: | RCT | Beobachtungsstudien mit RWD | |
|---|---|---|---|
| | | Mit TTE | Ohne TTE |
| „Prevalent user bias" [27] | Gering,<br>Behandlung beginnt mit Randomisierung | Gering,<br>durch Ausrichtung am Startzeitpunkt vermieden | Hoch,<br>wenn etwa zur Bestimmung der Exposition/Behandlung Informationen aus der Vergangenheit verwendet werden |
| „Immortal time bias" [30] | Gering,<br>Behandlungsgruppen werden bei Randomisierung zugewiesen | Gering,<br>durch Ausrichtung am Startzeitpunkt vermieden | Hoch,<br>wenn etwa zur Bestimmung der Exposition/Behandlung Informationen aus dem Follow-up, also aus der Zukunft, verwendet werden |
| Unklare Forschungsfrage [7] | Bei beiden Ansätzen gering,<br>durch Definition der zu vergleichenden Behandlungsstrategien | | Hoch,<br>wenn Expositionen etwa nicht eindeutig als hypothetische Interventionen definiert werden |
| Confounding durch Baseline-Variablen | Gering,<br>wird durch Randomisierung vermieden | Bei beiden Ansätzen hoch,<br>kann aber durch statistische Adjustierung behoben werden, wenn Daten hinreichend informativ | |
| Zeitabhängiges Confounding bei Per-protocol-Analysen [18, 21] | Bei allen Ansätzen hoch,<br>wenn etwa Behandlungen abgebrochen oder gewechselt werden (Nicht-Adhärenz);<br>kann aber durch statistische Adjustierung behoben werden (z. B. durch „inverse probability weighting"), wenn Daten hinreichend informativ | | |
| Mangelnde externe Validität | Hoch,<br>stark selektierte Studienpopulation unterscheidet sich u. U. stark von Zielpopulation | Abhängig von der Datenquelle<br>– Register-/Krankenkassendaten können gerade für Subgruppen informativ sein, die in RCT typischerweise ausgeschlossen werden, sowie etwa für das reale Versorgungsgeschehen.<br>– „volunteer bias" u. ä. muss bei anderen Datenquellen bedacht werden | |
| Kosten- und zeitintensiv | Hoch | Bei beiden Ansätzen gering, wenn vorhandene Daten genutzt werden können | |

*RCT* „randomized controlled trial", *RWD* „real world data", *TTE* Target-trial-Emulierung

beispielsweise durch robuste Bootstrap-Methoden [19].

## Klonen und künstliches Zensieren

Wie oben beschrieben, können mehrere Klone oder Kopien einer Person in die Studie eingeschlossen werden, wenn diese Person zu mehreren Zeitpunkten die Einschlusskriterien erfüllt. Eine weitere Form des Klonens kann sich innerhalb eines einzelnen Trials abspielen, wenn die tatsächliche Behandlung von Personen am Startzeitpunkt mit mehreren Behandlungsarmen konform ist. Beispielsweise wird in der ZEBra-Studie die Wirksamkeit des deutschen Mammographie-Screening-Programms (MSP) hinsichtlich Senkung der Brustkrebsmortalität untersucht [3, 22]. Hierbei können u. a. die drei folgenden Strategien definiert werden: 1) niemals Screening, 2) Screening mindestens zum Startzeitpunkt, 3) Screening zum Startzeitpunkt und danach regelmäßig in 2-Jahres-Abständen. Werden nun Trials pro Kalenderquartal emuliert, kann eine Frau, die sich im Anfangsquartal einer Screeningmammographie unterzieht, sowohl Strategie 2 als auch Strategie 3 zugewiesen werden. Anstatt sie zufällig einer der beiden Strategien zuzuweisen, werden ihre Daten „geklont" und beiden Strategien zugewiesen.

Im obigen Beispiel sind also die Populationen in den beiden aktiven Studienarmen zum Startzeitpunkt identisch. Personen werden dann aber während des Follow-up künstlich zensiert (d. h. ihr Follow-up wird künstlich beendet; [23]), sobald ihre beobachtete Behandlung von einer zugewiesenen Strategie abweicht. Im Mammographiebeispiel würde also eine Frau aus der Strategie „Screening zum Startzeitpunkt und danach regelmäßig" zensiert werden, falls sie nach dem vorgesehenen Intervall nicht erneut beim Screeningtermin erscheint. Hierbei sind Ausnahmen genau wie beim „target trial" zu berücksichtigen, z. B. wenn sie zwischenzeitlich eine Brustkrebsdiagnose erhält oder ein Alter erreicht, in dem Screeningmammographien nicht mehr vorgesehen sind [3].

Es ist wichtig zu beachten, dass dieses künstliche Zensieren eine Selektion darstellt, der die statistische Analyse Rechnung tragen muss. Wenn beispielsweise ein Faktor (wie etwa eine gesundheitliche Vorbelastung) sowohl die Abweichung von einer zugewiesenen Strategie (wie etwa Behandlungsabbruch) als auch die Zielvariable beeinflusst, könnte es dazu kommen, dass im Behandlungsarm Personen, bei denen dieser Faktor vorliegt, unterrepräsentiert sind. Dies würde wiederrum die Zielvariable im Behandlungsarm und damit das Ergebnis verfälschen. Das künstliche Zensieren muss daher durch eine geeignete Gewichtung, die solche Faktoren berücksichtig, ausgeglichen werden, ähnlich wie bei anderen Formen der informativen Zensierung. Diese Gewichtung erfolgt beispielsweise mit Hilfe von zeitveränderlichen Propensity Scores (PS) durch das sog. „inverse probability of censoring weighting" (IPCW; [19]).

## Annahmen und Voraussetzungen

In vielen Beobachtungsstudien wird ein starkes Augenmerk auf das Problem des Confounding gelegt. Da es keine Randomisierung gibt, können sich die Vergleichsgruppen hinsichtlich prognostischer Faktoren systematisch unterscheiden, wodurch es zu Verzerrungen beim Schätzen von Effekten kommt. Um diese Verzerrung auszugleichen, müssen die prognostischen Faktoren beobachtet sein und angemessen in das statistische Analysemodell einfließen. In dieser Arbeit möchten wir aber auf weitere Annahmen aufmerksam machen, die für alle Arten von Beobachtungsstudien erfüllt sein müssen, wenn das Ziel eine kausale Aussage ist, die aber oftmals zu wenig beachtet werden [15]. Wir beginnen wieder mit der Forschungsfrage. Diese sollte einer eindeutigen und realistischen Intervention entsprechen, die auch tatsächlich in den Daten beobachtbar ist. Es wäre z. B. problematisch, eine Intervention zu betrachten, die den Body Mass Index (BMI) verändert, ohne dabei klarzustellen, ob dies durch eine Veränderung der Ernährung, der körperlichen Aktivität, durch einen chirurgischen Eingriff oder Kombinationen dieser Faktoren geschieht, denn die Effekte der jeweiligen Intervention könnten sehr unterschiedlich ausfallen. Diese Annahme wird auch „(causal) consistency" genannt und soll im „target trial" durch das explizite Formulieren realistischer Behandlungsstrategien erfüllt werden. Ein wichtiger Aspekt ist auch die Positivitätsannahme [19]. In einem RCT kann jede teilnehmende Person durch die Randomisierung jeder Strategie zugeordnet werden – dies muss auch bei der Emulierung sichergestellt sein: Wenn die Einschlusskriterien erfüllt sind, muss es für jede Subgruppe Individuen in jedem Behandlungsarm geben (also einen positiven Anteil). Dies lässt sich leicht empirisch überprüfen [34]. Wenn die Positivität empirisch nicht gilt, kann dies zwar an einem zu kleinen Stichprobenumfang liegen, aber auch daran, dass die Einschlusskriterien ungeeignet oder die Behandlungsstrategien unrealistisch sind, was sich durch eine Anpassung der Forschungsfrage oder des TTE-Protokolls beheben lässt.

Grob lässt sich sagen, dass Annahmen immer dann ins Spiel kommen, wenn eine Abweichung der Emulierung vom „target trial" besteht. So wird z. B. im „target trial" die Zuordnung zu den Behandlungsarmen zufällig (randomisiert) stattfinden, was in der Emulierung mit Beobachtungsdaten durch Adjustierung für Confounding ersetzt werden muss – dafür ist die Annahme nötig, dass kein ungemessenes Confounding existiert. Eine erfolgreiche Adjustierung für gemessenes Confounding lässt sich z. B. anhand von „balance checks" überprüfen [29]. Allerdings ist dies nicht für ungemessenes Confounding möglich – hier lassen sich stattdessen teilweise quantitative Biasanalysen durchführen [24]. Zudem ist zu beachten, dass auch im RCT nur anfangs randomisiert wird; bei Nicht-Adhärenz ist es auch dann u. U. nötig, für zeitabhängiges Confounding zu adjustieren, so wie auch bei der TTE, wenn die zu vergleichenden Behandlungsstrategien über längere Zeit andauern. Neben dem künstlichen Zensieren in Verknüpfung mit entsprechenden Gewichten gibt es eine Reihe von alternativen statistischen Verfahren zur Berücksichtigung von zeitabhängigem Confounding [19, 28].

Aus den obigen Annahmen folgt, dass die Datengrundlage bestimmte Kriterien erfüllen muss: 1) Es müssen ausreichend Informationen zu relevanten Kovariablen vorliegen, zumindest in Form von Proxyvariablen. Diese Voraussetzung haben alle Beobachtungsstudien gemein. 2) Die relevanten Informationen in der Datenbank müssen über einen längeren Zeitraum ohne große Unterbrechungen und jeweils mit relativ präzisen Datumsangaben vorliegen. Dies ist beispielsweise in Forschungsdatenbanken mit Versichertendaten – wie der pharmakoepidemiologischen Forschungsdatenbank GePaRD (German Pharmacoepidemiological Research Database; [14]) – erfüllt. Allerdings kann es bei anderen RWD, beispielsweise aus Kohortenstudien, in denen Primärdaten im Abstand mehrerer Jahre erhoben werden, schwieriger sein. Aber auch in Fällen, in denen die Datengrundlage nicht für die TTE ausreicht, ist das Erstellen des Target-trial-Protokolls sinnvoll: Es macht transparent, welche Aspekte mit den verfügbaren Daten nicht abgebildet werden können und welche Auswirkungen dies auf die Studienergebnisse hat [7]. Beides ist zur Bewertung der Ergebnisse oder auch für die Planung von Sensitivitätsanalysen und zukünftiger Studien wichtig.

## Funktioniert TTE in der Praxis?

Um zu untersuchen, ob der Target-trial-Ansatz tatsächlich die Analyse von RWD systematisch verbessert, bieten sich Fragestellungen an, die sowohl in Beobachtungsstudien als auch in RCT untersucht wurden. Beispielsweise wurden unterschiedliche Analysen zum Effekt von Statinen auf das Risiko, an Krebs zu erkranken, miteinander verglichen [5]. In früheren Beobachtungsstudien wurde ein stark protektiver Effekt von Statinen auf das Krebsrisiko beschrieben, ohne dass es eine klare biologische Erklärung dafür gab. Spätere RCT konnten diesen scheinbar protektiven Effekt allerdings nicht replizieren. Mit einer TTE und Daten aus der Clinical Practice Research Database (CPRD) wurden schließlich folgende explizite Behandlungsstrategien verglichen [5]: (i) Statin-Therapie zu einem bestimmten Zeitpunkt starten und über den Follow-up beibehalten, es sei denn, dass sich eine Kontraindikation entwickelt; (ii) keine Statin-Therapie zum Startzeitpunkt und im Follow-up nur dann, wenn sich eine Indikation dazu entwickelt. Diese TTE fand keinen nennenswerten Effekt von Statin-Therapie auf die Krebsinzidenz. Die Designschwächen früherer Beobachtungsstudien waren, dass ein „prevalent user design" gewählt wurde, und eine Einteilung in den Behandlungsarm nur erfolgte, wenn die Therapie in den ersten 4 Follow-up-Jahren aufrechterhalten wurde. Als dieses fehlerhafte Design mit denselben CPRD-Daten implementiert wurde, ergab sich wieder ein ähnlicher, scheinbar protektiver Effekt. Dieser irreführende protektive Effekt ist also nicht durch unbeobachtetes Confounding, sondern durch ein mangelhaftes Studiendesign entstanden, was sich durch ein

TTE vermeiden ließ. In der Literatur sind mehrere Beispiele beschrieben, in denen erst durch TTE plausible Ergebnisse erzielt wurden, die mit dem Wissen aus RCTs bzw. sonstigem Wissen konsistent waren und die Ergebnisse früherer Beobachtungsstudien in Frage gestellt haben [1, 5, 10–12, 17].

Zwar kann das Target-trial-Prinzip nicht garantieren, dass es in einer Studie kein ungemessenes Confounding gibt, aber fast alle anderen Annahmen lassen sich zumindest nach geeigneter Anpassung des Protokolls erfüllen, bzw. man kann explizit begründen, dass sie bei gegebener Datensituation so gut wie möglich approximiert werden. Allerdings gibt es andere gute Gründe, warum die Ergebnisse aus Beobachtungsstudien nicht unbedingt den Ergebnissen aus RCT entsprechen. Eine Schwäche von klinischen Studien ist, dass die Studienpopulation stark eingeschränkt ist und bestimmte Subgruppen aus der Gesamtbevölkerung nicht vertreten oder unterrepräsentiert sind (z. B. Schwangere, Ältere). Wenn der Effekt in manchen Subgruppen anders (z. B. bei Älteren schwächer) ist, dann unterscheiden sich u. U. die Ergebnisse aus Beobachtungs- und klinischer Studie. Sensitivitätsanalysen können verwendet werden, um zu quantifizieren, wie stark sich Unterschiede im Studienprotokoll zwischen „target trial" und Emulierung auf die Ergebnisse auswirken [25]. Allerdings ist der Vorteil von Sekundärdaten und anderen RWD, dass sich spezielle Subgruppen untersuchen lassen, die in RCT unterrepräsentiert sind. In einer Studie, in der 10 RCTs mit Hilfe von Beobachtungsdaten repliziert wurden [10], waren die regulatorischen Entscheidungen in 6 von 10 Fällen identisch, und Beobachtungsstudien mit einem aktiven Vergleichsarm und einer vergleichbaren Indikation lieferten eher verlässliche Ergebnisse als placebokontrollierte Studien.

## Ableitbare Aussagen aus TTE

Zu guter Letzt möchten wir festhalten, dass TTE nicht als alternatives Design zu anderen Studiendesigns für Beobachtungsstudien, sondern als ergänzendes und übergreifendes Prinzip, in das bestimmte Designs und Methoden integriert werden können, verstanden werden sollte. TTE kann etwa sowohl mit prospektiven als auch eingebettete Fall-Kontroll-Designs, falls bestimmte Voraussetzungen erfüllt sind, kombiniert werden [6] und ist daher vielseitig und für verschiedenste Fragestellungen einsetzbar. Bestimmte Fragen lassen sich allerdings mit RWD generell nicht oder nur schwer beantworten. So kann beispielsweise der Vergleich eines Medikamentes mit einem Placebo nur in einem verblindeten RCT untersucht werden, nicht aber basierend auf RWD in einer Studie ohne Randomisierung oder Verblindung. Der Vorteil von TTE gegenüber klassischen Studiendesigns ohne explizite TTE liegt in der expliziten Herangehensweise, die den zentralen Herausforderungen bei der Analyse von RWD systematisch begegnet.

## Fazit für die Praxis

- **Sekundärdaten und andere RWD („real world data") sind eine wichtige Informationsquelle. Die Herausforderungen – im Vergleich zu RCTs („randomized controlled trials") – bei deren Analyse und Interpretation liegen zwar auch in der fehlenden Randomisierung, aber daneben gibt es eine Reihe von vermeidbaren Fehlerquellen, deren Relevanz bisher oft unterschätzt wurde.**
- **Die Target-trial-Emulierung (TTE) ist ein allgemeines Prinzip: Zunächst wird das Studienprotokoll für eine ideale randomisierte Studie aufgesetzt (der „target trial"). Diese ideale Studie wird dann basierend auf RWD emuliert.**
- **Das TTE-Prinzip hilft dabei, eine explizite und präzise Fragestellung zu formulieren, wodurch sich Ergebnisse klar kommunizieren lassen. TTE ist zudem mit verschiedenen Methoden der kausalen Inferenz kombinierbar.**
- **Die TTE beugt vielen vermeidbaren Fehlern bei der Analyse von Beobachtungsdaten systematisch vor. Unvermeidbare potenzielle Verzerrungsquellen werden transparent gemacht und sollten einer Sensitivitätsanalyse unterzogen werden.**

## Korrespondenzadresse

**Prof. Dr. Vanessa Didelez**
Abteilung für Biometrie und EDV, Leibniz Institut für Präventionsforschung und Epidemiologie – BIPS
Achterstr. 30, 28359 Bremen, Deutschland
didelez@leibniz-bips.de

**Funding.** Open Access funding enabled and organized by Projekt DEAL.

## Einhaltung ethischer Richtlinien

**Interessenkonflikt.** M. Braitmaier und V. Didelez geben an, dass kein Interessenkonflikt besteht.

Für diesen Beitrag wurden von den Autor/-innen keine Studien an Menschen oder Tieren durchgeführt. Für die aufgeführten Studien gelten die jeweils dort angegebenen ethischen Richtlinien.

## Literatur

1. Admon AJ, Donnelly JP, Casey JD et al (2019) Emulating a novel clinical trial using existing observational data. Predicting results of the PreVent study. Ann Am Thorac Soc 16:998–1007
2. Börnhorst C, Reinders T, Rathmann W et al (2021) Avoiding time-related biases: a feasibility study on antidiabetic drugs and pancreatic cancer applying the parametric g-formula to a large German healthcare database. Clin Epidemiol 13:1027–1038
3. Braitmaier M, Kollhorst B, Heinig M et al (2022) Effectiveness of mammography screening on breast cancer mortality—a study protocol for emulation of target trials using German health claims data manuscript submitted for publication
4. Braitmaier M, Schwarz S, Kollhorst B et al (2022) Screening colonoscopy similarly prevented distal and proximal colorectal cancer; A prospective study among 55–69-year-olds. J Clin Epidemiol

149:118–126. https://doi.org/10.1016/j.jclinepi.2022.05.024

5. Dickerman BA, Garcia-Albeniz X, Logan RW et al (2019) Avoidable flaws in observational analyses: an application to statins and cancer. Nat Med 25:1601–1606
6. Dickerman BA, Garcia-Albeniz X, Logan RW et al (2020) Emulating a target trial in case-control designs: an application to statins and colorectal cancer. Int J Epidemiol 49:1637–1646
7. Didelez V (2016) Commentary: should the analysis of observational data always be preceded by specifying a target experimental trial? Int J Epidemiol 45:2049–2051
8. Dommershuijsen LJ, Boon AJW, Ikram MK (2021) Probing the pre-diagnostic phase of Parkinson's disease in population-based studies. Front Neurol 12:702502
9. Fanaroff AC, Califf RM, Harrington RA et al (2020) Randomized trials versus common sense and clinical observation: JACC review topic of the week. J Am Coll Cardiol 76:580–589
10. Franklin JM, Patorno E, Desai RJ et al (2021) Emulating randomized clinical trials with nonrandomized real-world evidence studies: first results from the RCT DUPLICATE initiative. Circulation 143:1002–1013
11. Garcia-Albeniz X, Hsu J, Bretthauer M et al (2017) Screening colonoscopy to prevent colorectal cancer among medicare beneficiaries aged 70 to 79 years. Ann Intern Med 166:758–759
12. Garcia-Albeniz X, Hsu J, Hernan MA (2017) The value of explicitly emulating a target trial when using real world evidence: an application to colorectal cancer screening. Eur J Epidemiol 32:495–500
13. Harron K, Gamble C, Gilbert R (2015) E-health data to support and enhance randomised controlled trials in the United Kingdom. Clin Trials 12:180–182
14. Haug U, Schink T (2021) German pharmacoepidemiological research database (GepaRD). In: Sturkenboom M, Schink T (Hrsg) Databases for pharmacoepidemiological research. Springer, Cham, S 119–124
15. Hernan MA (2012) Beyond exchangeability: the other conditions for causal inference in medical research. Stat Methods Med Res 21:3–5
16. Hernan MA (2016) Does water kill? A call for less casual causal inferences. Ann Epidemiol 26:674–680
17. Hernan MA, Alonso A, Logan R et al (2008) Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease. Epidemiology 19:766–779
18. Hernan MA, Hernandez-Diaz S (2012) Beyond the intention-to-treat in comparative effectiveness research. Clin Trials 9:48–55
19. Hernan MA, Robins JM (2020) Causal inference: what if. Chapman & Hall/CRC, Boca Raton
20. Hernan MA, Robins JM (2016) Using big data to emulate a target trial when a randomized trial is not available. Am J Epidemiol 183:758–764
21. Howe CJ, Cole SR, Lau B et al (2016) Selection bias due to loss to follow up in cohort studies. Epidemiology 27:91–97
22. Institut Für Epidemiologie Und Sozialmedizin Der Universität Münster (2021) ZEBra-MSP Evaluation der Brustkrebsmortalität im deutschen Mammographie-Screening-Programm. https://www.medizin.uni-muenster.de/epi/forschung/projekte/zebra-msp.html. Zugegriffen: 25.5.2022
23. Joffe MM (2001) Administrative and artificial censoring in censored regression models. Stat Med 20:2287–2304
24. Lash TL, Fox MP, Fink AK (2009) Applying quantitative bias analysis to epidemiologic data. Springer, New York
25. Lodi S, Phillips A, Lundgren J et al (2019) Effect estimates in randomized trials and observational studies: comparing apples with apples. Am J Epidemiol 188:1569–1577
26. Pigeot I, Kollhorst B, Didelez V (2021) Secondary data for pharmacoepidemiological research—making the best of it! Gesundheitswesen 83:S69–S76
27. Ray WA (2003) Evaluating medication effects outside of clinical trials: new-user designs. Am J Epidemiol 158:915–920
28. Robins JM (1986) A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. Math Model 7:1393–1512
29. Stuart EA, Lee BK, Leacy FP (2013) Prognostic score-based balance measures can be a useful diagnostic for propensity score methods in comparative effectiveness research. J Clin Epidemiol 66:S84–S90.e1
30. Suissa S (2008) Immortal time bias in pharmacoepidemiology. Am J Epidemiol 167:492–499
31. Suissa S, Azoulay L (2012) Metformin and the risk of cancer: time-related biases in observational studies. Diabetes Care 35:2665–2673
32. Suissa S, Moodie EE, Dell'aniello S (2017) Prevalent new-user cohort designs for comparative drug effect studies by time-conditional propensity scores. Pharmacoepidemiol Drug Saf 26:459–468
33. Thomas LE, Yang S, Wojdyla D et al (2020) Matching with time-dependent treatments: a review and look forward. Stat Med 39:2350–2370
34. Zhou Y, Matsouaka RA, Thomas L (2020) Propensity score weighting under limited overlap and model misspecification. Stat Methods Med Res 29:3721–3756

## 7.4 Paper 3: Effectiveness of mammography screening on breast cancer mortality - a study protocol for emulation of target trials using German health claims data

This paper was published under a CC-BY NC 3.0 open access license in the Journal ”Clinical Epidemiology”. For details on how to cite the paper, refer to
https://doi.org/10.2147/CLEP.S376107

Clinical Epidemiology

Dovepress
open access to scientific and medical research

Open Access Full Text Article

STUDY PROTOCOL

# Effectiveness of Mammography Screening on Breast Cancer Mortality – A Study Protocol for Emulation of Target Trials Using German Health Claims Data

Malte Braitmaier[1], Bianca Kollhorst[1], Miriam Heinig[2], Ingo Langner[2], Jonas Czwikla[3], Franziska Heinze[3], Laura Buschmann[4], Heike Minnerup[4], Xabiér García-Albéniz[5,6], Hans-Werner Hense[4], André Karch[4], Hajo Zeeb[7,8], Ulrike Haug[2,8], Vanessa Didelez[1,9]

[1]Department of Biometry and Data Management, Leibniz Institute for Prevention Research and Epidemiology – BIPS, Bremen, Germany; [2]Department of Clinical Epidemiology, Leibniz Institute for Prevention Research and Epidemiology – BIPS, Bremen, Germany; [3]SOCIUM Research Center on Inequality and Social Policy, University of Bremen, Bremen, Germany; [4]Institute for Epidemiology and Social Medicine, Faculty of Medicine, Westfälische Wilhelms University of Münster, Münster, Germany; [5]Department of Epidemiology, Harvard T. H. Chan School of Public Health, Boston, MA, USA; [6]RTI Health Solutions, Barcelona, Spain; [7]Department of Prevention and Evaluation, Leibniz Institute for Prevention Research and Epidemiology – BIPS, Bremen, Germany; [8]Faculty of Human and Health Sciences, University of Bremen, Bremen, Germany; [9]Faculty of Mathematics and Computer Science, University of Bremen, Bremen, Germany

Correspondence: Vanessa Didelez, Leibniz Institute for Prevention Research and Epidemiology – BIPS, Department of Biometry and Data Management, Achterstraße 30, Bremen, 28359, Germany, Tel +49-421-56939, Fax +49-421-56941, Email didelez@leibniz-bips.de

**Background:** The efficacy of mammography screening in reducing breast cancer mortality has been demonstrated in randomized trials. However, treatment options - and hence prognosis – for advanced tumor stages as well as mammography techniques have considerably improved since completion of these trials. Consequently, the effectiveness of mammography screening under current conditions is unclear and controversial. The German mammography screening program (MSP), an organized population-based screening program, was gradually introduced between 2005 and 2008 and achieved nation-wide coverage in 2009.

**Objective:** We describe in detail a study protocol for investigating the effectiveness of the German MSP in reducing breast cancer mortality in women aged 50 to 69 years based on health claims data. Specifically, the proposed study aims at estimating per-protocol effects of several screening strategies on cumulative breast cancer mortality. The first analysis will be conducted once 10-year follow-up data are available.

**Methods and Analysis:** We will use claims data from five statutory health insurance providers in Germany, covering approximately 37.6 million individuals. To estimate the effectiveness of the MSP, hypothetical target trials will be emulated across time, an approach that has been demonstrated to minimize design-related biases. Specifically, the primary contrast will be in terms of the cumulative breast cancer mortality comparing the screening strategies of "never screen" versus "regular screening as intended by the MSP".

**Ethics and Dissemination:** In Germany, the utilization of data from health insurances for scientific research is regulated by the Code of Social Law. All involved health insurance providers as well as the responsible authorities approved the use of the health claims data for this study. The Ethics Committee of the University of Bremen determined that studies based on claims data are exempt from institutional review. The findings of the proposed study will be published in peer-reviewed journals.

**Keywords:** emulated target trial, cancer screening, effectiveness, claims data, mammography

## Introduction

### Background

Mammography screening aims at reducing breast cancer mortality through early diagnosis of asymptomatic, early-stage cancers.[1] The prognosis of breast cancer is considerably better when diagnosed at an early stage.[2–4] Several randomized

Received: 28 June 2022
Accepted: 12 October 2022
Published: 2 November 2022

clinical trials were conducted in the second half of the last century demonstrating a reduction in breast cancer mortality due to screening with mammography.[5,6]

In parallel to screening efforts increasing world-wide, novel treatment options for women with advanced breast cancer stages have also been introduced over the last two decades leading to improved survival rates, particularly for advanced stages without distant metastases.[7,8] Consequently, the reduction in breast cancer mortality due to screening might be lower if trials were conducted nowadays. However, mammography techniques have also improved such that the present sensitivity of imaging techniques might have resulted in greater mortality reductions as compared to the earlier trials.[9] Given that RCTs comparing mammography screening against no screening nowadays are no longer ethical, the analysis of observational data is the only option to obtain insights into the effectiveness of mammography screening under current conditions. A few large observational studies have been conducted on the effectiveness of mammography screening and indicated a reduction in breast cancer mortality in screened women.[10–12] To date, however, there has been no observational study on this research question using the novel principle of target trial emulation, which specifically aims to minimize common time-related and other biases.

In Germany, an organized mammography screening program (MSP) was introduced from 2005 to 2008, achieving nation-wide coverage in 2009. All women aged 50 to 69 years, with German residency, are centrally invited biennially by mail to attend screening at one of the 94–95 certified mammography screening units.[1] Participation rates in the German MSP are around 50% per screening round and 83% of women in a survey said they had participated at least once over a 10-year time frame.[13,14] Information on whether and when invitations were issued is not available due to data protection reasons; screening attendance, however, can be identified using specific health insurance claims codes.

## Objectives

The proposed observational study is part of a larger research effort commissioned by the German Federal Office for Radiation protection to evaluate whether mammography screening is beneficial in Germany. Within this research effort, our proposed study will estimate the effects of different screening strategies in the German mammography screening program on breast cancer mortality over a 10-year follow-up in women aged 50 to 69 at baseline. Specifically, three screening strategies will be compared: 1) Never screening 2) Screening at least at baseline, with free choice whether to undergo screening afterwards 3) Screening at baseline and then regularly every two years.

Two primary research questions will be addressed:

Research question 1: Does participation in the German MSP reduce breast cancer mortality in the population of all eligible women?

Research question 2: Does participation in the German MSP reduce breast cancer mortality in the subgroup of screening-affine women?

While question 1 addresses the ideal situation that all those eligible participate, question 2 is relevant as it concerns those women who are most likely to participate in the MSP. Both are regarded as primary research questions. The follow-up time of 10 years refers to the time point when the first analysis will be conducted. Re-analysis based on extended follow-up is planned.

For each research question we will consider the following two contrasts:

- Primary contrast: Strategy 1 (never screened) versus strategy 3 (regular screening).
- Secondary contrast: Strategy 1 (never screened) versus strategy 2 (screening at baseline).

Specifically, in view of competing events, we will assess the total effect[15] of the strategies in the primary analysis. The primary contrast reflects the original intention of the MSP and is relevant to the individual women who can decide whether to participate and adhere to the program or not. The secondary contrast addresses the effect of offering the MSP under the reality of imperfect adherence; it is therefore also relevant to public health decision makers.

# Materials and Methods

## Description of Data Sources

The German Pharmacoepidemiological Research Database (GePaRD) and the BARMER data warehouse (DWH) will be the main data sources for this study. GePaRD comprises health claims data from four German statutory health insurance (SHI) providers, with data on approximately 25 million individuals who have been insured with one of the participating SHI providers since 2004 or later.[16] The BARMER DWH covers approximately 12.6 million individuals who were insured with BARMER between 2006 and 2017.[17] In Germany, health insurance is mandatory, with 87% of the population being insured with an SHI provider (11% of the population are insured with a private insurance provider and further government schemes exist, eg for soldiers or refugees).[18] The health claims data contain basic demographic information, codes for outpatient drug prescriptions, outpatient physician contacts, in- and outpatient operations, procedures, and diagnoses. Outpatient procedures and diagnoses are coded on a quarterly basis, while exact dates are available for inpatient codes and outpatient services. Reimbursed drugs are identified based on Anatomical Therapeutic Chemical (ATC) codes, diagnoses are identified based on International Classification of Diseases, tenth revision, German modification (ICD-10-GM) codes and procedures and services based on Operation and Procedure classification (OPS) codes and Uniform Assessment Standard (EBM) codes.

All regions of Germany are represented in the data from the involved SHIs. GePaRD and BARMER data will be analyzed separately for reasons of data protection. Similar data from further health insurance providers might be added to increase sample size if and when their use for this project will be approved.

Data starting in 2004 for GePaRD and 2006 for BARMER will be used for this study. For the first analysis, data up to and including 2018 will be used. The follow-up will be extended as soon as further data years are available.

## Study Design

To address the research questions, we use a target trial emulation approach.[19,20] While any observational study might suffer from bias due to uncontrolled confounding, awareness has recently increased for biases (often time-related) due to deviation from basic principles of study design. These latter, "self-inflicted" biases, can be avoided or reduced by emulating, as best as possible, the design of a hypothetical randomized trial that would ideally answer the research question (the target trial).[21] For our proposed mammography study, the protocol of the hypothetical target trial and its emulation with health claims data are described in Table 1. Multiple consecutive trials will be emulated, with one trial starting on the first day of each calendar quarter, to make full use of the longitudinal database. At the core of target trial emulation is the alignment of eligibility checks, assignment to treatment strategies, and start of follow-up at time-zero, ie baseline of each trial. A lack of such alignment is likely to entail erroneous conclusions.[21,22] Therefore, eligibility criteria will be assessed at the baseline of each emulated trial. As a woman may qualify for multiple trials (starting in different quarters), her individual data will be copied (or "cloned") and included in every trial for which she is eligible[23] (see Figure S1). Furthermore, at each baseline, a woman's data might fit with more than one screening strategy. Again, information from this woman will be copied and one clone will be assigned to each screening strategy she fits. Hence one person can contribute to several trials and within one trial to several screening strategies. This cloning approach reduces time-related biases,[23,24] while maximizing statistical efficiency.[22] Data from all emulated trials and all clones within each trial will be pooled and analyzed jointly. The respective analysis dataset will then contain information on m clones across all trials originating from n women (ie $m \geq n$). Randomization is emulated by adjustment for confounding (more details are given below and in Supplement 1). Each of the emulated trials has its own baseline, defined as the first day of the calendar quarter of trial start. Pre-baseline covariates are based on information before this day, while follow-up and outcome variables are based on information starting with this day, again ensuring alignment at time zero.

## Eligibility Criteria

Individuals must satisfy the eligibility criteria listed in the emulated trial column in Table 1. Eligibility will be assessed at the baseline of each emulated trial.

**Table 1** Tabular Study Protocol for the Ideal Target Trial and the Approximation by Our Emulated Trial

| Component | Target Trial | Emulated Trial |
|---|---|---|
| Aim | To estimate the effect, if any, of different mammography-based screening strategies on breast cancer mortality in the German population aged 50–69. | Same |
| Eligibility | To be eligible, women must:<br>● Be 50 to 69 years old.<br>● Have no history of breast cancer, carcinoma in situ of the breast or unspecified lumps in the breast.<br>● Be naïve to screening or diagnostic mammography and other imaging of the breast (in order to avoid selection of the study population according to prior screening history).<br>● Be permanently living in Germany. | To be eligible, women must:<br>● Not have missing information on sex, age, and region of residency.<br>● Be 50 to 69 years old.<br>● Be continuously insured for the 3 years before trial start.<br>● Have no coded diagnosis of breast cancer, carcinoma in situ of the breast or unspecified lumps in the breast) ever before baseline.<br>● Have no coded screening or diagnostic mammography or other imaging of the breast within 3 years before baseline.<br>● Be permanently living in Germany.<br>● For research question 2: have had at least one of the following preventive services coded during 3 years before trial start: screening colonoscopy, pap test or breast examination, health check-up 35, fecal occult blood test, influenza vaccine, skin cancer screening (in order to identify screening-affine women). |
| Screening strategies | 1. Never undergo screening.<br>2. Screening at least at baseline.<br>3. Regular screening (two-year intervals).<br>Women are retained under their strategy if they receive a breast cancer diagnosis or if they stop regular screening (strategy 3) at age 70 or older. Receiving a screening mammogram will be considered non-adherence for strategy 1. Under all strategies, diagnostic mammograms are allowed when clinically indicated. | Same |
| Assignment to study arms | Randomly to one study arm.<br>Randomization is unblinded. | Women are assigned to screening strategies based on observed screening behavior in baseline quarter.<br>We assume random assignment within the levels of the baseline covariates described in the Supplement. |
| Follow-up | Start: Treatment assignment.<br>End: Death, loss to follow-up or end of study period at 2018, whichever occurs first. | Same, except start is the first day of the quarter of trial start. Length of follow-up is 10 years. |
| Outcome | Death from breast cancer. | Same (as determined either by the cause of death algorithm or by record linkage). |
| Causal contrast | Per protocol (PP) effect. | Observational analogue of PP effect. Adjustment for baseline and time-varying post-baseline confounding is necessary. |
| Statistical analysis | Women are artificially censored when they deviate from their assigned screening strategy as follows:<br>● No screening: Censored when a screening mammography occurs.<br>● Screening at baseline: No censoring based on screening participation after baseline quarter.<br>● Regular screening: Censored when no subsequent screening mammography was coded, unless the woman turned 70 or received a breast cancer diagnosis by the tenth quarter after the last screening mammography.<br>The analysis is adjusted for non-adherence using baseline and post-baseline variables (eg via inverse probability weighting). | Same, except that data from each eligible woman receiving screening in the baseline quarter is cloned and assigned to screening strategies 2–3. Randomization will be emulated via adjustment for baseline confounders. Bias due to artificial censoring will be adjusted for using post-baseline confounders. |

**Note:** The emulated trial is purely observational and exposure to the screening strategies is therefore based on the observed (appropriately censored) participation profiles. Sequential trials are emulated in each calendar quarter from 2009 to 2016, with each trial applying the eligibility criteria at its respective baseline.

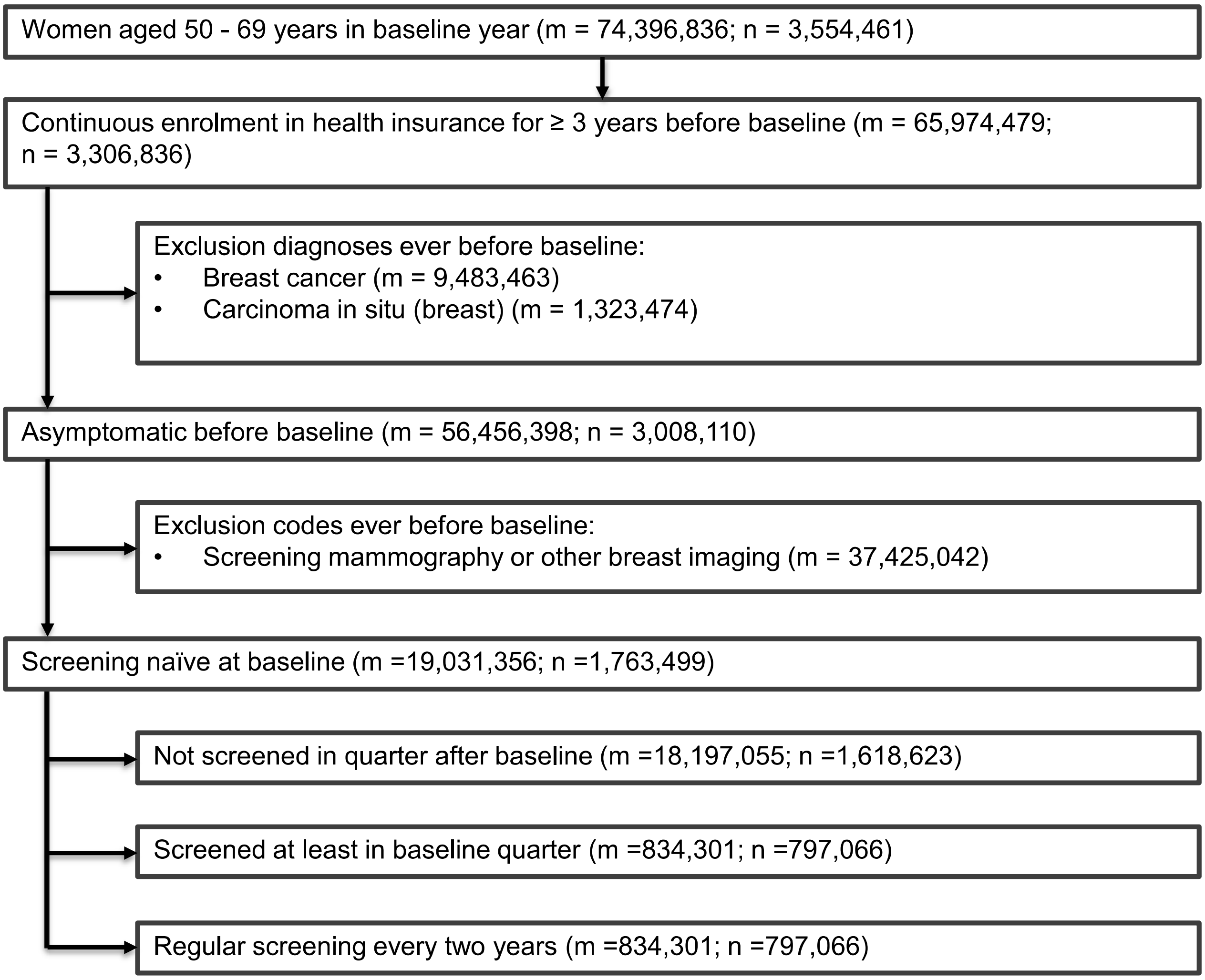

Women aged 50 - 69 years in baseline year (m = 74,396,836; n = 3,554,461)
Continuous enrolment in health insurance for ≥ 3 years before baseline (m = 65,974,479; n = 3,306,836)
Exclusion diagnoses ever before baseline:
• Breast cancer (m = 9,483,463)
• Carcinoma in situ (breast) (m = 1,323,474)
Asymptomatic before baseline (m = 56,456,398; n = 3,008,110)
Exclusion codes ever before baseline:
• Screening mammography or other breast imaging (m = 37,425,042)
Screening naïve at baseline (m =19,031,356; n =1,763,499)
Not screened in quarter after baseline (m =18,197,055; n =1,618,623)
Screened at least in baseline quarter (m =834,301; n =797,066)
Regular screening every two years (m =834,301; n =797,066)

**Figure 1** Flow chart of subject disposition. m refers to all clones across all emulated trials while n refers to women. GePaRD data from 2004–2016 was used; no information on the study outcome was available at the time of analysis. Clones were assigned to screening strategies as illustrated in the Supplement. Active screening strategies have identical sample sizes, since they only differ in the screening sustained over time and, therefore, in the censoring process. For the estimation of the subgroup effect in women with at least one other preventive service during three years before baseline, a further decrease in sample size of 32.6% was observed (screening naïve at cohort entry - m: 12,819,058, n: 1,503,094).

## Sample Size

An overview of sample sizes (using GePaRD data from 2004 to 2016) is given in the flow chart in Figure 1.

## Screening Strategies, Cloning and Artificial Censoring

The mammography screening strategies to be compared will be:

No screening: Under this strategy a woman never undergoes screening (control strategy).

Screening at baseline: Under this strategy, a woman undergoes screening in the baseline quarter and may or may not attend further screenings afterwards.

Regular screening: Under this strategy, a woman undergoes screening in the baseline quarter and in regular two-year intervals thereafter as long as she is in the age range of screening.

Assignment of a woman's data to a strategy has to be done carefully using only baseline information, ie without "looking into the future". Thus, as explained above, her data will be cloned and one clone will be assigned to each strategy with which the woman's behavior is consistent in the baseline quarter, resulting in multiple clones (see Supplement Figure S1 for illustration, and reference[20] for a methodological introduction). Note that women who died in the baseline quarter without starting screening are consistent with and therefore assigned to all strategies. This avoids accumulation of early breast cancer deaths in the no screening strategy and, thereby, avoids bias. The same applies to women who received a breast cancer diagnosis in the baseline quarter without starting screening. Women who undergo screening within the baseline quarter are only cloned into the active screening strategies, since their observed screening behavior at baseline is not consistent with the control strategy. Furthermore, they are cloned into both active screening strategies, which avoids immortal time bias.[21]

Clones will be artificially censored at the start of the first calendar quarter during which their observed screening behavior deviates from the assigned treatment strategy (Supplement Figures S2–S4). Artificial censoring describes the analyst's decision to ignore any future data for this subject, just as if it were missing.[25] For example, a woman's data will be censored in the "never screened" strategy at the time when she receives a screening, and a woman's data will be censored from the "regular screening" strategy at the time when she misses a regular screening. Regular screenings are defined as screenings taking place between the fifth and tenth quarters (ie 12th to 30th month) after the quarter of the previous screening.

Artificial censoring can introduce selection bias if (time-varying) factors influence both deviation from the assigned strategy and the outcome. For instance, a woman may start taking hormone replacement therapy (HRT) and, due to the increased breast cancer risk associated with this medication, also be advised to start regular mammography screening. At that point, she deviates from the "no screening" strategy and would be artificially censored at the time of her first screening mammogram in that strategy. Thus, artificial censoring will be more likely for women using HRT than for women not using HRT, so that censoring might induce selection bias. However, this bias is avoided by weighting with the inverse probability of censoring taking HRT (and other relevant time-varying factors) into account;[26] note that this is equivalent to adjusting for time-varying confounding. Women will not be artificially censored under any strategy after receiving a breast cancer diagnosis or after turning 70. No artificial censoring occurs for the "screening at baseline" strategy as any behavior, regular, irregular or lack of further screening after baseline, is compatible with this strategy. Note that within all strategies, diagnostic mammography may take place at any time, as required or indicated, and does not lead to artificial censoring as we aim to assess the added benefit of the MSP.

## Exposure

While invitation to screening is not captured in the data, utilization of screening mammography can be identified via a unique EBM code and, thus, is distinguishable from utilization of diagnostic mammography. Women who have not attended screening yet or never attend screening will be included in the control strategy.

## Outcome

Information on cause of death is not recorded in health claims data. For the majority of the study population, breast cancer deaths will therefore be identified via an algorithm that uses available information in claims data in the year of death. The algorithm has been developed in a sample for which both claims data and the official cause of death were directly linked. The initial version of the algorithm, described by Langner et al,[27] showed a sensitivity of 91.3% and a specificity of 97.4%, and is currently being further optimized, eg by also considering information on cancer treatment. For study participants living in the federal states of North Rhine-Westphalia, Bavaria, and Lower Saxony official cause of death records will be directly available by linkage to the cancer registry of the respective federal state.

## Covariates

Covariates for confounder adjustment were selected following subject matter knowledge and considerations about the causal relationships between covariates, exposure and outcome. An illustration of relevant causal patterns is given in Figure S5 in the Supplement. Details on how confounding as a potential source of bias is considered and how relevant covariates are captured in the data are provided in Supplements 4 and 5, and a preliminary list of covariates is given in Table S1 in Supplement 6. The list of covariates used in the final analysis will be finalized before data on the study outcome becomes available. All baseline covariates that can vary over time will be re-assessed at each time point, ie on a quarterly basis. All of these updated covariate values will be used to estimate inverse probability weights for artificial censoring and competing events censoring (the latter only applies to sensitivity analyses). We will apply the usual model diagnostics and carry out balance checks. In addition to these covariates, further variables will be assessed to describe the study cohort.

## Missing Data

Individuals with missing core demographic information (ie age, sex, and region of residency) will be excluded from the study. We expect this to be a negligible proportion of women.

We assume that prescriptions, diagnoses, and procedures not coded in our data did not take place. Since no information other than codes from the databases is available, this assumption cannot be verified. Over-the-counter prescriptions and medical services that are not reimbursed by health insurance providers are not coded in our database.

## Loss to Follow-Up

Loss to follow-up may occur due to interruption of continuous enrolment, or end of insurance coverage. Interruptions in insurance coverage are very rare in Germany, particularly in the age group relevant for this study.[28] We therefore assume that loss to follow-up is neither related to screening participation nor to the risk of breast cancer. Women are censored at loss to follow-up.

## Addressing Potential Sources of Bias

As explained under “Study design”, the target trial emulation principle, combined with cloning and artificial censoring, ensures the alignment at time zero and thus mitigates many typical design-related biases in observational studies. Under “Covariates” we further address how information in the claims data can be used to adjust for confounding. A systematic overview and further explanation is provided in Supplement 4, addressing the topics “Confounding”, “Healthy screenee bias”, “Competing events”, “Time-related biases”, “Misclassification” and “Identifying assumptions”.

## Primary Analysis

The primary analysis will consist of an estimation of the per-protocol effect of the screening strategies on breast cancer mortality, both in the overall population (research question 1) and in a subgroup of screening affine women (research question 2). One trial will be emulated for each calendar quarter from 2009 to 2016 with one woman possibly contributing to multiple screening strategies per trial. This means that the baseline of the first trial is January 1st, 2009 and follow-up extends until end of data availability. The baseline of the second trial is April 1st, 2009 and follow-up extends until end of data availability. Thus, one trial is emulated per quarter, until the last emulated trial starts on October 1st, 2016. Data from all these emulated trials will be pooled and analyzed jointly. Clones will be artificially censored as described above and reweighted with suitable inverse probability weights. In our main analysis, we estimate the total effect on breast cancer mortality, ie the effect when death from other causes is not eliminated.[15] Adjusted cumulative incidence functions (CIF) will be estimated using a pooled logistic regression (for details on the statistical methods used, see Supplement 1).

The comparison of the effect of screening strategies will be done in terms of differences in CIF, ie the effect will be observed at each point of follow-up. For the comparison of strategy 1 (never screened) with strategy 3 (regular screening), the above standardization to the empirical distribution of baseline confounders will use the confounder distribution of the entire study population, ie we estimate the average treatment effect (ATE). For the comparison of strategy 1 with strategy 2 (screening at least at baseline), on the other hand, the confounder distribution among treated women will be used, ie we will estimate the average treatment effect on the treated (ATT). The ATT will be more informative to answer the health policy question of whether offering the screening given imperfect adherence (ie contrast between strategies 1 and 2) and given the confounder distribution in women who decide to undergo screening is effective in lowering breast cancer mortality. Confidence intervals will be based on a person-level bootstrap to account for cloning. For a more detailed description of the statistical methods used here, see references.[15,19,29] The analyses may use a random sample of controls only (ie from the never screened strategy), if computationally prohibitive otherwise.

Furthermore, alternative adjustment methods may be used to adjust for baseline confounding instead of the above-described standardization if the bootstrap sampling becomes computationally prohibitive.

## Sub-Group and Secondary Analyses

While primary research question 1 refers to the entire study population, primary research question 2 will assess the effect of screening in the sub-group of screening affine women. These are defined as having attended at least one of the following preventive services during the three years preceding baseline: pap test or breast examination (identified via a single claim code), health check-up after age 35, skin cancer screening, screening colonoscopy, fecal occult blood test, influenza vaccination. By choosing a more restricted and homogeneous study population for primary research question 2, we aim at minimizing residual confounding while being aware that the effect within this special group may be different than in the larger population.[30]

As a secondary analysis, stratification by calendar year at baseline will be carried out in order to account for the implementation phase of the MSP. We will group all clones from emulated trials with baseline before or in 2011 in one stratum and all others in another stratum. The choice of the cut-off year 2011 is based on baseline characteristics in preliminary analyses (data not shown) and results in one stratum with highly variable age structure (implementation phase until 2011) and one stratum with more homogeneous age structure.

Furthermore, a restricted analysis without women who have a coded family history of breast cancer will be conducted, in order to obtain a subpopulation excluding high-risk individuals.

## Sensitivity Analyses

The main analysis assesses the total effect of screening on breast cancer mortality, which encompasses the effect of screening on breast cancer mortality mediated by death due to other causes. This amounts to estimating the event-specific CIF for breast cancer death as event of interest.[15] An estimation of the direct effect, ie under a hypothetical intervention which eliminates all competing events (ie death from other causes), will be conducted in a sensitivity analysis.[15] A comparison with the main analysis will help assess the impact of competing events on any conclusions. Furthermore, the models from the primary analysis will be re-fitted, but with all-cause mortality as outcome variable.

Further sensitivity analyses, such as quantitative bias analysis regarding family history of breast cancer (see Supplement 4), will be added to the analysis. Results of all secondary and sensitivity analyses will be interpreted in an exploratory way.

# Discussion

We propose a design for an observational study in Germany that aims to investigate whether mammography screening reduces breast cancer mortality. The first data analysis will be conducted once 10-year follow-up data are available; extension of follow-up is planned. To the best of our knowledge, there is currently no other study using the principle of target trial emulation to address this research question. García-Albéniz et al used target trial emulation to investigate the continuation of screening mammography after age 70.[31] There have been other large observational studies investigating screening mammography and breast cancer mortality that were conceptually different from our proposed study. Furthermore, these studies did not have individual-level confounder information, nor an unscreened control group and they did not employ a per-protocol design.[10,11] The key contribution of our proposed study will be an up-to-date assessment of mammography screening effectiveness in a real-world German population, complementing evidence from earlier randomized trials in other countries. For many reasons, we do not expect exact agreement of our results with those from previous RCTs (see eg Groenwold for a discussion[32]), but instead aim at complementing past studies with the best possible evidence currently available on whether screening mammography affects breast cancer mortality in the German population. The chosen screening strategies will inform individual women's choices as well as policy makers. The proposed study design carefully accounts for potential sources of bias and ambiguity. In particular, we also conduct analyses restricted to screening affine women, which is expected to minimize healthy-screenee bias and thus leads to a high internal validity. Moreover, the large size of our database constitutes a clear strength of our proposed study.

While our study focuses on the effectiveness of mammography screening on breast cancer mortality, we are fully aware that mammography screening also has harmful effects. Overdiagnoses are considered to be one of the major harms of mammography screening including subsequent treatment of overdiagnosed cases.[33,34] Mammography screening programs have been implemented in many countries because it is assumed that the benefits of mammography on breast cancer mortality outweigh these harms, but there is an ongoing debate with some scientists questioning this.[33,35,36] Given that a very long follow-up of up to 30 years would be required to address overdiagnoses,[37] our study cannot contribute to the debate on overdiagnoses. By focusing on the question whether there is a benefit of mammography screening under current conditions in Germany, our study addresses one part of the evaluation required by law. According to German law, any screening method to detect non-communicable diseases entailing exposure to radiation must be assessed both regarding the ability to detect the disease at an early state and thereby improve prognosis and regarding the harm to benefit ratio (§84 of the German Radiation Protection Law, "Strahlenschutzgesetz/StrlSchG").

While the design of our study addresses several sources of bias, it is still limited by other issues of observational analyses. While we have carefully considered all plausible sources of confounding as detailed in Supplement 4, some unmeasured (baseline or time-dependent) confounding that cannot be mitigated with information based on claims codes cannot be ruled out. In particular, the role of family history of breast cancer, which is only partly observed, will therefore be assessed in quantitative bias analyses. Additionally, for a part of the study population the official cause of death is not available and will instead be identified based on an algorithm that has previously been validated through data linkage. The former version of this algorithm has already shown high sensitivity and specificity and is currently being further optimized.[27] Furthermore, imprecision in the date of some codes is present in the data, as outpatient diagnoses are coded on a quarterly basis. We mitigated this by processing all data on a quarterly basis. This, however, introduces the limitation of potential residual time-related biases in quarterly trial emulation. With these limitations in mind, we are confident that our study represents the best analysis currently possible on the effectiveness of the mammography screening program in Germany.

Given data availability, we expect to publish the results of this study by the end of 2024.

## Abbreviations

ATC, anatomical therapeutic chemical; ATE, average treatment effect; ATT, average treatment effect on the treated; CIF, cumulative incidence function; DWH, data warehouse; EBM, uniform assessment standard [German: Einheitlicher Bewertungsmaßstab]; GePaRD, German Pharmacoepidemiological Research Database; HRT, hormone replacement therapy; ICD-10-GM, international classification of diseases, tenth revision, German modification; MSP, mammography screening program; OPS, operation and procedure classification [German: Operationen- und Prozedurenschlüssel]; SHI, statutory health insurance.

## Ethics Approval and Informed Consent

Pseudonymized health care data may be used for scientific purposes without individual informed consent under German law if the data protection of individuals is not compromised or the public interest of the research project outweighs the data protection interest of individuals (see §75 SGB X). All SHI providers approved the use of their data for this study. Approval of the use of GePaRD data has also been given by the German Federal Office for Social Security and the Senator for Health, Women and Consumer Protection in Bremen. All published results will display aggregated data only, so that identification of individuals will be impossible.

## Data Sharing Statement

In Germany, use of personal data is protected by the Federal Data Protection Act and particularly the use of claims data for research is regulated by the Code of Social Law. Researchers must apply for a project-specific permit from the statutory health insurance providers which then need an approval from their governing authorities. The use of the data on which this publication is based was only allowed for BIPS employees within the framework of the specified project and limited to a pre-defined time span. Researchers who want to access the data on which this publication is based need to ask for new approval by the statutory health insurance providers DAK-Gesundheit (service@dak.de), die Techniker (service@tk.de), hkk Krankenkasse (info@hkk.de) and AOK Bremen/Bremerhaven (info@hb.aok.de) which upon granting approval will ask their respective authorities for approval. Please contact gepard@leibniz-bips.de for help with this process.

## Acknowledgments

We are grateful to Klaus Giersiepen for his valuable contributions to this study. Furthermore, we thank Anja Gabbert and Inga Schaffer for the programming of the GePaRD flow chart. We thank all participating health insurance providers for their cooperation. Specifically, we thank the AOK Bremen/Bremerhaven, DAK-Gesundheit, hkk Krankenkasse, and die Techniker (TK) for making available their data for the flow chart shown in this study protocol. A preliminary version of this study protocol was published in a project status report on the institutional website of the study sponsor: https://doris.bfs.de/jspui/bitstream/urn:nbn:de:0221-2022030131647/3/BfS_2022_3617S42402.pdf.

## Author Contributions

All authors made substantial contributions to conception and design, acquisition of data, or analysis and interpretation of data; took part in drafting the article or revising it critically for important intellectual content; agreed to submit to the current journal; gave final approval of the version to be published; and agree to be accountable for all aspects of the work.

## Funding

The project was/is funded by the Federal Office for Radiation Protection from funds of the Federal Ministry for the Environment, Nature Conservation and Nuclear Safety, the Federal Ministry of Health and the Kooperationsgemeinschaft Mammographie (supported by the National Association of Statutory Health Insurance Funds and the National Association of Statutory Health Insurance Physicians), grant number: 3617S42402 and 3621S42410.

## Disclosure

The authors declare no competing interests.

## References

1. Biesheuvel C, Weigel S, Heindel W. Mammography screening: evidence, history and current practice in Germany and other European countries. *Breast Care*. 2011;6(2):104–109. doi:10.1159/000327493
2. John S, Broggio J Cancer survival in England: national estimates for patients followed up to 2017. Newport: Office for National Statistics; 2019.
3. Sant M, Allemani C, Capocaccia R, et al. Stage at diagnosis is a key explanation of differences in breast cancer survival across Europe. *Int J Cancer*. 2003;106(3):416–422. doi:10.1002/ijc.11226
4. Erdmann F, Spix C, Katalinic A, et al. Krebs in Deutschland für 2017/2018; 2021. Available from: https://edocrkide/handle/176904/9042. Accessed August 24, 2022.
5. Nelson HD, Fu R, Cantor A, Pappas M, Daeges M, Humphrey L. Effectiveness of breast cancer screening: systematic review and meta-analysis to update the 2009 U.S. preventive services task force recommendation. *Ann Intern Med*. 2016;164(4):244–255. doi:10.7326/M15-0969
6. Marmot MG, Altman DG, Cameron DA, Dewar JA, Thompson SG, Wilcox M. The benefits and harms of breast cancer screening: an independent review. *Br J Cancer*. 2013;108(11):2205–2240. doi:10.1038/bjc.2013.177
7. Guarneri V, Conte PF. The curability of breast cancer and the treatment of advanced disease. *Eur J Nucl Med Mol Imaging*. 2004;31(Suppl 1): S149–S161. doi:10.1007/s00259-004-1538-5
8. Jansen L, Holleczek B, Kraywinkel K, et al. Divergent patterns and trends in breast cancer incidence, mortality and survival among older women in Germany and the United States. *Cancers*. 2020;12(9):2419. doi:10.3390/cancers12092419
9. Chiarelli AM, Edwards SA, Prummel MV, et al. Digital compared with screen-film mammography: performance measures in concurrent cohorts within an organized breast screening program. *Radiology*. 2013;268(3):684–693. doi:10.1148/radiol.13122567
10. Duffy SW, Tabar L, Yen AM, et al. Mammography screening reduces rates of advanced and fatal breast cancers: results in 549,091 women. *Cancer*. 2020;126(13):2971–2979. doi:10.1002/cncr.32859
11. Coldman A, Phillips N, Wilson C, et al. Pan-Canadian study of mammography screening and mortality from breast cancer. *J Natl Cancer Inst*. 2014;106(11). doi:10.1093/jnci/dju261
12. Lauby-Secretan B, Loomis D, Straif K. Breast-cancer screening - viewpoint of the IARC working group. *N Engl J Med*. 2015;373(15):1479. doi:10.1056/NEJMc1508733
13. Kääb-Sanyal V, Hand E. Jahresbericht evaluation 2019 deutsches mammographie-screening-programm. Kooperationsgemeinschaft Mammographie; 2021.
14. Schmuker C, Zok K. Informierte Teilnahme an Früherkennungsuntersuchungen: Ergebnisse einer Befragung unter GKV-Versicherten [Informed participation in screening examinations: Results of a survey among individuals enrolled in statutory health insurance]. *Versorgungsreport Früherkennung*. 2019:31–78. doi:10.32745/9783954664023-2

15. Young JG, Stensrud MJ, Tchetgen EJT, Hernan MA. A causal framework for classical statistical estimands in failure-time settings with competing events. *Stat Med*. 2020;39(8):1199–1236. doi:10.1002/sim.8471
16. Haug U, Schink T. German pharmacoepidemiological research database (GePaRD). In: Sturkenboom M, Schink T, editors. *Databases for Pharmacoepidemiological Research*. Springer; 2021:119–124.
17. Czwikla J, Urbschat I, Kieschke J, Schussler F, Langner I, Hoffmann F. Assessing and explaining geographic variations in mammography screening participation and breast cancer incidence. *Front Oncol*. 2019;9:909. doi:10.3389/fonc.2019.00909
18. Busse R, Blumel M, Knieps F, Barnighausen T. Statutory health insurance in Germany: a health system shaped by 135 years of solidarity, self-governance, and competition. *Lancet*. 2017;390(10097):882–897. doi:10.1016/S0140-6736(17)31280-1
19. Hernan MA, Robins JM. *Causal Inference - What if*. Chapman & Hall/CRC; 2020.
20. Hernan MA, Robins JM. Using big data to emulate a target trial when a randomized trial is not available. *Am J Epidemiol*. 2016;183(8):758–764. doi:10.1093/aje/kwv254
21. Hernan MA, Sauer BC, Hernandez-Diaz S, Platt R, Shrier I. Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses. *J Clin Epidemiol*. 2016;79:70–75. doi:10.1016/j.jclinepi.2016.04.014
22. Garcia-Albeniz X, Hsu J, Hernan MA. The value of explicitly emulating a target trial when using real world evidence: an application to colorectal cancer screening. *Eur J Epidemiol*. 2017;32(6):495–500. doi:10.1007/s10654-017-0287-2
23. Zhao SS, Lyu H, Yoshida K. Versatility of the clone-censor-weight approach: response to “trial emulation in the presence of immortal-time bias”. *Int J Epidemiol*. 2021;50(2):694–695. doi:10.1093/ije/dyaa223
24. Maringe C, Benitez Majano S, Exarchakou A, et al. Reflection on modern methods: trial emulation in the presence of immortal-time bias. Assessing the benefit of major surgery for elderly lung cancer patients using observational data. *Int J Epidemiol*. 2020;49(5):1719–1729. doi:10.1093/ije/dyaa057
25. Joffe MM. Administrative and artificial censoring in censored regression models. *Stat Med*. 2001;20(15):2287–2304. doi:10.1002/sim.850
26. Robins JM, Hernan MÁ, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology*. 2000;11(5):550–560. doi:10.1097/00001648-200009000-00011
27. Langner I, Ohlmeier C, Haug U, Hense HW, Czwikla J, Zeeb H. Implementation of an algorithm for the identification of breast cancer deaths in German health insurance claims data: a validation study based on a record linkage with administrative mortality data. *BMJ Open*. 2019;9(7): e026834. doi:10.1136/bmjopen-2018-026834
28. Pigeot I, Ahrens W. Establishment of a pharmacoepidemiological database in Germany: methodological potential, scientific value and practical limitations. *Pharmacoepidemiol Drug Saf*. 2008;17(3):215–223. doi:10.1002/pds.1545
29. Dickerman BA, Garcia-Albeniz X, Logan RW, Denaxas S, Hernan MA. Avoidable flaws in observational analyses: an application to statins and cancer. *Nat Med*. 2019;25(10):1601–1606. doi:10.1038/s41591-019-0597-x
30. Schneeweiss S, Patrick AR, Sturmer T, et al. Increasing levels of restriction in pharmacoepidemiologic database studies of elderly and comparison with randomized trial results. *Med Care*. 2007;45(10Supl 2):S131–S42. doi:10.1097/MLR.0b013e318070c08e
31. Garcia-Albeniz X, Hernan MA, Hsu J. Continuation of annual screening mammography and breast cancer mortality in women older than 70 years. *Ann Intern Med*. 2020;173(3):247. doi:10.7326/L20-0827
32. Groenwold RHH. Trial emulation and real-world evidence. *JAMA Netw Open*. 2021;4(3):e213845. doi:10.1001/jamanetworkopen.2021.3845
33. Loberg M, Lousdal ML, Bretthauer M, Kalager M. Benefits and harms of mammography screening. *Breast Cancer Res*. 2015;17:63. doi:10.1186/ s13058-015-0525-z
34. Baum M. Harms from breast cancer screening outweigh benefits if death caused by treatment is included. *BMJ*. 2013;346:f385. doi:10.1136/bmj.f385
35. Herrmann C, Vounatsou P, Thurlimann B, Probst-Hensch N, Rothermundt C, Ess S. Impact of mammography screening programmes on breast cancer mortality in Switzerland, a country with different regional screening policies. *BMJ Open*. 2018;8(3):e017806. doi:10.1136/bmjopen-2017-017806
36. Biller-Andorno N, Juni P. Abolishing mammography screening programs? A view from the Swiss Medical Board. *N Engl J Med*. 2014;370 (21):1965–1967. doi:10.1056/NEJMp1401875
37. Lee CI, Etzioni R. Missteps in current estimates of cancer overdiagnosis. *Acad Radiol*. 2017;24(2):226–229. doi:10.1016/j.acra.2016.05.020

**Supplement to:**

**Effectiveness of mammography screening on breast cancer mortality - a study protocol for emulation of target trials using German health claims data**

# Contents

# 1 Details on the statistical analysis

Discrete-time cumulative incidence functions (CIFs) will be estimated using the following approach. Let $A_{never}$ (reference), $A_{once}$, and $A_{regular}$ be indicator variables for the screening strategies "never screened", "screened at least at baseline", and "screened at baseline and every two years afterwards". The discrete-time (cause specific) hazard is modelled using pooled logistic regression adjusted for baseline covariates:

$$logit\left(\mathbb{P}\left(Y_{t+1}\middle|\bar{Y}_t = 0, \bar{C}_t = 0, \bar{D}_t = 0, A_{once}, A_{regular}, X\right)\right)$$

$$= f_1(\theta_1', t) + f_2(\theta_2', t, A_{once}) + f_3\left(\theta_3', t, A_{regular}\right) + \theta_4 A_{once} + \theta_5 A_{regular} + \theta_6' X'.$$

The above model includes flexible functions $f(.)$ of time $t$, regression coefficients $\theta$ for (transformed) time and, possibly, interaction terms between time and screening strategy. The functions $f(.)$ will be determined by visual inspection so that the unadjusted parametric CIF estimated via pooled logistic modelling approximates the non-parametric Aalen-Johansen curves reasonably well. The binary variable $Y_t$ denotes the outcome event breast cancer death at time $t$. The binary variable $C_t$ denotes censoring status at time $t$ and the binary variable $D_t$ contains the event status of the competing event (death by other causes) at time $t$. Baseline covariates and interactions between covariates are denoted by $X$. The prime notation $(.)'$ denotes vectors. The history of a variable is denoted by overbars as $\overline{(.)}$. The above model is a marginal structural model and contains baseline covariates, but no time-varying covariates. Adjustment for time-varying confounding by $X_t$ is achieved by inverse probability weighting, where time-varying weights are calculated for each screening strategy $A \epsilon \left\{A_{never}, A_{once}, A_{regular}\right\}$ separately as

$$W_t^A = \prod_{k=1}^{t} \frac{1}{\widehat{\mathbb{P}}(A_k|\bar{A}_{k-1}, \bar{X}_k, \bar{Y}_{k-1} = \bar{C}_{k-1} = 0)},$$

truncating weights at the 99th percentile. Here $A_k$ is the actual screening status at time k and is, by definition, consistent with the strategy $A$ as individuals will otherwise be censored. For efficiency the above weights can be replaced by stabilized weights (see Cain et al. (2010) for a description of stabilized weights). Analogous weights are used for censoring due to competing events when estimating the direct effect. Below, upper indices refer to counterfactuals, e.g. the probability of breast cancer death under screening even if a portion of the study subjects did not experience screening, i.e. exposure is set to a value possibly contrary to the observed exposure (Hernan & Robins, 2020). The cumulative incidence function $\widehat{CIF}_{i,t}^{A=a}$ for clone $i = 1, \dots, m$, at time point $t$ under screening strategy $A = a$ will then be estimated using one of the approaches (i.e. based on modelling either subdistribution or cause-specific hazard) described in Young et al. (2020), depending on computational cost. This cumulative incidence will be standardized to the empirical distribution of baseline confounders as

$$\widehat{CIF}_t^{A=a} = \frac{1}{m}\sum_{i=1}^{m} \widehat{CIF}_{i,t}^{A=a}.$$

As a function of time *t*, the above cumulative incidence function allows an assessment of how the effect of screening evolves over the whole of follow-up.

# 2 Illustration of assignment to screening strategies

| | |
|---|---|
| S: | screening in 1st quarter |
| C: | breast cancer diagnosis in 1st quarter |
| D: | death in 1st quarter |
| S&C: | screening & breast cancer diagnosis in 1st quarter |
| S&D: | screening & death in 1st quarter |

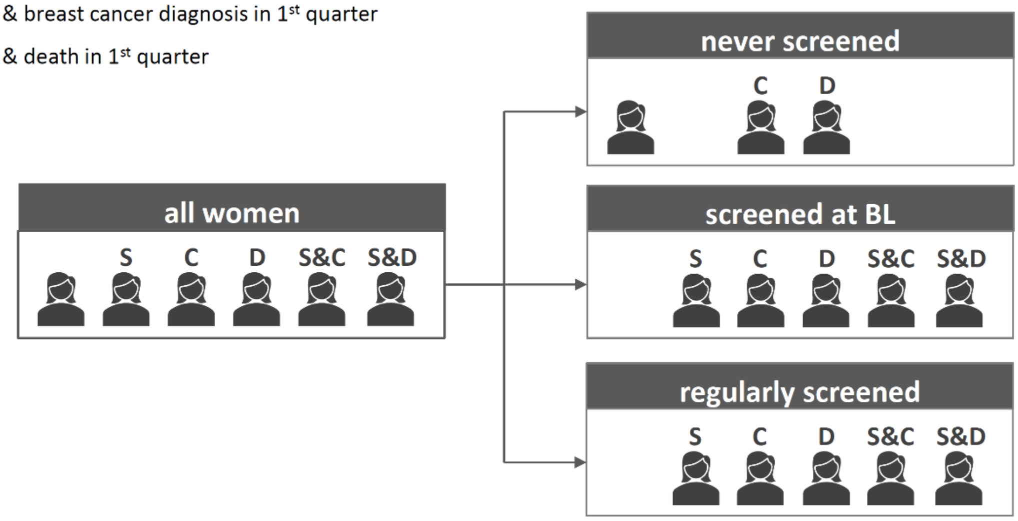


**Figure S1:** Illustration of cloning of women into the screening strategies. Assignment of clones to screening strategies is based on screening behaviour from the calendar quarter of baseline. Women with a breast cancer diagnosis or recorded death in the first quarter are cloned into all screening strategies, since they were compliant with all screening strategies until the diagnosis/death occurred.

# 3 Illustrations of artificial censoring schemes per screening strategy

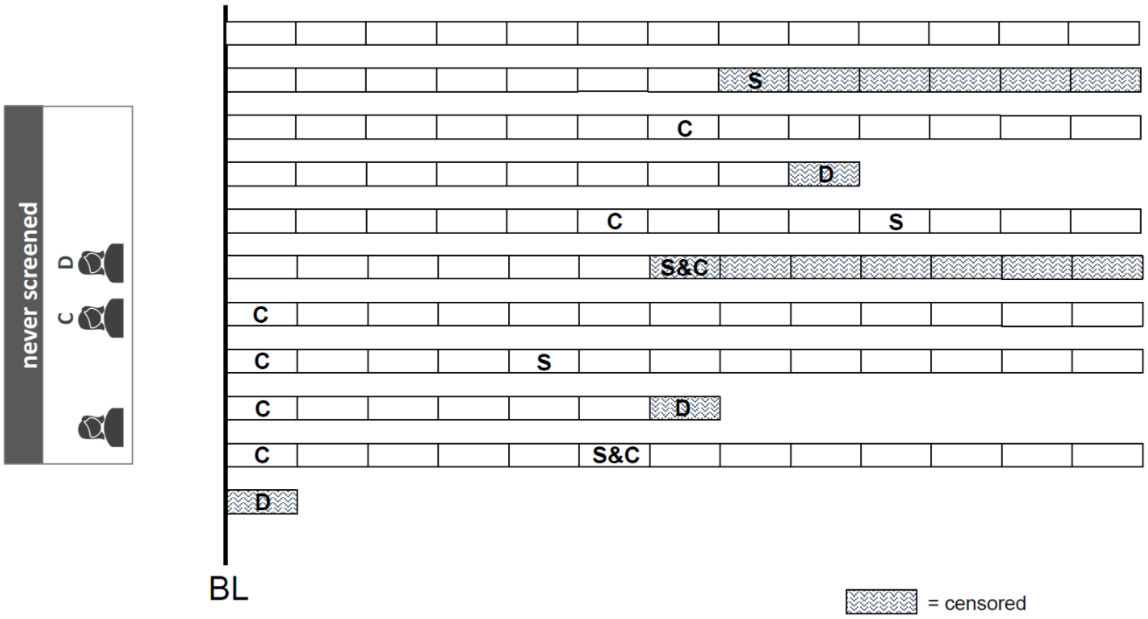


**Figure S2:** Illustration of artificial censoring scheme under screening strategy "never screened". Follow-up time is discretized into calendar quarters, with rectangles denoting individual quarters. The rationale for censoring is described in depth in the main body of the paper. Note that when a woman is censored, the time of censoring is set to the beginning of the calendar quarter that led to censoring. In the above illustration, the last woman is censored at baseline because she dies in the baseline quarter, i.e. she is censored at time point 0 with reason of censoring being death. C = breast cancer, D = death, S = screening, S&C = screening and cancer in the same quarter.

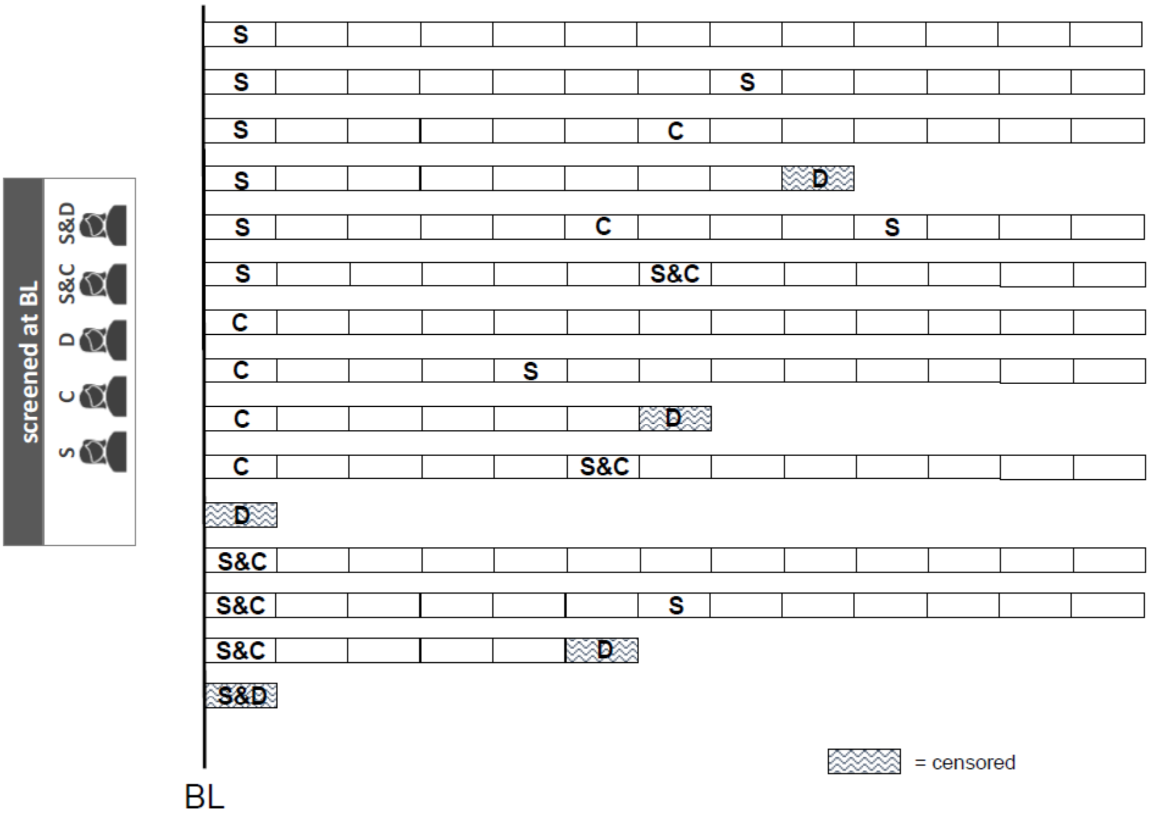


**Figure S3:** Illustration of artificial censoring scheme under screening strategy "screened at baseline". Follow-up time is discretized into calendar quarters, with rectangles denoting individual quarters. The rationale for censoring is described in depth in the main body of the paper. C = breast cancer, D = death, S = screening, S&C = screening and cancer in the same quarter, S&D = screening and death in the same quarter.

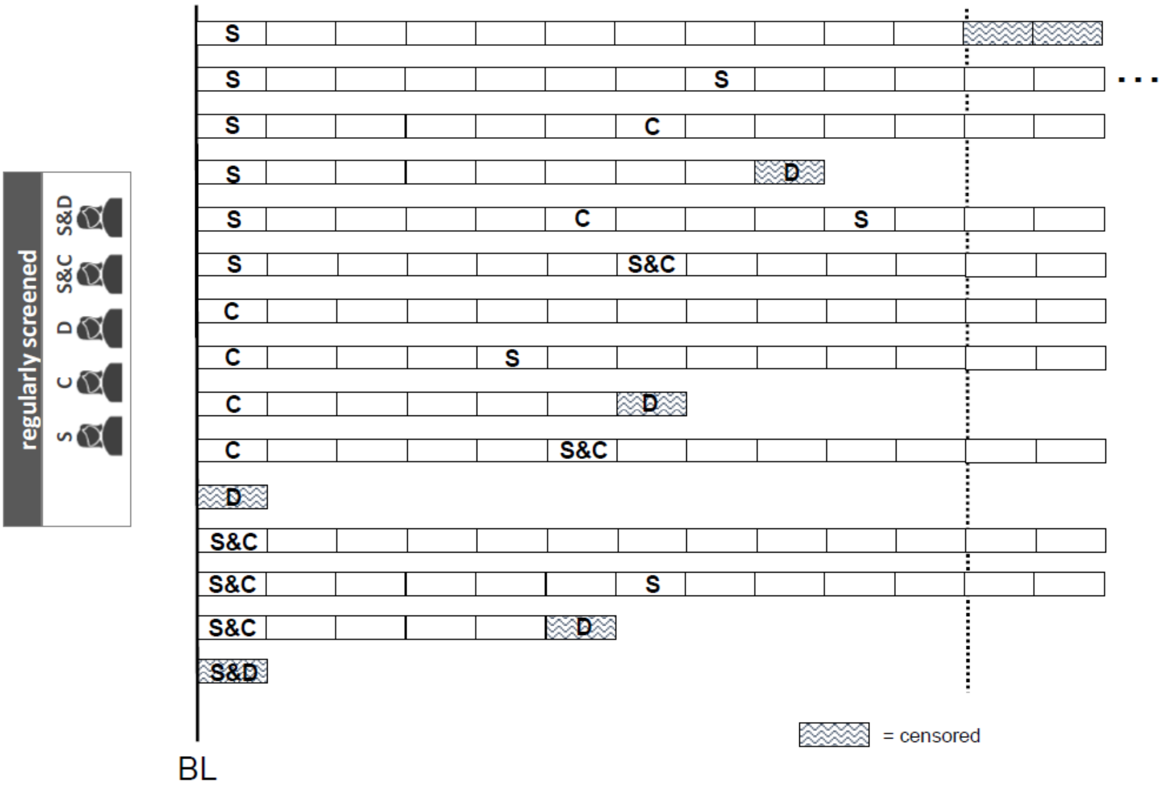


**Figure S4:** Illustration of artificial censoring scheme under screening strategy “regularly screened every two years”. Follow-up time is discretized into calendar quarters, with rectangles denoting individual quarters. A regular screening is defined as having taken place between one year to ten quarters after the previous screening. The rationale for censoring is described in depth in the main body of the paper. C = breast cancer, D = death, S = screening, S&C = screening and cancer in the same quarter, S&D = screening and death in the same quarter. The dotted line indicates the end of the time period in which the second screening would need to take place.

# 4 Addressing potential sources of bias

*Confounding*: Covariates used to adjust for confounding will be derived at baseline and during follow-up. Their selection is based on subject matter knowledge and available literature. Risk factors for breast cancer were considered relevant, even though the outcome variable is breast cancer mortality, since developing breast cancer is a necessary antecedent for breast cancer death. Figure S5 illustrates the causal considerations for covariate selection. Adjustment for confounding will be carried out via standardization and inverse probability weighting.

Given that claims data are not collected for research purposes, direct information on relevant confounders is not always available or only available for extreme cases (e.g. heavy smoking, alcohol abuse). We aim to minimize this problem by using indirect information on these confounders (e.g. diseases resulting from exposure to these risk factors such as smoking-related diseases, or diseases resulting mainly from unhealthy behaviour such as obesity) as well as proxy variables for a health-seeking behaviour (e.g. utilization of preventive services, educational attainment). With respect to family history of breast cancer, the information is restricted to the ICD-10-GM code Z80.3 (“malignant neoplasm of the breast in the family”). It is not clear whether it is primarily coded in patients with a hereditary breast cancer syndrome rather than in those with a “simple” family history. The observed low proportion of women with Z80 codes (Braitmaier et al. 2022) indicates that it might only be used in high-risk subjects who would not be the target group of normal MSP screening. We therefore plan to conduct sensitivity analyses excluding women with this code. In addition, we will conduct a quantitative bias analysis to estimate the impact of unmeasured confounding regarding a “simple” family history of breast cancer.

For some risk factors, no information will be available in our data, for example age at menarche, parity, age at first full-term pregnancy, breastfeeding, age at menopause, height, breast density, exposure to radiation (unrelated to mammography). However, we argue that these risk factors are relatively unknown to the public and it is therefore reasonable to assume that they do not influence the decision to undergo screening.

*Healthy screenee bias*: Individuals volunteering for screening are generally healthier than individuals who choose not to undergo screening (Weiss & Rossing 1996). In addition to adjustment for confounding, we will address this specific issue by carrying out a subgroup analysis within screening-affine women, defined by their pre-baseline use of other preventive services (research question 2). This subpopulation is more homogenous regarding health seeking behaviour, and we expect an increased internal validity albeit at the cost of generalizability. Therefore, both effects, the one in the full study sample and the one in the subgroup of screening-affine women, will be important for the evaluation of the screening programme.

*Competing events*: Death due to causes other than breast cancer is a competing event for the outcome of interest. We will compare the total effect (where death due to other causes is not treated as eliminable) with the direct effect of screening (where the competing event is treated as eliminable and thus censored with appropriate inverse probability of censoring weights, IPCW). Note that adjustment for confounding of the direct effect must also include common causes of the competing event and the study outcome, e.g. by including comorbidities (Young et al. 2020).

*Time-related biases*: Immortal time and other biases will be minimized by aligning eligibility checks and treatment assignment at time zero, i.e. baseline (Dickerman et al. 2019). Furthermore, women whose screening behaviour in the first quarter after trial start is consistent with more than one screening strategy will be copied and one clone will be assigned to each eligible screening strategy, i.e. women who undergo screening in the baseline quarter will be assigned to all active screening strategies. An alternative, but less efficient approach would be to randomly assign each person to exactly one of the eligible strategies (Garcia-Albeniz et al. 2020). Given that some information in the database used for this study is only available on a quarterly basis (e.g. outpatient diagnosis codes), it is impossible to break down the information into smaller time intervals than quarters. However, the length of follow-up required to observe the effect of screening is large (approx. 7 - 10 years) (Jatoi & Miller 2003). We therefore argue that the extent of bias due to the time units is negligible, as a delay of diagnosis of three months is unlikely to influence the screening effect.

*Misclassification*: Health claims data is primarily generated for reimbursement purposes and, therefore, some diagnosis codes might be used inappropriately for the underlying condition or over-used (e.g. diagnosis codes in the outpatient setting). To minimize misclassification, we define most of the diseases based on algorithms that, for example, combine different sources of information (e.g. diagnosis codes in combination with therapy), only use codes with a high validity (such as inpatient diagnosis codes) or only consider codes if recorded repeatedly. There may still be some misclassification of morbidity, but we consider this type of misclassification unlikely to differ between groups and negligible in our analysis. Risk factors that have a delayed impact on breast cancer may not be measured adequately due to a limited length of the available look-back period. For instance, HRT might influence breast cancer risk only after several years. Thus, a woman who stopped HRT treatment five years before baseline would be misclassified as “no HRT use” if her look-back period in the data is only three years. We will systematically describe the available look-back period (stratified by age at baseline) to assess whether this could be a relevant misclassification. Finally, misclassification of the outcome variable of breast cancer related deaths might occur since this variable is not directly available for much of the data and must be derived based on an algorithm. Langner et al. (2019) reported a sensitivity of 91.3 % and a specificity of 97.4 % for a former version of this algorithm, which is currently being further optimized and will be validated again based on a sample for which the official cause of death is available.

*Identifying assumptions*: We make the usual assumptions for causal inference from observational data, namely consistency, sequential exchangeability given observed covariates, and positivity. Consistency is fulfilled when the screening strategies being assessed are well-defined and correspond to the screening behaviour observed in the data, e.g. the outcome for a woman who happens to never undergo screening is the same as if she had been assigned to never undergo screening in the target trial. Sequential exchangeability is fulfilled when the observed screening behaviour of a woman at time *t* is independent of her potential outcomes under the strategies given the measured covariates prior to *t;* this can be thought of as no unmeasured baseline or time-varying confounding. Positivity is fulfilled when the probability of observing a screening strategy is greater than zero for all strategies in all covariate strata (Young et al. 2020, Hernan & Robins 2020). Furthermore, censoring competing events to obtain the direct effect requires an assumption of no unmeasured common causes of the different event types.

# 5 Illustration of causal considerations for covariate selection

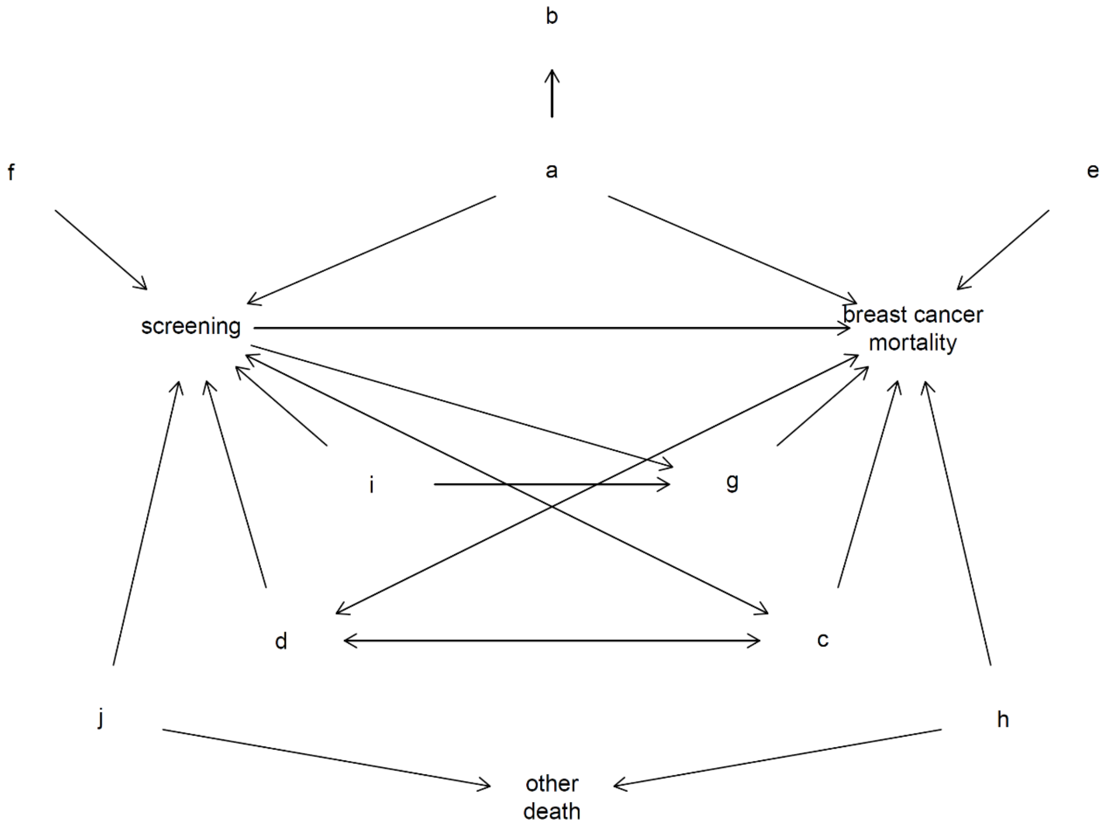


**Figure S5:** Illustration of variable groups considered for covariate adjustment and their causal connections. Note that this is a simplified graph, ignoring the longitudinal aspect of the study. A directed edge from one variable to another means that the first variable is a direct cause of the second. Screening is the exposure, breast cancer death is the outcome, and other death is a competing event. A bi-directed edge can be interpreted as presence of latent variables between the two connected variables. Variables "a" are common causes of screening and outcome. Variables "b" are proxies for those of category "a". Variables "c" are causes of the outcome that are associated with exposure. Variables "d" are causes of the exposure that are associated with the outcome. Variables "e" are causes of the outcome that are not associated with exposure. Variables "f" are causes of the exposure that are not associated with the outcome. Variables "g" are post-screening variables that are mediators between exposure and outcome. Variables "h" have a causative effect both on the competing event and the outcome. Variables "i" are causes of exposure and mediators. Variables "j" are confounders between exposure and the competing event. Variables "f" should not be included for adjustment, as this can lead to bias-amplification in case of residual unobserved confounding. Variables "g" (e.g. treatment after screening) should not be included for adjustment, as they are on the causal path from exposure to outcome. Variables "a", "b" (if "a" is unmeasured), "c", "d", "h" (only for estimating the direct effect, not for the total effect), "i", and "j" should be included for adjustment to mitigate confounding. Variables "e" are not needed for adjustment but can be included to increase precision of estimation. The variable groups (except "f") are not mutually exclusive, and in fact many variables will fit into more than one of these groups. An example of a covariate of the category "a" would be previous use of menopausal hormone therapy, as this is a known risk factor for breast cancer and physicians might advise women with this risk factor to attend screening. An example of a covariate of the category "j" would be presence of palliative care. An example for "d" might be educational attainment as it may affect awareness of screening and is strongly associated with direct risk factors "c" of breast cancer mortality; educational attainment can also be seen as type "b" proxy for further unmeasured confounders.

# 6 List of covariates

In Table S1 below, we give an overview of variables used to adjust for confounding. Time-varying covariates will be re-assessed on a quarterly basis. Variables might be added to this list of covariates, if indicated by subject matter knowledge. The list of covariates used in the final analysis will be finalized before data on the study outcome becomes available. Note that this is just an alphabetical list of covariates that will be defined based on the information in the database. Content-wise, a discussion on how confounding as a potential source of bias is considered and how relevant covariates are captured in the data is provided in Supplement 4. Furthermore, Figure S5 illustrates the causal considerations for covariate selection.

The covariates in Table S1 are mostly implemented as binary (time-dependent) variables. For most of the variables, algorithms considering different types of information (e.g. diagnosis codes in combination with therapy) will be developed or have been developed, with the aim of maximizing validity and thus minimizing misclassification (see also Supplement 4).

**Table S1:** Relevant covariates for confounder adjustment.

| **variable/variable group** | **time-varying** |
|---|---|
| Acute hemorrhagic stroke | yes |
| Acute ischemic stroke | yes |
| Acute myocardial infarction | yes |
| Age at baseline | no |
| Alcohol abuse | yes |
| Anaemia | yes |
| Anticoagulant therapy | yes |
| Antihypertensive therapy | yes |
| Antiplatelet therapy | yes |
| Benign neoplasm of breast | yes |
| Breast disorders (benign mammary dysplasia, inflammatory disorders of breast, hypertrophy of breast, unspecified lump in breast, other disorders) | yes |
| Bronchial asthma | yes |
| Cachexia | yes |
| Chronic obstructive pulmonary disease (COPD) | yes |
| Coronary heart disease | yes |
| Dementia | yes |
| Diabetes with end organ damage | yes |
| Drug abuse | yes |
| Drug-treated (arterial) hypertension | yes |
| Educational attainment | no |
| Family history of breast cancer* | yes |
| Glaucoma | yes |
| Heart failure | yes |
| Hemiplegia | yes |
| Hepatitis B or C | yes |
| Hip fracture | yes |
| HIV therapy | yes |

| | |
|---|---|
| Hormone replacement therapy | yes |
| Lipid-lowering therapy | yes |
| Liver diseases including chronic viral hepatitis | yes |
| Mental diseases | yes |
| Number of hospitalizations | yes |
| Number of non-screening mammographies | yes |
| Number of outpatient physician contacts | yes |
| Number of prescriptions | yes |
| Number of screening mammographies | yes |
| Obesity/adiposity | yes |
| Other cancers | yes |
| Palliative care | yes |
| Severe liver disease | yes |
| Terminal renal disease | yes |
| Tobacco abuse | yes |
| Treated diabetes | yes |
| Treatment for hypothyroidism | yes |
| Treatment for osteoporosis | yes |
| Treatment with antidepressants | yes |
| Treatment with antipsychotics | yes |
| Treatment with immunosuppressive drugs | yes |
| Treatment with opioids | yes |

* Given that information on family history is limited, additional methods will be taken to consider this confounder (see manuscript and Supplement 4).

# 7 References

Braitmaier M, Sarina S, Kollhorst B, et al. (2022) Screening colonoscopy similarly prevented distal and proximal colorectal cancer; A prospective study among 55-69-year-olds; Journal of Clinical Epidemiology; 149: 118-126

Cain LE, Saag MS, Petersen M, et al. (2016) Using observational data to emulate a randomized trial of dynamic treatment-switching strategies: an application to antiretroviral therapy; Int J Epidemiol; 45(5):2038-2049

Dickerman BA, Garcia-Albeniz X, Logan RW, et al. (2019) Avoidable flaws in observational analyses: an application to statins and cancer; Nat Med; 25(10): 1601 – 1606

Garcia-Albeniz X, Hernan MA, Hsu J (2020) Continuation of annual screening mammography and breast cancer mortality in women older than 70 years; Ann Intern Med; 173(3): 247

Hernan MA, Robins JM (2020) Causal Inference – What If. Boca Raton, FL: Chapman & Hall/CRC

Jatoi I, Miller AB (2003) Why is breast cancer mortality declining?; Lancet Oncol; 4(4): 251 – 254

Langner I, Ohlmeier C, Haug U (2019) Implementation of an algorithm fort he identification of breast cancer deaths in German health insurance claims data: a validation study based on record linkage with administrative mortality data; BMJ Open; 9(7): e026834

Weiss NS, Rossing MA (1996) Healthy screenee bias in epidemiological studies of cancer incidence; Epidemiology; 7(3): 319 – 322

Young JG, Stensrud MJ, Tchetgen Tchetgen EJ, et al. (2020) A causal framework for classical statistical estimands in failure-time settings with competing events; Statistics in Medicine; 39(8): 1199 – 1236

## 7.5 Paper 4: 13-year colorectal cancer risk after low-quality, high-quality and no screening colonoscopy: a cohort study

This paper was under review for publication in the Journal of Clinical Epidemiology when this thesis was submitted. The manuscript was removed from the published version of this thesis as to avoid conflicts with future copyright agreements. Only the draft abstract is printed here.

## 13-year colorectal cancer risk after low-quality, high-quality and no screening colonoscopy: a cohort study

**Short title:** Colonoscopy quality and colorectal cancer

Sarina Schwarz[1], Malte Braitmaier[2], Christian Pox[3], Bianca Kollhorst[2], Vanessa Didelez[2,4], Ulrike Haug[1,5]

1. Department of Clinical Epidemiology, Leibniz Institute for Prevention Research and Epidemiology – BIPS, Bremen, Germany
2. Department of Biometry and Data Management, Leibniz Institute for Prevention Research and Epidemiology – BIPS, Bremen, Germany
3. Department of Medicine St. Joseph-Stift Bremen, Bremen, Germany
4. Faculty of Mathematics and Computer Science, University of Bremen, Germany
5. Faculty of Human and Health Sciences, University of Bremen, Germany

## Abstract

**Objective:** A low-quality colonoscopy has been shown to be less effective in reducing colorectal cancer (CRC) incidence than a high-quality colonoscopy, but the comparison with no screening colonoscopy is lacking. We aimed to compare the 13-year risk of developing CRC between persons with I) a high-quality screening colonoscopy, II) a low-quality screening colonoscopy and III) without a screening colonoscopy. **Study Design and Setting:** A healthcare database ( 20% of the German population) was used to emulate a target trial with three arms: High-quality screening colonoscopy (highQualSC) vs. low-quality screening colonoscopy (lowQualSC) vs. no screening colonoscopy (noSC) at baseline. The quality of screening colonoscopy was categorized based on the polyp detection rate of the examining physician (cut-off of 21.8%). We included persons aged 55 to 69 years at average CRC risk and CRC screening naïve at baseline. We estimated adjusted cumulative CRC incidence over 13 years of followup. **Results:** The highQualSC arm comprised 142,960 persons, the lowQualSC arm 62,338 persons and the noSC arm 124,040 persons. The adjusted 13-year CRC risk was 1.77% in the highQualSC arm, 2.09% in the lowQualSC arm and 2.74% in the noSC arm. Compared to the noSC arm, the adjusted relative risk was 0.76 (95% CI: 0.70–0.84) in the lowQualSC arm and 0.65 (95% CI: 0.60–0.69) in the highQualSC arm. **Conclusion:** Our study shows that a low-quality screening colonoscopy is also effective in reducing CRC incidence compared to no screening colonoscopy. However, the effect is about one third less than that of a high-quality screening colonoscopy.

## 7.6 Paper 5: Misleading and avoidable: design-induced biases in observational studies evaluating cancer screening—the example of screening colonoscopy

This paper was posted as a preprint to medRxiv.org (DOI: 10.1101/2024.04.29.24306522) under a CC-BY 4.0 license. The manuscript is printed below.

# Misleading and avoidable: design-induced biases in observational studies evaluating cancer screening—the example of site-specific effectiveness of screening colonoscopy

**Authors**: Malte Braitmaier, Sarina Schwarz, Vanessa Didelez, Ulrike Haug

**Author information:**

Malte Braitmaier, Department of Biometry and Data Management, Leibniz Institute for Prevention Research and Epidemiology – BIPS, Bremen, Germany, ORCID: 0000-0001-7534-4068

Sarina Schwarz, Department of Clinical Epidemiology, Leibniz Institute for Prevention Research and Epidemiology – BIPS, Bremen, Germany, ORCID: 0000-0002-7926-2032

Vanessa Didelez, Department of Biometry and Data Management, Leibniz Institute for Prevention Research and Epidemiology – BIPS, Bremen, Germany & Faculty of Mathematics and Computer Sciences, University of Bremen, Bremen, Germany, ORICD: 0000-0001-8587-7706

Ulrike Haug, Department of Clinical Epidemiology, Leibniz Institute for Prevention Research and Epidemiology – BIPS, Bremen, Germany & Faculty of Human and Health Sciences, University of Bremen, Bremen, Germany, ORCID: 0000-0002-1886-2923

Corresponding author: Ulrike Haug, haug@leibniz-bips.de

## Abstract

**Objective**: Observational studies evaluating the effectiveness of cancer screening are often biased due to an inadequate design where I) the assessment of eligibility, II) the assignment to screening vs. no screening and III) the start of follow-up are not aligned at time zero (baseline). Such flaws can entail misleading results but are avoidable by designing the study following the principle of target trial emulation (TTE). We aimed to illustrate this by addressing the research question whether screening colonoscopy is more effective in the distal vs. the proximal colon.

**Methods**: Based on a large German health care database (20% population coverage), we assessed the effect of screening colonoscopy in preventing distal and proximal CRC over 12 years of follow-up in 55–69-year-old persons at average CRC risk. We applied four different study designs and compared the results: cohort study with / without alignment at time zero, case control study with / without alignment at time zero.

**Results**: In both analyses with alignment at time zero, screening colonoscopy showed a similar effectiveness in reducing the incidence of distal and proximal CRC (cohort analysis: 32% (95% CI: 27% - 37%) vs. 28% (95% CI: 20% - 35%); case-control analysis: 27% vs. 33%). Both analyses without alignment at time zero suggested a difference in site-specific performance: Incidence reduction regarding distal and proximal CRC, respectively, was 65% (95% CI: 61% - 68%) vs. 37% (95% CI: 31% - 43%) in the cohort analysis and 77% (95% CI: 67% - 84%) vs. 46% (95% CI: 25% - 61%) in the case-control analysis.

**Conclusions**: Our study demonstrates that violations of basic design principles can substantially bias the results of observational studies on cancer screening. In our example, it falsely suggested a much stronger preventive effect of colonoscopy in the distal vs. the proximal colon. The difference disappeared when the same data were analyzed using a TTE approach, which is known to avoid such design-induced biases.

## Introduction

Randomized controlled trials (RCT) are the gold standard for evaluating the effectiveness of cancer screening. However, existing RCTs in this field do not answer all relevant research questions. For screening colonoscopy, for example, an RCT has recently been published (NordICC trial) demonstrating its effectiveness in reducing colorectal cancer (CRC) incidence overall [1], but it was not powered to compare the effectiveness in the distal vs. the proximal colon.

Complementary evidence from observational studies is therefore needed. Apart from potential confounding, there is a high risk of bias and thus of misleading results if such studies are inadequately designed. Indeed, several observational studies have reported a markedly stronger preventive effect of screening colonoscopy in the distal as compared to the proximal colon [2, 3, 4], while a cohort study designed following the principle of target trial emulation (TTE) showed a similar effectiveness of screening colonoscopy in the distal and the proximal colon [5]. We argued that the difference by site in the former studies was due to biases induced by non-alignment at “time zero”, i.e. at baseline. This means that I) the assessment of eligibility, II) the assignment to study arms and III) the start of follow-up were not aligned as they would be in an RCT and as it would be ensured in an observational study designed based on the principle of TTE [6]. Specifically, previous studies often defined exposure based on pre- or post-baseline information on colonoscopy. As we further argued, this lack of alignment in previous studies led to overestimating the effectiveness of screening colonoscopy. Due to the different age pattern of distal and proximal CRC, this bias affected distal CRC more than proximal CRC, i.e. the difference in effectiveness by site was an artefact.

To demonstrate this, we compared different study designs with and without alignment at time zero aiming to investigate the question of site-specific effectiveness of screening colonoscopy in reducing CRC incidence. For the two designs without alignment we used a cohort study design, where the assignment to study arms occurs *before* time zero (pre-baseline), and a nested case control study design, where the assignment to study arms occurs *after* time zero (post-baseline). The current paper is part of a growing literature identifying violations of alignment at time zero as a potential source of major bias in observational studies [6, 7, 8].

## Methods

### *Data source and study population*

We used the German Pharmacoepidemiological Research Database (GePaRD) which comprises claims data from four statutory health insurance providers in Germany and covers about 20% of the German population [9]. In GePaRD, information on utilization of screening colonoscopy, offered in Germany to persons aged 55 or older since 2002 (since 2019 also to men aged 50-54), is distinguishable from diagnostic colonoscopy. As previously described, the data source enables the valid identification of incident CRCs [10]. Furthermore, it contains appropriate information to apply in- and exclusion criteria and to adjust for confounding as relevant to the research question on the effectiveness of screening colonoscopy in reducing CRC incidence [5]. For the present study, we used data from 2004 to 2020.

Based on this data source, we applied four different study designs to address the research question, specifically a cohort and a case-control study design, each with and without alignment at time zero. The study designs without alignment at time zero were inspired by published examples [2, 11, 12, 13], and were partly complemented by sensitivity analyses. For each of these four studies, persons were selected from the same population. Specifically, the source population was a cohort of persons aged 55–69 at baseline, who were continuously insured for at least three years before baseline.

### *Cohort study without alignment at time zero*

The cohort started in 2009 (baseline). Similar to a previous study [2], individuals were assigned to the screening colonoscopy arm if they had a screening colonoscopy any time before baseline, including the baseline quarter. Individuals were assigned to the control arm if they did not undergo screening colonoscopy any time before baseline, including the baseline quarter. In a sensitivity analysis, we considered both screening and diagnostic colonoscopies for the assignment to the study arms, because some of the previous studies did not distinguish between these examinations. Eligibility criteria were checked at baseline and the outcome variable (incident CRC) was assessed beginning with baseline (start of follow-up). Persons were followed up until end of study period (end of 2020), end of continuous insurance coverage, death or CRC diagnosis, whichever occurred first. We also conducted sensitivity analyses starting the cohort in 2010 and 2011, respectively.

When using such a study design, the assessment of eligibility and the start of follow-up are aligned, but the assignment to the screening and the control arm is based on a period *before* time zero (pre-baseline). Specifically, individuals in the colonoscopy arm had the examination in the past (i.e. they were assigned to the screening arm based on past exposure) rather than *at* time zero.

*Cohort study with alignment at time zero*

As described previously [5], we emulated sequential trials for each calendar quarter from 2007 to 2011. The emulation of sequential target trials makes full use of the information from longitudinal data without violating principles of study design by using pre- or post-baseline information for the assignment to study arms. At the baseline quarter of each trial, eligibility was assessed and individuals with previous screening colonoscopy or CRC diagnosis were excluded. Individuals were then assigned to the screening arm if they underwent a screening colonoscopy *in the baseline quarter* of the respective trial and to the control arm otherwise. Individuals were followed up until end of study period (end of 2020, i.e. follow-up was longer than in our previous analysis), end of continuous insurance coverage, death or CRC diagnosis, whichever occurred first. This study design made sure that assessment of eligibility criteria, assignment to the screening and control arm, and start of follow-up were aligned at time zero as would be the case in an RCT.

*Case-control study without alignment at time zero*

We applied a case-control design frequently used in the published literature [11, 12, 13, 14, 15, 16]. Essentially, CRC cases are identified (date of diagnosis corresponds to index date) and matched with controls free of CRC at index date. Then screening colonoscopy use *ever before* or within a certain time period *before* the index date is assessed in cases and controls, i.e. colonoscopies leading to CRC diagnosis are not considered as exposure in this type of study. Here, we selected all individuals from the source population entering the cohort in 2009 with a CRC diagnosis in 2018-2020. For each case we matched up to five controls on age (+/- one year) and sex (sampling without replacement). The exposure variable was then defined as any screening colonoscopy between 2009 and the index date, i.e. exposure to colonoscopy use was assessed within 10-12 years *before* the index date. Colonoscopies conducted in the six months before CRC diagnosis were not considered in defining the exposure. As mentioned above, this approach corresponds to published case-control studies which ignore colonoscopies conducted as part of the diagnostic process leading to the current diagnosis [11, 12, 13, 14, 15, 16]. In general, it is a fundamental characteristic of traditional case-control studies to assess exposure before disease onset. In a sensitivity analysis, we considered both screening and diagnostic colonoscopies for the assignment to exposure groups. Again, we also conducted sensitivity analyses using the years 2010 and 2011 for cohort entry, i.e. the source population underlying this nested case-control study.

In the case-control design we used here (nested within a cohort), the assessment of eligibility and the start of follow-up were aligned, while the assignment to the screening and the control arm occurred *after* time zero (post-baseline) instead of *at* time zero. Note that in case-control studies not nested in a cohort, there typically are additional misalignments [11, 14]. Specifically, eligibility is assessed at index date and the start of follow-up is unclear.

*Case-control study with alignment at time zero*

Following the approach described by Dickerman et al. [17], a case-control study was nested within the original cohort of sequential emulated target trials, and colonoscopy use was assessed at baseline of each emulated trial. We included CRC patients with an incident CRC diagnosis at any point during follow-up (until 2020) and then used risk set sampling to match up to five controls to each case. We sampled matched controls with replacement, i.e. the same control could be matched to more than one case. Matching variables were the same as above. The key difference to the case-control study without alignment is that exposure assignment was based on information available at the start of the emulated trial, i.e. at time zero, instead of information occurring after time zero. This approach has been shown to avoid self-inflicted biases in the same way as a prospective study using TTE [17].

### *Data analysis*

For the cohort studies, we estimated cumulative incidence functions (CIF) via pooled logistic regressions, which were adjusted for baseline confounders via inverse probability of treatment weighting. Effects were estimated as adjusted relative risks (RR) at the end of follow-up based on these CIFs. As previously shown, adjustment yielded satisfactory covariate balance and a negative control analysis did not indicate any residual confounding [5]. Confidence intervals were estimated via person-level bootstrap. For the case-control studies, effects were estimated as adjusted odds ratios (ORs) obtained via conditional logistic regression. For the case-control analysis with alignment, no confidence intervals could be obtained due to computational limitations: The emulation of sequential trials with repeated cohort entry would require bootstrapping, where matching is repeated for every bootstrap sample, resulting in run times of several months.

## Results

### *Cohort study without alignment at time zero*

We selected a random sample of 200,000 individuals in the control arm and 200,000 individuals in the screening colonoscopy arm. The adjusted relative risk after 12 years of follow-up was 0.35 for distal CRC and 0.63 for proximal CRC (Table 1). The adjusted cumulative incidence curves are given in Fig. **1**. As shown in Supplement 1, results were similar when the year 2010 or the year 2011 was used as baseline. In sensitivity analyses considering both screening and diagnostic colonoscopies as exposure, the adjusted 12-year relative risk was 0.40 for distal CRC and 0.66 for proximal CRC (Supplement 2).

*Cohort study with alignment at time zero*

Overall, 192,054 persons were included in the screening colonoscopy arm. The 5% random sample (restriction due to computational limitations) of controls assigned to the no screening arm included 116,452 persons (1,241,071 non-unique). The adjusted relative risk after 12 years of follow-up was 0.68 for distal CRC and 0.72 for proximal CRC (Table 1). Figure 1 shows the adjusted cumulative incidence curves for distal and proximal CRC. The distribution of screen-detected and post-colonoscopy CRCs (i.e. non-screen-detected CRCs) by site is shown in Supplement 6.

*Case-control study without alignment at time zero*

Overall, 446 cases with distal CRC matched to 2,230 controls and 302 cases with proximal CRC matched to 1,510 controls were included. The adjusted ORs for distal and proximal CRC were 0.23 and 0.54, respectively (Table 2). When the year 2010 or the year 2011 was used to define the source population, the difference by site was similar (Supplement 1). The sensitivity analysis considering both screening and diagnostic colonoscopy as exposure yielded similar results; the adjusted ORs for distal and proximal CRC were 0.20 for distal CRC and 0.44 for proximal CRC, respectively (Supplement 2).

*Case-control study with alignment at time zero*

Overall, 8,382 cases with distal CRC matched to 40,925 controls and 4,463 cases with proximal CRC matched to 22,175 controls were included. The adjusted ORs for distal and proximal CRC were 0.73 and 0.67, respectively (Table 2).

## Discussion

To the best of our knowledge, our study is the first to systematically compare different study designs to assess the effectiveness of screening colonoscopy in reducing CRC incidence in the distal vs. the proximal colon. Our cohort and case-control analyses with alignment at time zero showed no relevant difference in the effectiveness by site. Using study designs without alignment at time zero led to an overestimation of the effectiveness of screening colonoscopy overall. The overestimation affected distal CRCs considerably more than proximal CRCs , i.e. purely by design there appeared to be a difference in effectiveness by site. This finding held up in sensitivity analyses varying data years and the type of examinations considered for the exposure definition (only screening or also diagnostic colonoscopy). Our findings demonstrate that the difference in the effectiveness of colonoscopy by site reported by previous observational studies was due to bias introduced by inadequate study design.

As illustrated in Supplement 3 using directed acyclic graphs, the bias underlying studies using pre-baseline information on colonoscopy for the assignment to study arms can be expressed as a form of collider stratification bias [18, 19]. To give an intuitive explanation, let us revisit the study by Guo et al. [2, 5]: At baseline, patients were asked about past colonoscopy use and—based on this information—assigned as exposed or unexposed to colonoscopy. Persons reporting a prior CRC diagnosis at baseline were excluded [2]. Given that colonoscopy is one of the main tools by which CRC is diagnosed, this process removes individuals with previously diagnosed CRC from the exposed group, i.e. it enriches the exposed group with individuals who are known to be free of CRC. No such selection process takes place in the unexposed group. This leads to a lower prevalence of preclinical CRC at baseline in the exposed as compared to the unexposed group. As a consequence, this selection reduces the number of CRCs occurring during follow-up in the exposed group as compared to the unexposed group and thus leads to overestimation of the effect of screening on CRC incidence. As the vast majority of CRCs diagnosed at an age when persons are typically included into screening studies are in the distal colon [20] while proximal CRCs become more common at older age, this bias mainly affects results for distal CRC, i.e. as mentioned above there appeared to be a difference in effectiveness by site purely by design. We note that in addition to the initial exposure assignment, Guo et al. also used an updated exposure variable in a Cox model with time-dependent covariates. However, this does not correct the initial selection issue at the start of follow-up.

The above argument applies to studies using pre-baseline information for the assignment to exposure groups. Many other studies used post-baseline information for the assignment to exposure groups, also inducing bias. We illustrated this by the case-control study without alignment at time zero: Whenever after baseline CRC is detected in a person at his or her first colonoscopy, as is the case for most screen-detected CRCs, this person is assigned to the unexposed group as there was no previous colonoscopy and the actual colonoscopy detecting the CRC is not considered as prior exposure. This enriches the unexposed group with CRCs and thus leads to overestimation of the effectiveness of screening. As the majority of screen-detected CRCs are in the distal colon, this bias predominantly affects CRCs in the distal colon and thus leads to an artificial difference in the effectiveness of colonoscopy by site (see also Supplement 4). In our case-control study design embedded in an emulated target trial with alignment at time zero, in which screen-detected CRCs are correctly assigned, no relevant difference in the effectiveness of colonoscopy by site was observed. Of note, misalignment due to post-baseline exposure assignment is typical of but not limited to case-control designs on cancer screening. It can also occur in inadequately designed cohort studies and is not overcome by using a time-varying exposure variable in a hazard model. This is explained in more detail in Supplement 5 based on the example of the study by Nishihara et al. [3]

In summary and more generally, both study designs without alignment at time zero have in common that there are mechanisms that lead to inappropriate consideration of screen-detected CRCs, i.e. in the screening arm there was no peak in CRC incidence immediately after baseline as it would be the case in an RCT. Of course, this overestimates the impact of screening on CRC incidence, particularly for distal CRC, as illustrated in Figure 1. The flawed approaches ignore the fact that a screening colonoscopy sometimes comes too late to prevent CRC. Following the publication of the NordICC study, there was a discussion whether it is appropriate to include persons with preclinical CRC, causing the peak at baseline, in a prevention trial [21, 22]. However, from a public health perspective, it is important to also take into account CRCs that are not prevented by screening in order to avoid overestimating the effectiveness of CRC screening at the population level. Apart from this, it should be noted that studies without alignment at time zero do not provide a valid answer to the question regarding the size of the preventive effect of colonoscopy in persons free of CRC at baseline.

It should be noted that, although we focus our discussion on biases most relevant for site-specific effectiveness of screening colonoscopy, misalignment at time zero should also be avoided for many other reasons. Rasouli et al. [23] demonstrated that time related issues such as prevalent user bias or time-varying confounding are a threat to case-control designs not embedded in an emulated target trial. Also Dickerman et al. showed—based on case-control studies investigating the impact of statins on CRC risk—the biases inherent to traditional case-control studies and the potential of avoiding bias and wrong conclusions if the study is designed following the principle of TTE [17]. Similarly, there are many examples of biases other than those we discussed here that are inherent to cohort studies without alignment at time zero [8].

Our findings have several implications. First, regarding research on CRC screening, previous studies suggesting a lower effectiveness of colonoscopy in the proximal colon stimulated a search for reasons that may explain the occurrence of post-colonoscopy CRCs specifically in the proximal colon. It was suggested that one main reason relates to sessile serrated lesions as they are more difficult to detect and more often occur in the proximal colon [24]. While we do not question the important role of these lesions, our findings may encourage a broadening of the discussion of potential reasons leading to post-colonoscopy CRCs. Indeed, in our emulated target trial on screening colonoscopy, the proportion of post-colonoscopy CRCs located in the distal vs. the proximal colon was rather similar (Supplement 6). A one-sided focus on lesions that occur more frequently in the proximal colon therefore seems too narrow regarding the identification of lesions possibly leading to post-colonoscopy CRCs.

Our results also have implications beyond the specific research question of our study. Observational data are often used to evaluate the effectiveness of cancer screening. They represent a valuable data source to complement RCT evidence in this field, as RCTs on cancer screening are scarce, were often conducted many years ago and are typically not powered to estimate, for example, subgroup-specific effects or differences by cancer subtype. However, our study illustrates that—in addition to appropriate control of confounding—it is of key importance to design these studies in a way to ensure alignment at time zero. This means that assessment of eligibility, assignment to the screening and control arm and start of follow-up must be aligned. Otherwise, there is a high risk of bias.

Specific strengths of our study include the systematic comparison of different study designs as well as the comprehensive sensitivity analyses. Given that all analyses were conducted using the same data source and referred to the same setting, there is no heterogeneity regarding, for example, the study variables or setting-related factors such as the uptake of surveillance colonoscopy or colonoscopy quality. This strengthens our conclusion that differing results of the analyses with and without alignment at time zero are exclusively due to the study design.

It should be noted that our findings apply to the population aged 55-69, covering the typical screening age range of CRC screening. Whether screening colonoscopy is equally effective in the distal and proximal colon in older age groups cannot be answered by our study, nor did we address the endpoint CRC mortality. These research questions were beyond our study's scope, as our primary objective was to illustrate the relevance of design-induced biases and the possibility to avoid them using TTE, exemplified by investigating site-specific effectiveness of screening colonoscopy in reducing CRC incidence.

In conclusion, our study demonstrates that violation of alignment at time zero can substantially bias the results of observational studies on cancer screening. In our example, it falsely suggested an almost doubled preventive effect of colonoscopy in the distal vs. the proximal colon. The difference disappeared when the same data were analyzed using a TTE approach, which is known to avoid design-induced biases.

## References

1 Bretthauer M, Loberg M, Wieszczy P, Kalager M, Emilsson L, Garborg K*, et al.* Effect of colonoscopy screening on risks of colorectal cancer and related death. N Engl J Med 2022;**387**:1547-56.

2 Guo F, Chen C, Holleczek B, Schottker B, Hoffmeister M, Brenner H. Strong reduction of colorectal cancer incidence and mortality after screening colonoscopy: prospective cohort study from Germany. Am J Gastroenterol 2021;**116**:967-75.

3 Nishihara R, Wu K, Lochhead P, Morikawa T, Liao X, Qian ZR*, et al.* Long-term colorectal-cancer incidence and mortality after lower endoscopy. N Engl J Med 2013;**369**:1095-105.

4 Brenner H, Stock C, Hoffmeister M. Effect of screening sigmoidoscopy and screening colonoscopy on colorectal cancer incidence and mortality: systematic review and meta-analysis of randomised controlled trials and observational studies. BMJ 2014;**348**:g2467.

5 Braitmaier M, Schwarz S, Kollhorst B, Senore C, Didelez V, Haug U. Screening colonoscopy similarly prevented distal and proximal colorectal cancer: a prospective study among 55-69-year-olds. J Clin Epidemiol 2022;**149**:118-26.

6 Garcia-Albeniz X, Hsu J, Hernan MA. The value of explicitly emulating a target trial when using real world evidence: an application to colorectal cancer screening. Eur J Epidemiol 2017;**32**:495-500.

7 Wakabayashi R, Hirano T, Laurent T, Kuwatsuru Y, Kuwatsuru R. Impact of "time zero" of Follow-Up Settings in a Comparative Effectiveness Study Using Real-World Data with a Non-user Comparator: Comparison of Six Different Settings. Drugs Real World Outcomes 2023;**10**:107-17.

8 Hernan MA, Sauer BC, Hernandez-Diaz S, Platt R, Shrier I. Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses. J Clin Epidemiol 2016;**79**:70-5.

9 Haug U, Schink T. German Pharmacoepidemiological Research Database (GePaRD). In: Sturkenboom M, Schink T, eds. Databases for pharmacoepidemiolpogical research. Cham, Switzerland: Springer, 2021:119-24.

10 Schwarz S, Hornschuch M, Pox C, Haug U. Colorectal cancer after screening colonoscopy: 10-year incidence by site and detection rate at first repeat colonoscopy. Clin Transl Gastroenterol 2023;**14**:e00535.

11 Baxter NN, Goldwasser MA, Paszat LF, Saskin R, Urbach DR, Rabeneck L. Association of colonoscopy and death from colorectal cancer. Ann Intern Med 2009;**150**:1-8.

12 Mulder SA, van Soest EM, Dieleman JP, van Rossum LG, Ouwendijk RJ, van Leerdam ME*, et al.* Exposure to colorectal examinations before a colorectal cancer diagnosis: a case-control study. Eur J Gastroenterol Hepatol 2010;**22**:437-43.

13 Doubeni CA, Weinmann S, Adams K, Kamineni A, Buist DS, Ash AS*, et al.* Screening colonoscopy and risk for incident late-stage colorectal cancer diagnosis in average-risk adults: a nested case-control study. Ann Intern Med 2013;**158**:312-20.

14 Brenner H, Chang-Claude J, Seiler CM, Rickert A, Hoffmeister M. Protection from colorectal cancer after colonoscopy: a population-based, case-control study. Ann Intern Med 2011;**154**:22-30.

15 Kahi CJ, Pohl H, Myers LJ, Mobarek D, Robertson DJ, Imperiale TF. Colonoscopy and colorectal cancer mortality in the veterans affairs health care system: a case-control study. Ann Intern Med 2018;**168**:481-8.

16 Baxter NN, Warren JL, Barrett MJ, Stukel TA, Doria-Rose VP. Association between colonoscopy and colorectal cancer mortality in a US cohort according to site of cancer and colonoscopist specialty. J Clin Oncol 2012;**30**:2664-9.

17 Dickerman BA, Garcia-Albeniz X, Logan RW, Denaxas S, Hernan MA. Emulating a target trial in case-control designs: an application to statins and colorectal cancer. Int J Epidemiol 2020;**49**:1637-46.

18 Greenland S. Quantifying biases in causal models: classical confounding vs collider-stratification bias. Epidemiology 2003;**14**:300-6.

19 Hernan MA, Hernandez-Diaz S, Robins JM. A structural approach to selection bias. Epidemiology 2004;**15**:615-25.

20 Meza R, Jeon J, Renehan AG, Luebeck EG. Colorectal cancer incidence trends in the United States and United kingdom: evidence of right- to left-sided biological gradients with implications for screening. Cancer Res 2010;**70**:5419-29.

21 Song M, Bretthauer M. Interpreting epidemiologic studies of colonoscopy screening for colorectal cancer prevention: understanding the mechanisms of action is key. Eur J Epidemiol 2023;**38**:929-31.

22 Brenner H, Heisser T, Cardoso R, Hoffmeister M. When gold standards are not so golden: prevalence bias in randomized trials on endoscopic colorectal cancer screening. Eur J Epidemiol 2023;**38**:933-7.

23 Rasouli B, Chubak J, Floyd JS, Psaty BM, Nguyen M, Walker RL*, et al.* Combining high quality data with rigorous methods: emulation of a target trial using electronic health records and a nested case-control design. BMJ 2023;**383**:e072346.

24 Crockett SD, Nagtegaal ID. Terminology, molecular features, epidemiology, and management of serrated colorectal neoplasia. Gastroenterology 2019;**157**:949-66 e4.

## Statements & Declarations

### *Funding*

BIPS intramural funding

### *Competing interests*

MB, SS, VD and UH are working at an independent, non-profit research institute, the Leibniz Institute for Prevention Research and Epidemiology – BIPS. Unrelated to this study, BIPS occasionally conducts studies financed by the pharmaceutical industry. These are post-authorization safety studies (PASS) requested by health authorities. The design and conduct of these studies as well as the interpretation and publication are not influenced by the pharmaceutical industry. The study presented was not funded by the pharmaceutical industry and was performed in line with the ENCePP Code of Conduct.

### *Author contributions*

MB: Conceptualization, data curation, formal analysis, investigation, methodology, software, visualization, writing – original draft, writing – review and editing; SS: Conceptualization, investigation, writing – review and editing; VD: Conceptualization, investigation, methodology, supervision, writing – review and editing; UH: Conceptualization, funding acquisition, investigation, methodology, project administration, resources, supervision, writing – original draft, writing – review and editing.

### *Ethics approval & consent to participate*

In Germany, the utilisation of health insurance data for scientific research is regulated by the Code of Social Law. All involved health insurance providers as well as the German Federal Office for Social Security and the Senator for Health, Women and Consumer Protection in Bremen as their responsible authorities approved the use of GePaRD data for this study. Informed consent for studies based on claims data is required by law unless obtaining consent appears unacceptable and would bias results, which was the case in this study. According to the Ethics Committee of the University of Bremen studies based on GePaRD are exempt from institutional review board review.

### *Data availability*

As we are not the owners of the data we are not legally entitled to grant access to the data of the German Pharmacoepidemiological Research Database. In accordance with German data protection regulations, access to the data is granted only to employees of the Leibniz Institute for Prevention Research and Epidemiology – BIPS on the BIPS premises and in the context of approved research projects. Third parties may only access the data in cooperation with BIPS and after signing an agreement for guest researchers at BIPS.

### *Acknowledgements*

We thank Anja Gabbert and Inga Schaffer for the programming of analysis datasets.

## Tables and Figures

*Table 1: Results of cohort study designs without and with alignment at time zero (adjusted for baseline covariates).*

| | *Control group* | *Screening group* | *Adjusted relative risk* | *(95% CI)* |
|---|---|---|---|---|
| Design without alignment at time zero | | | | |
| Number at risk | 200,000 | 200,000 | | |
| Number of CRC cases | | | | |
| Distal CRC | 2,472 | 829 | 0.35 | (0.32-0.39) |
| Proximal CRC | 1,290 | 823 | 0.63 | (0.57-0.69) |
| Design with alignment at time zero | | | | |
| Number at risk | 1,241,071 | 192,054 | | |
| Number of CRC cases | | | | |
| Distal CRC | 16,750 | 1,678 | 0.68 | (0.63-0.73) |
| Proximal CRC | 8,548 | 919 | 0.72 | (0.65-0.80) |

*Table 2: Results of case-control study designs without and with alignment at time zero.*

| *Site* | *Case status* | | *Adjusted OR* | |
|---|---|---|---|---|
| | *Cases* | *Controls* | | *(95% CI)§* |
| Design without alignment at time zero | | | | |
| Number of distal CRCs / controls | 446 | 2,230 | | |
| Thereof exposed | 36 | 653 | 0.23 | (0.16-0.33) |
| Number of proximal CRCs / controls | 302 | 1,510 | | |
| Thereof exposed | 54 | 434 | 0.54 | (0.39-0.75) |
| Design with alignment at time zero | | | | |
| Number of distal CRCs / controls | 8,382 | 40,925 | | |
| Thereof exposed | 799 | 5,695 | 0.73 | |
| Number of proximal CRCs / controls | 4,463 | 22,175 | | |
| Thereof exposed | 409 | 3,013 | 0.67 | |

§: Confidence intervals could not be obtained for the case-control analysis with alignment at time zero due to computational limitation (see methods section).
OR: Odds Ratio

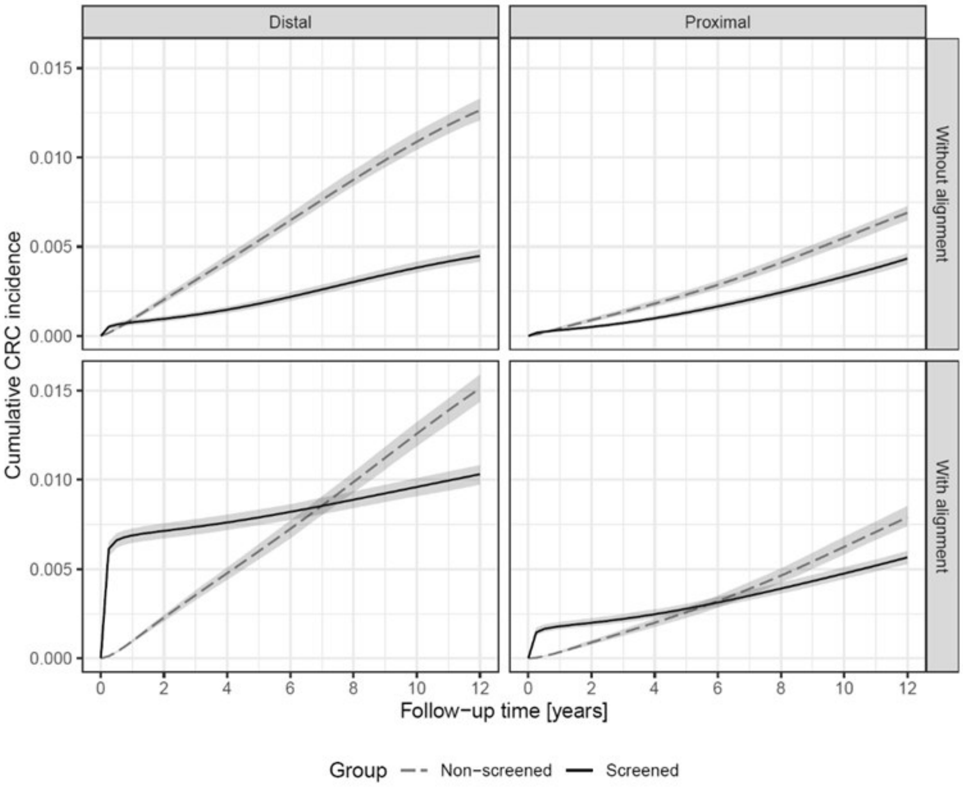


***Fig. 1**: Adjusted cumulative incidence functions for distal and proximal CRC from the cohort study design without alignment at time zero (top row) and the cohort study design with alignment at time zero (bottom row)*

# Supplement

## Supplement 1: Cohort and case-control study without alignment at time zero for different baseline years

As mentioned in the methods section, for the study designs without alignment at time zero, we selected individuals from the source population entering the cohort in 2009. In sensitivity analyses, we varied the baseline year, i.e. individuals entering the cohort in 2010 and 2011, respectively. The respective results are shown in Table S1 and Figure S1 for the cohort study and in Table S2 for the case-control study. For comparison, also the results of the base case analysis (baseline year 2009) are shown.

*Table S1: Results of cohort study designs without alignment at time zero for different baseline years.*

| ***Baseline year*** | | ***Control group*** | ***Screening group*** | ***Adjusted relative risk*** | ***(95% CI)*** |
|---|---|---|---|---|---|
| | Number at risk | 200,000 | 200,000 | | |
| | Number of CRC cases | | | | |
| 2009 | Distal, n | 2,472 | 829 | 0.35 | (0.32-0.39) |
| | Proximal, n | 1,290 | 823 | 0.63 | (0.57-0.69) |
| 2010 | Distal, n | 2,101 | 709 | 0.35 | (0.32-0.39) |
| | Proximal, n | 1,131 | 702 | 0.61 | (0.56-0.68) |
| 2011 | Distal, n | 2,048 | 642 | 0.33 | (0.30-0.37) |
| | Proximal, n | 1,056 | 640 | 0.54 | (0.49-0.59) |

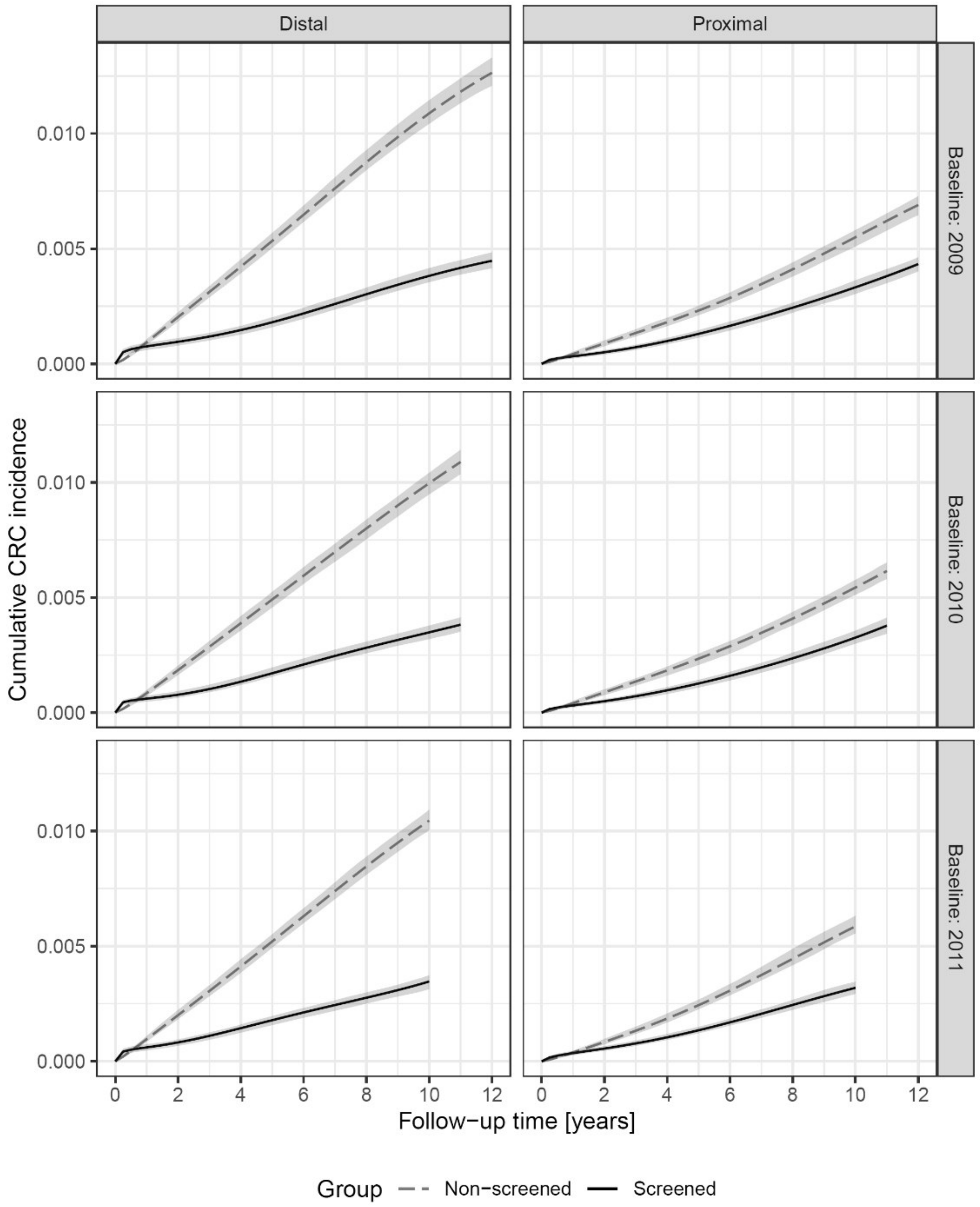


*Figure S1: Adjusted cumulative incidence functions for distal and proximal CRC from the cohort study design without alignment at time zero for different baseline years.*

*Table S2: Results of case-control designs without alignment at time zero for different baseline years.*

| ***Site*** | ***Case status*** | | ***Adjusted OR*** | |
|---|---|---|---|---|
| | ***Cases*** | ***Controls*** | | ***(95% CI)*** |
| Baseline year 2009 | | | | |
| Number of distal CRCs / controls | 446 | 2,230 | | |
| Thereof exposed | 36 | 653 | 0.23 | (0.16-0.33) |
| Number of proximal CRCs / controls | 302 | 1,510 | | |
| Thereof exposed | 54 | 434 | 0.54 | (0.39-0.75) |
| Baseline year 2010 | | | | |
| Number of distal CRCs / controls | 430 | 2,150 | | |
| Thereof exposed | 29 | 607 | 0.19 | (0.13-0.29) |
| Number of proximal CRCs / controls | 264 | 1,320 | | |
| Thereof exposed | 38 | 345 | 0.46 | (0.32-0.68) |
| Baseline year 2011 | | | | |
| Number of distal CRCs / controls | 408 | 2,040 | | |
| Thereof exposed | 24 | 500 | 0.20 | (0.13-0.31) |
| Number of proximal CRCs / controls | 254 | 1,270 | | |
| Thereof exposed | 31 | 298 | 0.47 | (0.31-0.71) |

## Supplement 2: Cohort and case-control study without alignment at time zero: considering both screening and diagnostic colonoscopy for the assignment to exposure groups

As mentioned in the methods section regarding the study designs without alignment at time zero, only screening colonoscopies were considered for the assignment to exposure groups in the base case analysis. In a sensitivity analysis, we considered both screening and diagnostic colonoscopies for the exposure assignment. The respective results are shown in Table S4 and Figure S2 for the cohort study design and in Table S5 for the case-control study.

*Table S4: Results of cohort study designs without alignment at time zero: sensitivity analyses considering both screening and diagnostic colonoscopy for the assignment to exposure groups. For comparison, also the results of the base case analysis are shown.*

| | *Control group* | *Screening group* | *Adjusted relative risk* | *(95% CI)* |
|---|---|---|---|---|
| Base case analysis | | | | |
| Number at risk | 200,000 | 200,000 | | |
| Number of CRC cases | | | | |
| Distal, n | 2,472 | 829 | 0.35 | (0.32-0.39) |
| Proximal, n | 1,290 | 823 | 0.63 | (0.57-0.69) |
| Sensitivity analysis | | | | |
| Number at risk | 200,000 | 200,000 | | |
| Number of CRC cases | | | | |
| Distal, n | 2,657 | 982 | 0.40 | (0.37-0.44) |
| Proximal, n | 1,348 | 893 | 0.66 | (0.60-0.72) |

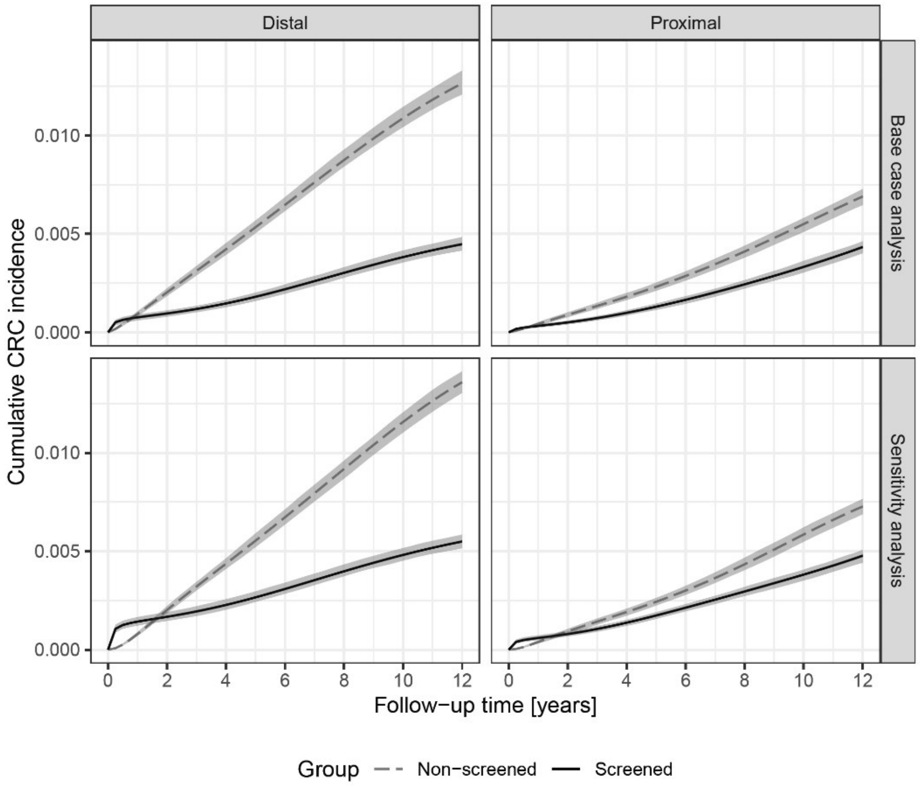


*Figure S2: Adjusted cumulative incidence functions for distal and proximal CRC from the cohort study design without alignment at time zero: sensitivity analysis considering both screening and diagnostic colonoscopy for the assignment to exposure groups. For comparison, also the cumulative incidence functions of the base case analysis are shown.*

*Table S5: Results of case-control designs without alignment at time zero: sensitivity analyses considering both screening and diagnostic colonoscopy for the assignment to exposure groups. For comparison, also the results of the base case analysis are shown.*

| ***Site*** | ***Case status*** | | ***Adjusted OR*** | |
|---|---|---|---|---|
| | ***Cases*** | ***Controls*** | | ***(95% CI)*** |
| Base case analysis | | | | |
| Number of distal CRCs / controls | 446 | 2,230 | | |
| Thereof exposed | 36 | 653 | 0.23 | (0.16-0.33) |
| Number of proximal CRCs / controls | 302 | 1,510 | | |
| Thereof exposed | 54 | 434 | 0.54 | (0.39-0.75) |
| Sensitivity analysis | | | | |
| Number of distal CRCs / controls | 446 | 2,230 | | |
| Thereof exposed | 60 | 1,020 | 0.20 | (0.15-0.26) |
| Number of proximal CRCs / controls | 302 | 1,510 | | |
| Thereof exposed | 82 | 685 | 0.44 | (0.33-0.58) |

## Supplement 3: Structural explanation of the bias inherent to study designs using pre-baseline information for the assignment to exposure groups

For simplicity, we divide time into three periods $t \in \{-1, 0, 1\}$ with $t = -1$ being the pre-baseline period, $t = 0$ the baseline and $t = 1$ the post-baseline or follow-up period. Let $E_t \in \{0, 1\}$ described a person's exposure to screening colonoscopy at time $t$. Let $P_t \in \{0, 1\}$ indicate the presence of colorectal precursors at time $t$ and $C_t \in \{0, 1\}$ the onset of preclinical CRC by time $t$. Let $Y_t \in \{0, 1\}$ indicate a diagnosis of colorectal cancer by time $t$. Finally, let $S = 1$ denote selection into the study cohort.

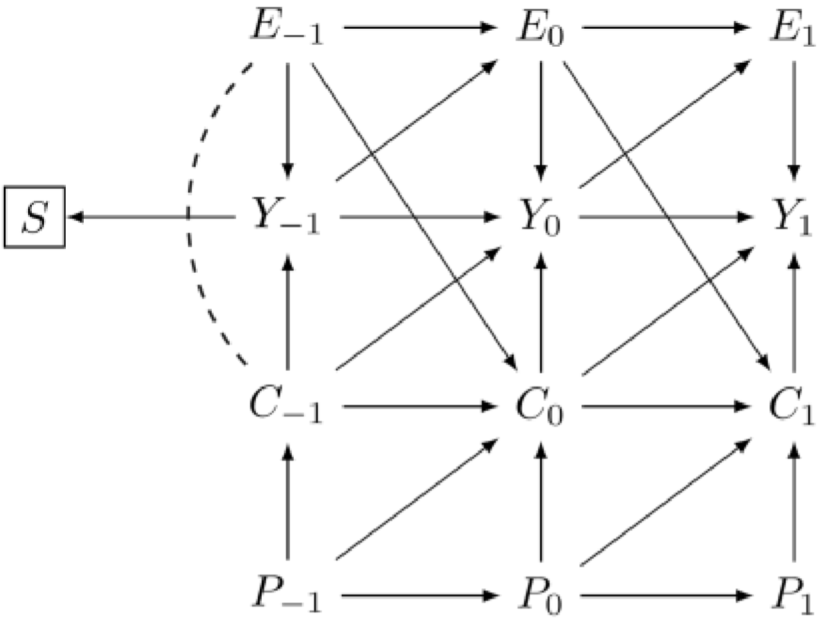


*Figure S3: DAG of bias resulting from violation of alignment at time zero in the form of exposure assessment based on pre-baseline information.*

As shown in Figure S3, at time point $t$ the causal mechanism that leads to a diagnosis of CRC is as follows: Precursors $P_t$ lead to the development of CRC $C_t$, which in turn progress to the outcome of interest, CRC diagnosis $Y_t$. At the same time, exposure to screening colonoscopy $E_t$ leads to CRC diagnosis $Y_t$ at the same time point, if the disease is present. Furthermore, exposure at time $t$ prevents disease onset at the later time $t + 1$ by removing precursor stages present at time $t$.

Importantly, the variable $Y_t$ is a collider variable on the path $P_t \rightarrow C_t \rightarrow Y_t \leftarrow E_t$. When cohort selection $S$ is based on this collider, a non-causal association is introduced between $E_t$ and $C_t$. If the cohort selection process excludes individuals with CRC diagnosis before baseline ($Y_{-1}$) while including individuals with past exposure $E_{-1}$ in the exposed group of the analysis dataset, the unexposed group appears to have a higher CRC incidence. Individuals who were screened in the past and had prevalent CRC received a diagnosis and were filtered out of the study cohort. Individuals who were screened in the past and did not have prevalent CRC are included in the exposed group. No such selection takes place in the unexposed group, where individuals must not have had any screening colonoscopy before baseline. Therefore, there is a non-causal association between exposure before baseline and prevalent, undiagnosed CRC before baseline. This non-causal association means that there are now open backdoor paths from exposure before baseline to the study outcome at later time points. The resulting bias, therefore, can be expressed as a form of collider stratification bias.

Importantly, the strength of the bias will depend on the prevalence of $C_{-1}$. If, conceptually, the prevalence of CRC before baseline were to approach zero, no such selection would take place. In the age group under study here, the prevalence of proximal CRC before baseline will be much lower than the prevalence of distal CRC before baseline, which means that this bias will impact the effect estimate for distal CRC more severely.

## Supplement 4: Illustration of the mechanism underlying the misallocation of screen-detected CRCs in case-control studies without alignment at time zero

Figure S4 illustrates the mechanism that underlies the misallocation of screen-detected CRCs in case-control studies without alignment at time zero, resulting in an overestimate of the effectiveness of screening colonoscopy. First, let us imagine a hypothetical RCT investigating the effectiveness of screening colonoscopy on CRC incidence. At baseline, screening-naïve persons are randomly assigned to either the screening or the control arm. Analysing this data as a case-control study without alignment at time zero would mean that for CRCs occurring in both arms, it is assessed whether they had a colonoscopy *before* CRC diagnosis. Given that screen-detected CRCs did not have a colonoscopy *before* CRC diagnosis, they are assigned (post-baseline, i.e. after randomization) to the control arm and are thereby classified as unexposed. This overestimates the effectiveness of screening given that CRCs accumulate in the control group (unexposed group). Given that screen-detected CRCs are more frequent in the distal colorectum, the resulting bias affects distal CRC more severely than proximal CRC.

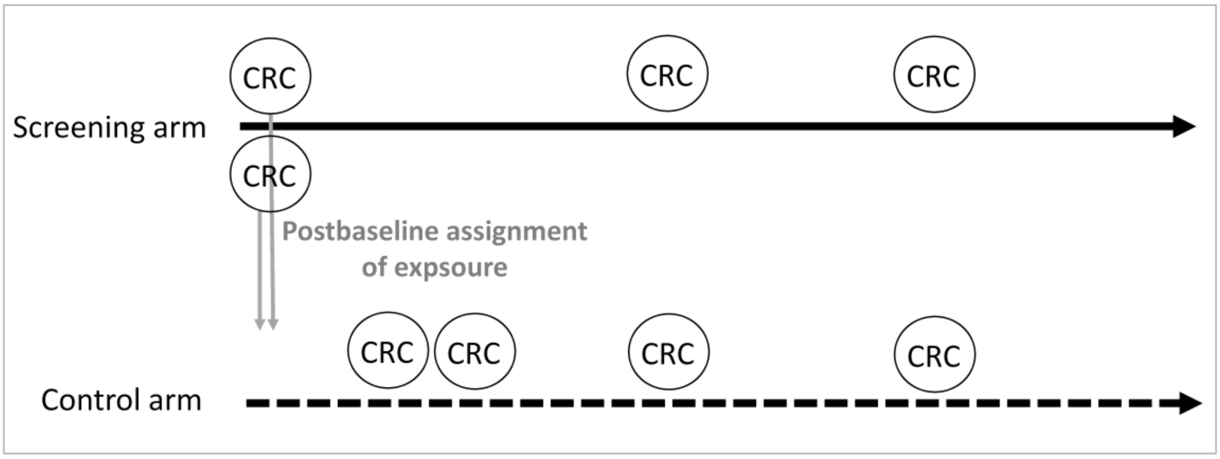


*Figure S4: Illustration of the mechanism of misallocation of screen-detected CRCs in case-control studies without alignment at time zero*

Of note, in published case-control studies investigating the effectiveness of screening colonoscopy based on primary data, selection bias in the control arm (higher prevalence of screening colonoscopy as compared to the general population) can—as an additional mechanism—also contribute to overestimating the effectiveness of screening colonoscopy, but it is not expected that this bias leads to a difference in the effectiveness by site.

In our case-control study without alignment at time zero, there was a second mechanism leading to overestimating the effectiveness of screening colonoscopy due a compromise we had to make because of the left truncation of our data. Specifically, we had to select CRC cases diagnosed in 2018-2020 from those entering the cohort in 2009 (see methods section) in order to be able to assess exposure in the 10 years prior to CRC diagnosis. CRCs diagnosed between 2009 and 2017 in the context of screening, which are more often in the distal than in the proximal colon, were not included in the final set of cases, i.e. distal CRCs exposed to screening colonoscopy were underrepresented in the final set of cases. We conducted additional analyses to disentangle the effect of both mechanism (data not shown), which did not change our conclusion, i.e. that the mechanism described in Figure S4 (also) leads to an artificial difference in the effectiveness of colonoscopy by site.

## Supplement 5: Bias due to post-baseline information for exposure assignment in a cohort study

In the cohort study by Nishihara et al. the assessment of eligibility criteria (e.g. no prior cancer except for nonmelanoma skin cancer, no prior endoscopy) as well as the start of follow-up was in 1988 (baseline). As part of a questionnaire administered every 2 years, participants were then asked whether they had undergone either sigmoidoscopy or colonoscopy and, if so, the reason for the investigation and whether there was a diagnosis of colorectal polyps. This means that the assignment to exposure groups used information after the assessment of eligibility and the start of follow-up, and it was updated every two years, i.e. post-baseline information was used to determine exposure. The outcome was the incidence of colorectal cancer, which was compared between participants without a lower endoscopy (control group), participants with a polypectomy, participants with a negative sigmoidoscopy and participants with a negative colonoscopy.

The mechanism described for the case-control study without alignment at time zero also applies to this design. In each two-year time interval CRCs detected in persons who had their first colonoscopy during this two-year time interval are—per definition—assigned to the unexposed group as they had no colonoscopy prior to CRC diagnosis. This overestimates the effectiveness of screening because CRCs are filtered to the unexposed group. As the majority of screen-detected CRCs are in the distal colon, this bias predominantly affects CRCs in the distal colon and thus leads to an artificial difference in the effectiveness of colonoscopy by site.

## Supplement 6: Post-colonoscopy CRC diagnoses

For the cohort analysis with alignment at time zero, we quantified the occurrence of post-colonoscopy CRC diagnoses occurring in the screening arm and assessed their site distribution. CRC diagnoses with a screening colonoscopy in the same calendar quarter or in the 180 days before CRC diagnosis were considered screen-detected and were not counted as post-colonoscopy CRC. The frequencies and percentages are given in the below Table:

| *Site* | *N* | *%* |
|---|---|---|
| Distal CRC | 541 | 39.3 |
| Proximal CRC | 633 | 46.0 |
| Both/unknown | 203 | 14.7 |
| Total | 1377 | |

# References

Aalen, O. O., Cook, R. J., and Røysland, K. (2015). Does Cox analysis of a randomized survival study yield a causal treatment effect? *Lifetime Data Analysis*, 21:579–593.

Abadie, A. and Imbens, G. W. (2008). On the failure of the bootstrap for matching estimators. *Econometrica*, 76(6):1537–1557.

Andersen, P. K., Geskus, R. B., De Witte, T., and Putter, H. (2012). Competing risks in epidemiology: possibilities and pitfalls. *International Journal of Epidemiology*, 41(3):861–870.

Arrospide, A., Rue, M., Van Ravesteyn, N. T., Comas, M., Larrañaga, N., Sarriugarte, G., and Mar, J. (2015). Evaluation of health benefits and harms of the breast cancer screening programme in the basque country using discrete event simulation. *BMC Cancer*, 15(1):1–11.

Assi, H. A., Khoury, K. E., Dbouk, H., Khalil, L. E., Mouhieddine, T. H., and El Saghir, N. S. (2013). Epidemiology and prognosis of breast cancer in young women. *Journal of Thoracic Disease*, 5(Suppl 1):S2.

Austin, P. C. (2009). Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in Medicine*, 28(25):3083–3107.

Austin, P. C. (2011). Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharmaceutical Statistics*, 10(2):150–161.

Austin, P. C. (2014). A comparison of 12 algorithms for matching on the propensity score. *Statistics in Medicine*, 33(6):1057–1069.

Austin, P. C. (2016). Variance estimation when using inverse probability of treatment weighting (IPTW) with survival analysis. *Statistics in Medicine*, 35(30):5642–5655.

Baum, M. (2013). Harms from breast cancer screening outweigh benefits if death caused by treatment is included. *The BMJ*, 346.

Baxter, N. N., Goldwasser, M. A., Paszat, L. F., Saskin, R., Urbach, D. R., and Rabeneck, L. (2009). Association of colonoscopy and death from colorectal cancer. *Annals of Internal Medicine*, 150(1):1–8.

Baxter, N. N., Warren, J. L., Barrett, M. J., Stukel, T. A., and Doria-Rose, V. P. (2012). Association between colonoscopy and colorectal cancer mortality in a US cohort according to site of cancer and colonoscopist specialty. *Journal of Clinical Oncology*, 30(21):2664.

Bickel, P. J. and Freedman, D. A. (1981). Some asymptotic theory for the bootstrap. *The Annals of Statistics*, 9(6):1196–1217.

Biller-Andorno, N. and Jüni, P. (2014). Abolishing mammography screening programs? A view from the Swiss Medical Board. *Obstetrical & Gynecological Survey*, 69(8):474–475.

Börnhorst, C., Reinders, T., Rathmann, W., Bongaerts, B., Haug, U., Didelez, V., and Kollhorst, B. (2021). Avoiding time-related biases: a feasibility study on antidiabetic drugs and pancreatic cancer applying the parametric g-formula to a large German healthcare database. *Clinical Epidemiology*, 13:1027.

Braitmaier, M. and Didelez, V. (2022). Emulierung von „target trials“ mit Real-world-Daten. *Prävention und Gesundheitsförderung*, pages 1–8.

Braitmaier, M., Kollhorst, B., Heinig, M., Langner, I., Czwikla, J., Heinze, F., Buschmann, L., Minnerup, H., García-Albéniz, X., Hense, H.-W., et al. (2022a). Effectiveness of mammography screening on breast cancer mortality-a study protocol for emulation of target trials using German health claims data. *Clinical Epidemiology*, pages 1293–1303.

Braitmaier, M., Schwarz, S., Didelez, V., and Haug, U. (2024). Misleading and avoidable: design-induced biases in observational studies evaluating cancer screening–the example of site-specific effectiveness of screening colonoscopy. *medRxiv*, pages 2024–04. DOI: 10.1101/2024.04.29.24306522.

Braitmaier, M., Schwarz, S., Kollhorst, B., Senore, C., Didelez, V., and Haug, U. (2022b). Screening colonoscopy similarly prevented distal and proximal colorectal cancer; a prospective study among 55-69-year-olds. *Journal of Clinical Epidemiology*, 149:118–126.

Brenner, H., Chang-Claude, J., Jansen, L., Knebel, P., Stock, C., and Hoffmeister, M. (2014a). Reduced risk of colorectal cancer up to 10 years after screening, surveillance, or diagnostic colonoscopy. *Gastroenterology*, 146(3):709–717.

Brenner, H., Chang-Claude, J., Seiler, C. M., Rickert, A., and Hoffmeister, M. (2011). Protection from colorectal cancer after colonoscopy: a population-based, case–control study. *Annals of Internal Medicine*, 154(1):22–30.

Brenner, H., Stock, C., and Hoffmeister, M. (2014b). Effect of screening sigmoidoscopy and screening colonoscopy on colorectal cancer incidence and mortality: systematic review and meta-analysis of randomised controlled trials and observational studies. *BMJ*, 348.

Bretthauer, M., Løberg, M., Wieszczy, P., Kalager, M., Emilsson, L., Garborg, K., Rupinski, M., Dekker, E., Spaander, M., Bugajski, M., et al. (2022). Effect of colonoscopy screening on risks of colorectal cancer and related death. *New England Journal of Medicine*, 387(17):1547–1556.

cancer.gov (2022). https://www.cancer.gov/types/breast/risk-fact-sheet#:~:text=in%20recent%20years%3F-,What%20is%20the%20average%20American%20woman[accessed: 17oct2022].

Caniglia, E. C., Robins, J. M., Cain, L. E., Sabin, C., Logan, R., Abgrall, S., Mugavero, M. J., Hernández-Díaz, S., Meyer, L., Seng, R., et al. (2019). Emulating a trial of joint dynamic strategies: an application to monitoring and treatment of HIV-positive individuals. *Statistics in Medicine*, 38(13):2428–2446.

Chiu, Y.-H., Chavarro, J. E., Dickerman, B. A., Manson, J. E., Mukamal, K. J., Rexrode, K. M., Rimm, E. B., and Hernán, M. A. (2021). Estimating the effect of nutritional interventions using observational data: the American Heart Association's 2020 dietary goals and mortality. *The American Journal of Clinical Nutrition*, 114(2):690–703.

Chiu, Y.-H., Huybrechts, K. F., Patorno, E., Yland, J. J., Cesta, C. E., Bateman, B. T., Seely, E. W., Hernán, M. A., and Hernández-Díaz, S. (2024). Metformin use in the first trimester of pregnancy and risk for nonlive birth and congenital malformations: emulating a target trial using real-world data. *Annals of Internal Medicine*. DOI: 10.7326/M23-2038.

Cole, S. R. and Hernán, M. A. (2004). Adjusted survival curves with inverse probability weights. *Computer Methods and Programs in Biomedicine*, 75(1):45–49.

D'Agostino, R. B., Lee, M. L., Belanger, A. J., Cupples, L. A., Anderson, K., and Kannel, W. B. (1990). Relation of pooled logistic regression to time dependant Cox regression analysis: the Framingham Heart Study. *Statistics in Medicine*, 9(12):1501–1515.

Dahabreh, I. J., Robertson, S. E., Tchetgen, E. J., Stuart, E. A., and Hernán, M. A. (2019). Generalizing causal inferences from individuals in randomized trials to all trial-eligible individuals. *Biometrics*, 75(2):685–694.

Danaei, G., Rodríguez, L. A. G., Cantero, O. F., Logan, R., and Hernán, M. A. (2013). Observational data for comparative effectiveness research: an emulation of randomised trials of statins and primary prevention of coronary heart disease. *Statistical Methods in Medical Research*, 22(1):70–96.

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap methods and their application*. Number 1. Cambridge University Press.

DiCiccio, T. J. and Efron, B. (1996). Bootstrap confidence intervals. *Statistical Science*, 11(3):189–228.

Dickerman, B. A., García-Albéniz, X., Logan, R. W., Denaxas, S., and Hernán, M. A. (2019). Avoidable flaws in observational analyses: an application to statins and cancer. *Nature Medicine*, 25(10):1601–1606.

Dickerman, B. A., García-Albéniz, X., Logan, R. W., Denaxas, S., and Hernán, M. A. (2023). Evaluating metformin strategies for cancer prevention: a target trial emulation using electronic health records. *Epidemiology*, 34(5):690–699.

Didelez, V. (2016). Commentary: Should the analysis of observational data always be preceded by specifying a target experimental trial? *International Journal of Epidemiology*, 45(6):2049–2051.

Dorn, H. F. (1953). Philosophy of inferences from retrospective studies. *American Journal of Public Health and the Nations Health*, 43(6_Pt_1):677–683.

Doubeni, C. A., Weinmann, S., Adams, K., Kamineni, A., Buist, D. S., Ash, A. S., Rutter, C. M., Doria-Rose, V. P., Corley, D. A., Greenlee, R. T., et al. (2013). Screening colonoscopy and risk for incident late-stage colorectal cancer diagnosis in average-risk adults: a nested case-control study. *Annals of Internal Medicine*, 158(5_Part_1):312–320.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26.

Efron, B. and Hastie, T. (2021). *Computer age statistical inference, student edition: algorithms, evidence, and data science*, volume 6. Cambridge University Press.

Elmunzer, B. J., Hayward, R. A., Schoenfeld, P. S., Saini, S. D., Deshpande, A., and Waljee, A. K. (2012). Effect of flexible sigmoidoscopy-based screening on incidence and mortality of colorectal cancer: a systematic review and meta-analysis of randomized controlled trials. *PLoS Medicine*, 9(12):e1001352.

Emilsson, L., García-Albéniz, X., Logan, R. W., Caniglia, E. C., Kalager, M., and Hernán, M. A. (2018). Examining bias in studies of statin treatment and survival in patients with cancer. *JAMA Oncology*, 4(1):63–70.

Faries, D., Zhang, X., Kadziola, Z., Siebert, U., Kuehne, F., Obenchain, R. L., and Haro, J. M. (2020). *Real world health care data analysis: causal methods and implementation using SAS*. SAS Institute.

Fine, J. P. and Gray, R. J. (1999). A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association*, 94(446):496–509.

Fox, M. P., MacLehose, R. F., and Lash, T. L. (2022). *Applying quantitative bias analysis to epidemiologic data*. Springer.

Franklin, J. M., Glynn, R. J., Martin, D., and Schneeweiss, S. (2019). Evaluating the use of nonrandomized real-world data analyses for regulatory decision making. *Clinical Pharmacology & Therapeutics*, 105(4):867–877.

Franklin, J. M., Patorno, E., Desai, R. J., Glynn, R. J., Martin, D., Quinto, K., Pawar, A., Bessette, L. G., Lee, H., Garry, E. M., et al. (2021). Emulating randomized clinical trials with nonrandomized real-world evidence studies: first results from the RCT DUPLICATE initiative. *Circulation*, 143(10):1002–1013.

Fu, E. L. (2023). Target trial emulation to improve causal inference from observational data: what, why, and how? *Journal of the American Society of Nephrology*, 34(8):1305–1314.

g-ba.de (2024). https://www.g-ba.de/presse/pressemitteilungen-meldungen/1133/[accessed: 20jun2024].

García-Albéniz, X., Hernán, M. A., Logan, R. W., Price, M., Armstrong, K., and Hsu, J. (2020). Continuation of annual screening mammography and breast cancer mortality in women older than 70 years. *Annals of Internal Medicine*, 172(6):381–389.

García-Albéniz, X., Hsu, J., Bretthauer, M., and Hernán, M. A. (2019). Estimating the effect of preventive services with databases of administrative claims: reasons to be concerned. *American Journal of Epidemiology*, 188(10):1764–1767.

García-Albéniz, X., Hsu, J., Bretthauer, M., and Hernán, M. A. (2017a). Effectiveness of screening colonoscopy to prevent colorectal cancer among Medicare beneficiaries aged 70 to 79 years: a prospective observational study. *Annals of Internal Medicine*, 166(1):18–26.

García-Albéniz, X., Hsu, J., and Hernán, M. A. (2017b). The value of explicitly emulating a target trial when using real world evidence: an application to colorectal cancer screening. *European Journal of Epidemiology*, 32(6):495–500.

Goetghebeur, E., le Cessie, S., De Stavola, B., Moodie, E. E., Waernbaum, I., and the topic group Causal Inference (TG7) of the STRATOS initiative (2020). Formulating causal questions and principled statistical answers. *Statistics in Medicine*, 39(30):4922–4948.

Greenland, S. (2000). An introduction to instrumental variables for epidemiologists. *International Journal of Epidemiology*, 29(4):722–729.

Greenland, S. (2003). Quantifying biases in causal models: classical confounding vs collider-stratification bias. *Epidemiology*, pages 300–306.

Grodstein, F., Manson, J. E., and Stampfer, M. J. (2006). Hormone therapy and coronary heart disease: the role of time since menopause and age at hormone initiation. *Journal of Women's Health*, 15(1):35–44.

Guarneri, V. and Conte, P. F. (2004). The curability of breast cancer and the treatment of advanced disease. *European journal of nuclear medicine and molecular imaging*, 31(1):S149–S161.

Guo, F., Chen, C., Holleczek, B., Schöttker, B., Hoffmeister, M., and Brenner, H. (2021). Strong reduction of colorectal cancer incidence and mortality after screening colonoscopy: prospective cohort study from Germany. *American College of Gastroenterology| ACG*, 116(5):967–975.

Hannan, L. M., Jacobs, E. J., and Thun, M. J. (2009). The association between cigarette smoking and risk of colorectal cancer in a large prospective cohort from the United States. *Cancer Epidemiology, Biomarkers & Prevention*, 18(12):3362–3367.

Hansford, H. J., Cashin, A. G., Jones, M. D., Swanson, S. A., Islam, N., Dahabreh, I. J., Dickerman, B. A., Egger, M., Garcia-Albeniz, X., Golub, R. M., et al. (2023a). Development of the TrAnsparent ReportinG of observational studies Emulating a Target trial (TARGET) guideline. *BMJ Open*, 13(9):e074626.

Hansford, H. J., Cashin, A. G., Jones, M. D., Swanson, S. A., Islam, N., Douglas, S. R. G., Rizzo, R. R. N., Devonshire, J. J., Williams, S. A., Dahabreh, I. J., Dickerman, B. A., Egger, M., Garcia-Albeniz, X., Golub, R. M., Lodi, S., Moreno-Betancur, M., Pearson, S.-A., Schneeweiss, S., Sterne, J. A. C., Sharp, M. K., Stuart, E. A., Hernán, M. A., Lee, H., and McAuley, J. H. (2023b). Reporting of observational studies explicitly aiming to emulate randomized trials: a systematic review. *JAMA Network Open*, 6(9):e2336023.

Haug, U. and Schink, T. (2021). German pharmacoepidemiological research database (GePaRD). In Sturkenboom, M. and Schink, T., editors, *Databases for Pharmacoepidemiological Research*, pages 119–124. Springer, Cham.

Hernán, M. A. (2004). A definition of causal effect for epidemiological research. *Journal of Epidemiology & Community Health*, 58(4):265–271.

Hernán, M. A. (2010). The hazards of hazard ratios. *Epidemiology*, 21(1):13.

Hernán, M. A. (2018). How to estimate the effect of treatment duration on survival outcomes using observational data. *BMJ*, 360.

Hernán, M. A., Alonso, A., Logan, R., Grodstein, F., Michels, K. B., Stampfer, M. J., Willett, W. C., Manson, J. E., and Robins, J. M. (2008). Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease. *Epidemiology (Cambridge, Mass.)*, 19(6):766.

Hernán, M. A. and Hernández-Díaz, S. (2012). Beyond the intention-to-treat in comparative effectiveness research. *Clinical Trials*, 9(1):48–55.

Hernán, M. A. and Monge, S. (2023). Selection bias due to conditioning on a collider. *BMJ*, 381.

Hernán, M. A. and Robins, J. M. (2016). Using big data to emulate a target trial when a randomized trial is not available. *American Journal of Epidemiology*, 183(8):758–764.

Hernán, M. A. and Robins, J. (2020). *Causal inference: What if.* Chapman & Hall/CRC.

Hernán, M. A., Sauer, B. C., Hernández-Díaz, S., and Platt, R. (2016). Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses. *Journal of Clinical Epidemiology*, 79:70–75.

Heyard, R., Held, L., Schneeweiss, S., and Wang, S. V. (2024). Design differences and variation in results between randomised trials and non-randomised emulations: meta-analysis of RCT-DUPLICATE data. *BMJ Medicine*, 3(1).

Hoffman, K. L., Schenck, E. J., Satlin, M. J., Whalen, W., Pan, D., Williams, N., and Díaz, I. (2022). Comparison of a target trial emulation framework vs Cox regression to estimate the association of corticosteroids with COVID-19 mortality. *JAMA Network Open*, 5(10):e2234425–e2234425.

Howe, C. J., Cole, S. R., Lau, B., Napravnik, S., and Eron Jr, J. J. (2016). Selection bias due to loss to follow up in cohort studies. *Epidemiology (Cambridge, Mass.)*, 27(1):91.

ICH E9 (R1) (2020). ICH Harmonised Guideline Addendum on Estimands and Sensitivity Analysis in Clinical Trials to the Guideline on Statistical Principles for Clinical Trials E9(R1). *International Conference on Harmonisation*.

Imai, K. and Ratkovic, M. (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, pages 243–263.

Jansen, L., Holleczek, B., Kraywinkel, K., Weberpals, J., Schröder, C. C., Eberle, A., Emrich, K., Kajüter, H., Katalinic, A., Kieschke, J., et al. (2020). Divergent patterns and trends in breast cancer incidence, mortality and survival among older women in Germany and the United States. *Cancers*, 12(9):2419.

Joffe, M. M. (2001). Administrative and artificial censoring in censored regression models. *Statistics in Medicine*, 20(15):2287–2304.

Kahi, C. J., Pohl, H., Myers, L. J., Mobarek, D., Robertson, D. J., and Imperiale, T. F. (2018). Colonoscopy and colorectal cancer mortality in the Veterans Affairs health care system: a case–control study. *Annals of Internal Medicine*, 168(7):481–488.

Kaminski, M. F., Thomas-Gibson, S., Bugajski, M., Bretthauer, M., Rees, C. J., Dekker, E., Hoff, G., Jover, R., Suchanek, S., Ferlitsch, M., et al. (2017). Performance measures for lower gastrointestinal endoscopy: a European Society of Gastrointestinal Endoscopy (ESGE) quality improvement initiative. *United European Gastroenterology Journal*, 5(3):309–334.

Laanani, M., Coste, J., Blotière, P.-O., Carbonnel, F., and Weill, A. (2019). Patient, procedure, and endoscopist risk factors for perforation, bleeding, and splenic injury after colonoscopies. *Clinical Gastroenterology and Hepatology*, 17(4):719–727.

Labrecque, J. A. and Swanson, S. A. (2017). Target trial emulation: teaching epidemiology and beyond. *European Journal of Epidemiology*, 32:473–475.

Latouche, A., Allignol, A., Beyersmann, J., Labopin, M., and Fine, J. P. (2013). A competing risks analysis should report results on all cause-specific hazards and cumulative incidence functions. *Journal of Clinical Epidemiology*, 66(6):648–653.

Lau, B., Cole, S. R., and Gange, S. J. (2009). Competing risk regression models for epidemiologic data. *American Journal of Epidemiology*, 170(2):244–256.

Lauby-Secretan, B., Scoccianti, C., Loomis, D., Benbrahim-Tallaa, L., Bouvard, V., Bianchini, F., and Straif, K. (2015). Breast-cancer screening—viewpoint of the IARC Working Group. *New England Journal of Medicine*, 372(24):2353–2358.

Lee, B. K., Lessler, J., and Stuart, E. A. (2010). Improving propensity score weighting using machine learning. *Statistics in Medicine*, 29(3):337–346.

Lesko, C. R. and Lau, B. (2017). Bias due to confounders for the exposure-competing risk relationship. *Epidemiology*, 28(1):20.

Li, H., Wang, C., Chen, W.-C., Lu, N., Song, C., Tiwari, R., Xu, Y., and Yue, L. Q. (2022). Estimands in observational studies: some considerations beyond ICH E9 (R1). *Pharmaceutical Statistics*, 21(5):835–844.

Lipsitch, M., Tchetgen, E. T., and Cohen, T. (2010). Negative controls: a tool for detecting confounding and bias in observational studies. *Epidemiology*, 21(3):383.

Løberg, M., Lousdal, M. L., Bretthauer, M., and Kalager, M. (2015). Benefits and harms of mammography screening. *Breast Cancer Research*, 17(1):1–12.

Lowenfels, A. B. and Maisonneuve, P. (2005). Risk factors for pancreatic cancer. *Journal of Cellular Biochemistry*, 95(4):649–656.

Maisonneuve, P. and Lowenfels, A. B. (2015). Risk factors for pancreatic cancer: a summary review of meta-analytical studies. *International Journal of Epidemiology*, 44(1):186–198.

Manson, J. E., Hsia, J., Johnson, K. C., Rossouw, J. E., Assaf, A. R., Lasser, N. L., Trevisan, M., Black, H. R., Heckbert, S. R., Detrano, R., et al. (2003). Estrogen plus progestin and the risk of coronary heart disease. *New England Journal of Medicine*, 349(6):523–534.

Marmot, M. G., Altman, D., Cameron, D., Dewar, J., Thompson, S., and Wilcox, M. (2013). The benefits and harms of breast cancer screening: an independent review. *British Journal of Cancer*, 108(11):2205–2240.

Martinussen, T. (2022). Causality and the Cox regression model. *Annual Review of Statistics and Its Application*, 9:249–259.

McNabb, S., Harrison, T. A., Albanes, D., Berndt, S. I., Brenner, H., Caan, B. J., Campbell, P. T., Cao, Y., Chang-Claude, J., Chan, A., et al. (2020). Meta-analysis of 16 studies of the association of alcohol with colorectal cancer. *International Journal of Cancer*, 146(3):861–873.

Mulder, S. A., van Soest, E. M., Dieleman, J. P., van Rossum, L. G., Rob, J., van Leerdam, M. E., and Kuipers, E. J. (2010). Exposure to colorectal examinations before a colorectal cancer diagnosis: a case–control study. *European Journal of Gastroenterology & Hepatology*, 22(4):437–443.

Murray, E. J., Caniglia, E. C., and Petito, L. C. (2021). Causal survival analysis: a guide to estimating intention-to-treat and per-protocol effects from randomized clinical trials with non-adherence. *Research Methods in Medicine & Health Sciences*, 2(1):39–49.

Narod, S. A., Giannakeas, V., and Sopik, V. (2018). Time to death in breast cancer patients as an indicator of treatment response. *Breast Cancer Research and Treatment*, 172(3):659–669.

National Cancer Institute (2024). SEER Cancer Stat Facts: Female Breast Cancer. https://seer.cancer.gov/statfacts/html/breast.html. [Online; accessed 01-February-2024].

Nelson, H. D., Fu, R., Cantor, A., Pappas, M., Daeges, M., and Humphrey, L. (2016). Effectiveness of breast cancer screening: Systematic review and meta-analysis to update the 2009 US Preventive Services Task Force recommendation. *Annals of Internal Medicine*, 164(4):244–255.

Nishihara, R., Wu, K., Lochhead, P., Morikawa, T., Liao, X., Qian, Z. R., Inamura, K., Kim, S. A., Kuchiba, A., Yamauchi, M., et al. (2013). Long-term colorectal-cancer incidence and mortality after lower endoscopy. *New England Journal of Medicine*, 369(12):1095–1105.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, 82(4):669–688.

Peters, J., Janzing, D., and Schölkopf, B. (2017). *Elements of causal inference: foundations and learning algorithms*. The MIT Press.

Petito, L. C., García-Albéniz, X., Logan, R. W., Howlader, N., Mariotto, A. B., Dahabreh, I. J., and Hernán, M. A. (2020). Estimates of overall survival in patients with cancer receiving different treatment regimens: emulating hypothetical target trials in the surveillance, epidemiology, and end results (SEER)–Medicare linked database. *JAMA Network Open*, 3(3):e200452–e200452.

Pigeot, I. and Ahrens, W. (2008). Establishment of a pharmacoepidemiological database in Germany: methodological potential, scientific value and practical limitations. *Pharmacoepidemiology and Drug Safety*, 17(3):215–223.

Pirracchio, R. and Carone, M. (2018). The balance super learner: a robust adaptation of the super learner to improve estimation of the average treatment effect in the treated based on propensity score matching. *Statistical Methods in Medical Research*, 27(8):2504–2518.

Putter, H., Fiocco, M., and Geskus, R. B. (2007). Tutorial in biostatistics: competing risks and multi-state models. *Statistics in Medicine*, 26(11):2389–2430.

Putter, H., Schumacher, M., and van Houwelingen, H. C. (2020). On the relation between the cause-specific hazard and the subdistribution rate for competing risks data: the Fine-Gray model revisited. *Biometrical Journal*, 62(3):790–807.

Rasouli, B., Chubak, J., Floyd, J. S., Psaty, B. M., Nguyen, M., Walker, R. L., Wiggins, K. L., Logan, R. W., and Danaei, G. (2023). Combining high quality data with rigorous methods: emulation of a target trial using electronic health records and a nested case-control design. *BMJ*, 383.

Ray, W. A. (2003). Evaluating medication effects outside of clinical trials: new-user designs. *American Journal of Epidemiology*, 158(9):915–920.

Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period–application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9-12):1393–1512.

Robins, J. M., Hernan, M. A., and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560.

Rojas-Saunero, L. P., Young, J. G., Didelez, V., Ikram, M. A., and Swanson, S. A. (2023). Considering questions before methods in dementia research with competing events and causal goals. *American Journal of Epidemiology*, 192(8):1415–1423.

Rosenbaum, P. R. (1989). Optimal matching for observational studies. *Journal of the American Statistical Association*, 84(408):1024–1032.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Rubin, D. B. (2005). Causal inference using potential outcomes: design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331.

Rühl, J. and Friedrich, S. (2023). Resampling-based confidence intervals and bands for the average treatment effect in observational studies with competing risks. *arXiv preprint arXiv:2306.03453*.

Schisterman, E. F., Cole, S. R., and Platt, R. W. (2009). Overadjustment bias and unnecessary adjustment in epidemiologic studies. *Epidemiology*, 20(4):488.

Schmid, M. and Berger, M. (2021). Competing risks analysis for discrete time-to-event data. *Wiley Interdisciplinary Reviews: Computational Statistics*, 13(5):e1529.

Schmuker, C. and Zok, K. (2019). Informierte Teilnahme an Früherkennungsuntersuchungen: Ergebnisse einer Befragung unter GKV-Versicherten. *Günster C, Klauber J, Robra BP, Schmacke N, Schmucker C. Versorgungsreport Früherkennunged. Medizinisch Wissenschaftlichen Verlagsgesellschaft Berlin*, pages 31–78.

Schneeweiss, S. (2006). Sensitivity analysis and external adjustment for unmeasured confounders in epidemiologic database studies of therapeutics. *Pharmacoepidemiology and Drug Safety*, 15(5):291–303.

Schneeweiss, S. and Avorn, J. (2005). A review of uses of health care utilization databases for epidemiologic research on therapeutics. *Journal of Clinical Epidemiology*, 58(4):323–337.

Schwarz, S., Braitmaier, M., Pox, C., Kollhorst, B., Didelez, V., and Haug, U. (2024). 13-year colorectal cancer risk after low-quality, high-quality and no screening colonoscopy: a cohort study. *under review*.

Schwarz, S., Hornschuch, M., Pox, C., and Haug, U. (2023). Polyp detection rate and cumulative incidence of post-colonoscopy colorectal cancer in Germany. *International Journal of Cancer*, 152(8):1547–1555.

Shrank, W. H., Patrick, A. R., and Alan Brookhart, M. (2011). Healthy user and related biases in observational studies of preventive interventions: a primer for physicians. *Journal of General Internal Medicine*, 26:546–550.

Shrier, I. and Suissa, S. (2022). The quintessence of causal DAGs for immortal time bias: time-dependent models. *International Journal of Epidemiology*, 51(3):1028–1029.

Suissa, S. (2008). Immortal time bias in pharmacoepidemiology. *American Journal of Epidemiology*, 167(4):492–499.

Suissa, S. and Azoulay, L. (2012). Metformin and the risk of cancer: time-related biases in observational studies. *Diabetes Care*, 35(12):2665–2673.

Suresh, K., Severn, C., and Ghosh, D. (2022). Survival prediction models: an introduction to discrete-time modeling. *BMC Medical Research Methodology*, 22(1):207.

Tsiatis, A. A. (2006). *Semiparametric theory and missing data*. Springer.

VanderWeele, T. J. and Hernán, M. A. (2013). Causal inference under multiple versions of treatment. *Journal of Causal Inference*, 1(1):1–20.

Wang, A., Nianogo, R. A., and Arah, O. A. (2017). G-computation of average treatment effects on the treated and the untreated. *BMC Medical Research Methodology*, 17:1–5.

Wang, S. V., Lin, K. J., and Schneeweiss, S. (2024). Emulation of randomized trials of direct oral anticoagulants with claims data and implications for new Factor XI inhibitors. *Pharmacoepidemiology and Drug Safety*, 33(5):e5813.

Wang, S. V., Schneeweiss, S., Franklin, J. M., Desai, R. J., Feldman, W., Garry, E. M., Glynn, R. J., Lin, K. J., Paik, J., Patorno, E., et al. (2023). Emulation of randomized clinical trials with nonrandomized database analyses: results of 32 clinical trials. *JAMA*, 329(16):1376–1385.

Wang, Y.-T., Gou, Y.-W., Jin, W.-W., Xiao, M., and Fang, H.-Y. (2016). Association between alcohol intake and the risk of pancreatic cancer: a dose–response meta-analysis of cohort studies. *BMC Cancer*, 16:1–11.

Weedon-Fekjær, H., Vatten, L. J., Aalen, O. O., Lindqvist, B., and Tretli, S. (2005). Estimating mean sojourn time and screening test sensitivity in breast cancer mammography screening: new results. *Journal of Medical Screening*, 12(4):172–178.

Weiss, N. S. and Rossing, M. A. (1996). Healthy screenee bias in epidemiologic studies of cancer incidence. *Epidemiology*, pages 319–322.

Witte, J. and Didelez, V. (2019). Covariate selection strategies for causal inference: classification and comparison. *Biometrical Journal*, 61(5):1270–1289.

Wyss, R., Ellis, A. R., Brookhart, M. A., Girman, C. J., Jonsson Funk, M., LoCasale, R., and Stürmer, T. (2014). The role of prediction modeling in propensity score estimation: an evaluation of logistic regression, bCART, and the covariate-balancing propensity score. *American Journal of Epidemiology*, 180(6):645–655.

Young, J. G., Stensrud, M. J., Tchetgen Tchetgen, E. J., and Hernán, M. A. (2020). A causal framework for classical statistical estimands in failure-time settings with competing events. *Statistics in Medicine*, 39(8):1199–1236.

Yu, G.-H., Li, S.-F., Wei, R., Jiang, Z., et al. (2022). Diabetes and colorectal cancer risk: clinical and therapeutic implications. *Journal of Diabetes Research*, 2022.

Zahl, P.-H., Mæhlen, J., and Welch, H. G. (2008). The natural history of invasive breast cancers detected by screening mammography. *Archives of Internal Medicine*, 168(21):2311–2316.

Zhang, Z., Kim, H. J., Lonjon, G., Zhu, Y., et al. (2019). Balance diagnostics after propensity score matching. *Annals of Translational Medicine*, 7(1).

Zhao, S. S., Lyu, H., and Yoshida, K. (2021). Versatility of the clone-censor-weight approach: response to “trial emulation in the presence of immortal-time bias”. *International Journal of Epidemiology*, 50(2):694–695.

# List of Figures

# List of Tables

# List of abbreviations

| | |
|---|---|
| **ATE** | average treatment effect |
| **ATT** | average treatment effect on the treated |
| **ATU** | average treatment effect on the untreated |
| **ARR** | absolute risk reduction |
| **BMI** | body mass index |
| **CIF** | cumulative incidence function |
| **CRC** | colorectal cancer |
| **CRR** | causal relative risk |
| **DAG** | directed acyclic graph |
| **EHR** | electronic health records |
| **GePaRD** | German Pharmacoepidemiological Research Database |
| **HR** | hazard ratio |
| **iid** | independant and identically distributed |
| **IPCW** | inverse probability of censoring weighting |
| **IPTW** | inverse probability of treatment weighting |
| **IPW** | inverse probability weighting |
| **ITT** | intention-to-treat |
| **MSM** | marginal structural model |
| **PDR** | polyp detection rate |
| **PP** | per-protocol |
| **PS** | propensity score |
| **RCT** | randomized controlled trial |
| **RR** | relative risk |
| **RWD** | real-world data |
| **TTE** | target trial emulation |

# List of mathematical notation

**Indices:**

| | |
|---|---|
| $h = 1, ..., u$ | Ordered event times |
| $i = 1, ..., n$ | Individuals included in the study cohort |
| $j = 1, ..., m$ | Person-trials ($m \geq n$) |
| $k_r = 1, ..., K_r$ | Follow-up time of the $r$-th trial, with $K_r = T - t_r + 1$ |
| $r = 1, ..., R$ | Indicator for the $r$-th emulated trial |
| $t = 1, ..., T$ | Index for time point |
| $t_r$ | Start time of the $r$-th emulated trial on calendar time scale |

**Random variables:**

| | |
|---|---|
| $A$ | Exposure |
| $A_q \in \{0, 1\}$ | Dummy encoding of exposure, when more than two exposure levels are present |
| $B$ | Number of bootstrap samples |
| $C \in \{0, 1\}$ | Presence of colorectal cancer |
| $D \in \{0, 1\}$ | Competing event indicator |
| $E$ | Type of outcome event in competing events setting (e.g. $E = 1$ for CRC incidence and $E = 2$ for death) |
| $N \in \{0, 1\}$ | Negative control outcome |
| $P$ | Used to denote presence of precursors in the screening colonoscopy study ($P \in \{0, 1\}$) |
| $Q$ | Exposure strategy |
| $S \in \{0, 1\}$ | Selection into the cohort |
| $T > 0$ | Survival time |
| $U$ | Unobserved confounders |
| $X$ | Measured covariates |
| $Y \in \{0, 1\}$ | Outcome |

**Other symbols:**

| | |
|---|---|
| $\perp\!\!\!\perp$ | Independence |
| $\not\perp\!\!\!\perp$ | Dependence |

# Note regarding shortened published version (§12 (2))

The manuscript by Schwarz et al. [2024] was under review at the time of thesis submission. In accordance with § 12 (2) of the ”Promotionsordnung (Dr.-Ing.) and (Dr. rer. nat.) der Universität Bremen für den Fachbereich 3 (Mathematik, Informatik)”, the manuscript was removed from the published version of this thesis to avoid conflicts with future copyright agreements and only the draft abstract was printed in section 7.5.

# Appendices

# APPENDIX A

# Bootstrapping in emulated target trials

Target trial emulation often includes the data from one individual multiple times, either because of cloning into exposure strategies that are congruent at baseline or because of repeated study entry in sequential emulated trials. Naive parametric variance estimators are then not applicable since they do not adjust for dependencies in the analysis dataset. Furthermore, statistical methods in emulated target trials with survival outcomes are often complex, e.g. pooled logistic regression models [D'Agostino et al., 1990] are frequently used to model flexible cumulative incidence functions [Hernán and Robins, 2020], which in turn are used to estimate contrasts such as ATEs or marginal relative risks, making it difficult to obtain analytical solutions for variance estimation. Instead, bootstrapping is commonly applied to obtain valid variance estimates and confidence intervals [Hernán and Robins, 2020]. Alternative approaches, such as robust sandwich estimators, exist for some but not all statistical methods [Austin, 2016]. Faster bootstrap algorithms, such as the wild bootstrap, have been proposed for time to event settings with competing events [Rühl and Friedrich, 2023], but have not been extended to target trial emulation settings with cloned data. Therefore, the classic bootstrap approach is the only currently available method of estimating robust confidence intervals, when using pooled logistic regression in a target trial emulation with repeated inclusion of the same individual.

Bootstrapping is a general, computer-intense method of obtaining valid variance estimates for a large variety of estimators [Efron, 1979]. It is an assumption lean method, particularly regarding parametric assumptions of the distribution underlying the sampled data. However, some assumptions must be made. For instance, the data is assumed to be independent and identically distributed, i.e. $Z_i \overset{iid}{\sim} F$. Furthermore, the observed distribution function $\hat{F}(z_i)$ must be an unbiased estimator for the true underlying distribution function $F(Z_i)$. Next, the parameter of interest $\theta$ must be a smooth function of $F$. See, for instance, Davison and Hinkley [1997]; Efron [1979]; Efron and Hastie [2021] for an in-depth introduction to bootstrapping and Bickel and Freedman [1981] for some asymptotic theory and examples in which bootstrapping fails.

Conceptually, many data samples could be obtained from the underlying population, as to assess the distribution of the estimator, which would be informative with regards to the variance of the estimator. For example, a series of one hundred studies - always sampling from the same underlying population - would result in a distribution of one hundred estimates, which is informative regarding the variance of the estimator. Since repeating an experiment a hundred times is not feasible, the bootstrap instead uses the sample of available data. To illustrate this, assume a sample of $n$ observations $z_1, ..., z_n$ stemming from a random variable $Z \overset{iid}{\sim} N(0, 1)$. A parameter of interest, $\theta$, is defined by a function of $Z$ as $\theta = g(Z)$. The estimate of $\theta$ is then $\hat{\theta} = g(z)$. In some cases, an analytic solution might not be available to estimate a confidence interval for the estimate $\hat{\theta}$, in which case bootstrapping is an alternative to obtain a robust confidence interval. A bootstrap sample is obtained by randomly sampling, with replacement, from $\hat{F}(z)$ exactly $n$ times, resulting in the bootstrap sample $z^* = (z_1^*, ..., z_n^*)$. A bootstrap estimate of the target parameter $\theta$ is then $\tilde{\theta} = g(z^*)$. This process is repeated $B$ times to obtain a distribution of bootstrap estimates $\tilde{\theta}_1, ..., \tilde{\theta}_B$. The distribution of bootstrap estimates is used to derive standard errors or confidence intervals, e.g. via the percentile bootstrap taking the 2.5 and 97.5 percentiles as lower and upper confidence limits [Efron, 1979; DiCiccio and Efron, 1996].

A central assumption of bootstrapping is that samples are independant and identically distributed (iid). As described above, in emulated target trials usually the same individual is included more than once. Similarly, in PS matched analyses, the same individual is often included as control in multiple matching sets to increase statistical efficiency. Furthermore, due to the matched nature of the analysis data, observations within matching sets are not independent. The observations in the analysis dataset of such studies, then, are not iid. For matching estimators, Abadie and Imbens noted that the basic bootstrap, i.e. when sampling from the matched data, does not yield valid confidence intervals [Abadie and Imbens, 2008]. Instead, bootstrap samples need to be drawn from the underlying study population, i.e. before matching is done, and the process of matching and estimation be repeated for the so-obtained bootstrap samples. Similarly, when emulating target trials, bootstrap samples need to be drawn from the underlying study population and the entire process of trial emulation, estimation of weights, weighted outcome regression and estimation of resulting contrasts of interest must be repeated for each bootstrap sample to obtain valid confidence intervals [Murray et al., 2021; Hernán and Robins, 2020].

The derivation of percentile-based bootstrap confidence intervals in the context of the evaluation of screening colonoscopy is briefly described in Chapter 4.