Emulating a target trial in case-control designs: an application to statins and colorectal cancer

Barbra A. Dickerman, PhD*,[a] Xabier García-Albéniz, MD, PhD,[a,b] Roger W. Logan, PhD,[a] Spiros Denaxas, PhD,[c,d,e] Miguel A. Hernán, MD, DrPH[a,f,g]

Affiliations:
[a] Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA
[b] RTI Health Solutions. Barcelona, Spain.
[c] Institute of Health Informatics Research, University College London, London, UK
[d] Health Data Research UK (HDR UK) London, University College London, London, UK
[e] The Alan Turing Institute, London, UK
[f] Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA
[g] Harvard-MIT Division of Health Sciences and Technology, Boston, MA, USA

Correspondence: Barbra A. Dickerman, Harvard T.H. Chan School of Public Health, 677 Huntington Avenue, Department of Epidemiology, Boston, MA, 02115. Email: bad788@mail.harvard.edu.

Funding: This work was supported by the National Institutes of Health grants K99 CA248335 (B.A.D.) and P01 CA134294.

Conflict of interest: none declared

Word counts: Abstract, 248; Text, 3,072.

Data statement: This study is based in part on data from the Clinical Practice Research Datalink obtained under license from the UK Medicines and Healthcare Products Regulatory Agency. The data are provided by patients and collected by the UK National Health Service as part of their care and support. This study is also based in part on data from the Hospital Episode Statistics and Office of National Statistics, re-used with permission of The Health & Social Care Information Centre. The study was approved by the Medicines and Healthcare Products Regulatory Agency Independent Scientific Advisory Committee (protocol 16_221), under Section 251 (National Health Service Social Care Act 2006). The interpretation and conclusions contained in this study are those of the authors alone.

ABSTRACT

Background: Previous case-control studies have reported a strong association between statin use and lower cancer risk. It is unclear whether this association reflects a benefit of statins or is the result of design decisions that cannot be mapped to a (hypothetical) target trial (that would answer the question of interest).

Methods: We outlined the protocol of a target trial to estimate the effect of statins on colorectal cancer incidence among adults with low-density lipoprotein (LDL) cholesterol below 5 mmol/L. We then emulated the target trial using linked electronic health records of 752,469 eligible UK adults (CALIBER 1999-2016) under both a cohort design and case-control sampling of the cohort. We used pooled logistic regression to estimate intention-to-treat and per-protocol effects of statins on colorectal cancer with adjustment for baseline and time-varying risk factors via inverse-probability weighting. Finally, we compared our case-control effect estimates with those obtained using previous case-control procedures.

Results: Over the 6-year follow-up, 3596 individuals developed colorectal cancer. Estimated intention-to-treat and per-protocol hazard ratios were 1.00 (95% CI: 0.87, 1.16) and 0.90 (95% CI: 0.71, 1.12), respectively. As expected, adequate case-control sampling yielded the same estimates. By contrast, previous case-control analytic approaches yielded estimates that appeared strongly protective (odds ratio 0.57, 95% CI: 0.36, 0.91, for ≥5 vs. <5 years of statin use).

Conclusions: Our study demonstrates how to explicitly emulate a target trial using case-control data to reduce discrepancies between observational and randomized trial

evidence. This approach may inform future case-control analyses for comparative effectiveness research.

Keywords: Case-control, causal inference, comparative effectiveness, electronic health records, target trial

KEY MESSAGES

- Previous case-control studies have reported a strong association between statin use and lower cancer risk; it is unclear whether this reflects a benefit of statins or is the result of design decisions that cannot be mapped to a (hypothetical) target trial (that would answer the question of interest).
- A target trial can be emulated using case-control data by (1) specifying the protocol of the target trial that would have answered the causal question of interest, (2) defining the observational cohort study that explicitly emulates this target trial, and (3) sampling cases and controls from that cohort.
- This approach reduces bias in the effect estimates derived from case-control studies and minimizes discrepancies between observational and randomized trial evidence.
- Case-control analyses that deviate from this approach may lead to severe bias, particularly on the multiplicative scale.

INTRODUCTION

Many important clinical decisions must be made in the absence of evidence from randomized trials, which may be impractical or too lengthy to provide a timely answer. In these cases, we resort to analyses of observational data to emulate the target trial that we would have liked to conduct and provide the best available evidence to inform decision-making.[1,2]

The target trial approach has mostly been applied to cohort (follow-up) studies, but it can be readily extended to case-control studies when (i) the goal is to estimate relative (not absolute) risks or rates and (ii) information on treatments or confounders is not available for the entire cohort but can be obtained for a smaller subset of cases and controls.[3] It is well-known that an analysis of the entire cohort and an analysis of the case-control data (which is just an efficient sampling from the underlying cohort) are expected to yield identical results.[4] However, for these estimates to be equivalent (and meaningful), both the cohort analysis and case-control analysis must estimate the same quantities as the target trial. For example, if adjustment for time-varying confounding or selection bias due to loss to follow-up is required to emulate the target trial in the cohort, then such adjustment is also required to emulate the target trial using the case-control data.

Therefore, like any study that attempts to emulate a target trial, case-control designs generally require an explicit definition of the start of follow-up (time zero) as well as data on time-varying treatments and time-varying confounders from the start of follow-up. Deviations from the target trial may lead to bias in case-control studies as in cohort studies.

Consider the example of statins and cancer. Several case-control studies have reported a strong association between statin use and lower cancer risk.[5-10] For example, a case-control study reported a substantially lower risk of colorectal cancer among long-term statin users compared with shorter-term and nonusers.[6] The magnitude of this protective estimate is implausible, and it is not compatible with the estimates from meta-analyses of randomized trials (odds ratio for colon cancer 0.95, 95% CI: 0.73, 1.25).[11,12] This lower risk is also unlikely to be entirely explained by confounding, because the indications for statins (*e.g.*, elevated low-density lipoprotein [LDL] cholesterol) are not such strong drivers of colorectal cancer risk.

Here we estimate the effect of statins on colorectal cancer using observational data from electronic health records. We use a case-control design rather than, as we did previously,[13] a cohort design, and we add linkage of electronic health records from primary care, hospital, and death registries. To describe how a target trial can be emulated using case-control data, we first specify the protocol of the (hypothetical) target trial that would have answered the causal question of interest, then define the observational cohort study that explicitly emulates this target trial, and finally sample cases and controls from that cohort. We show that a case-control design that deviates from the target trial may lead to implausible estimates similar to those previously reported.

METHODS

Target trial specification

We specified the protocol of a target trial to estimate the effect of statins on colorectal cancer incidence among adults with LDL cholesterol below 5 mmol/L.[13] **Table 1** summarizes the key protocol components (see also **Appendix 1**). Briefly, the eligibility criteria include age ≥30, no history of cancer, no statin contraindication, no statin prescription within the past year, and LDL cholesterol <5 mmol/L. The treatment strategies to be compared are initiation of any statin therapy at baseline and continuation over follow-up until the development of a contraindication (hepatic impairment or myopathy) and no initiation of statin therapy over follow-up unless there is an indication (LDL cholesterol ≥5 mmol/L). Participants are followed for up to six years or until colorectal cancer diagnosis.

Target trial emulation

We explicitly emulated this target trial under both a cohort design and a case-control sampling of the cohort using observational data from the Clinical Practice Research Datalink, Hospital Episode Statistics and Office of National Statistics: population-based datasets comprised of longitudinal electronic health records from primary care, hospital and death registries, accessed through the CALIBER resource (see also **Appendix 1**).[14,15]

Cohort analysis

We mirrored each protocol component as closely as possible, with several modifications to accommodate our use of observational data (**Table 1**). For example, to assess baseline confounders, we required information on lab values measured during the past year and lifestyle factors during the past four years. We classified individuals into two groups according to their prescription records at baseline. We assumed these groups were exchangeable at baseline conditional on the covariates in **Table 2**. The analysis proceeded as for the target trial, with adjustment for these baseline covariates via standardization in an attempt to emulate randomization (see also **Appendix 2**)**.**

Case-control analysis

We sampled cases and controls from the assembled cohort of eligible individuals via incidence density sampling.[16] Cases were all individuals diagnosed with colorectal cancer over the study period. Controls were individuals who were alive, under follow-up, and free of colorectal cancer at the time of selection. To reduce differences due to random variability when comparing the cohort and case-control estimates, we randomly selected 1,000 controls per case (case-control studies are typically based on a much lower number of controls). The analysis of the case-control data proceeded as for the cohort analysis (see also **Appendix 3**). The odds ratio from the case-control data is an unbiased estimator of the rate ratio obtained from the full cohort.[4] Therefore, if the cohort analysis correctly estimates the hazard ratios from the target trial in **Table 1**, then the case-control analysis does too.

Deviations from the target trial

In separate analyses, we applied the analytic approach of a previous case-control study to our data to demonstrate how deviations from the target trial framework lead to bias. The previous study reported an odds ratio of 0.53 (95% CI: 0.38, 0.74) when comparing colorectal cancer cases and controls in terms of their statin use: ≥5 vs. <5 years.[6] To assess statin use and potential confounders, eligible cases and controls were interviewed in person by the research team. This study deviated from its corresponding target trial in several ways.

First, the analysis was restricted to eligible cases and controls who could be interviewed. That is, individuals had to remain alive and under follow-up for a period after being selected for the study. The length of the period between selection and interview is unknown, but the authors reported that 19.4% of eligible cases could not be located or approached because they had died or been lost to follow-up.[6] In our study, a similar 18.7% loss of eligible cases would require a three-month period between selection and interview. This three-month survival requirement does not exist in the target trial.

Second, cases and controls were classified based on their observed cumulative duration of statin therapy through the time of diagnosis (selection) for cases but through the time of interview (post-selection) for controls. Compared with the target trial, this approach corresponds to neither the intention-to-treat analysis (which assigns individuals to a treatment strategy based on baseline information only) nor the per-protocol analysis (which assigns individuals to a treatment strategy based on baseline information and then censors them at deviation from the baseline assignment). Further,

this case-control study used a longer period of potential statin use for controls (baseline to interview) than for cases (baseline to diagnosis).

Third, the analysis adjusted for covariates assessed at the time of interview. From a target trial perspective, this is equivalent to adjusting for variables measured at or after the end of follow-up. By contrast, a correct intention-to-treat analysis adjusts for baseline confounders and a correct per-protocol analysis adjusts for baseline and post-baseline (time-varying) confounders during the follow-up. Because this case-control study ignored time-varying confounders, the analysis did not need inverse-probability weighting.

Fourth, the study included cases and controls who were using statins before baseline (prevalent users) and used pre-baseline statin therapy to quantify total duration of use. These individuals would not be eligible for the target trial.

To assess the cumulative impact of these deviations from the target trial on the estimates, we sequentially implemented them in our own case-control analysis. First, we restricted our case-control analysis to individuals alive and under follow-up three months after selection. We also implemented an equivalent cohort analysis that excludes all monthly records within three months of death or censoring. As a sensitivity analysis, we examined a six-month (rather than three-month) survival requirement.

Second, we classified cases and controls by their cumulative duration of statin use (≥5 vs. <5 years) after baseline through selection for cases and through selection + three months for controls. Again, we implemented an equivalent cohort analysis that (1) excludes all monthly records within three months of death or censoring and (2)

assesses cumulative statin use through the current month for event person-months and through the current month + three months for non-event person months.

Third, we adjusted for covariates measured at the time of selection, instead of at baseline or later, by including them in the pooled logistic model. We were unable to use pre-baseline statin therapy to quantify total duration of use because we lacked complete pre-baseline histories for some individuals in the cohort.

Statins and all-cause mortality

To show the generality of our approach, we repeated these analyses for statin therapy and all-cause mortality. We selected all-cause mortality as an alternative outcome because the magnitude of the intention-to-treat effect of statins on all-cause mortality is known from randomized trials (risk ratio 0.86, 95% CI: 0.80, 0.93) and can be used as a benchmark.[17] We emulated a target trial using the same data, with additional eligibility criteria of no cardiovascular disease at baseline and an increased cardiovascular risk (defined as LDL cholesterol ≥3.4 mmol/L) and with up to 10 years of follow-up. Here, replicating a three-month survival requirement after selection only resulted in a loss of controls, not cases.

All analyses were conducted using SAS 9.4 (SAS Institute, Inc., Cary, NC, USA).

RESULTS

Figure 1 shows a flowchart of participant selection, and **Table 2** shows baseline characteristics of the 752,469 eligible individuals in the cohort analysis and the 3596 cases and 3,596,000 controls in the case-control analysis. Compared with statin non-initiators at baseline, statin initiators were, on average, older and had higher LDL cholesterol and BMI, and included a higher proportion of men, current smokers, antihypertensive and aspirin users, and individuals with cardiovascular disease and diabetes. Compared with controls, cases were, on average, older and included a higher proportion of men, former smokers, antihypertensive and aspirin users, and individuals with cardiovascular disease and diabetes.

Table 3 shows estimated 6-year risk differences and hazard ratios when emulating a target trial of statin therapy and colorectal cancer. In the full cohort, the estimated 6-year risk differences were 0% (95% CI: -0.1%, 0.2%) in the intention-to-treat analysis and -0.1% (95% CI: -0.2%, 0.1%) in the per-protocol analysis, and the estimated hazard ratios were 1.00 (95% CI: 0.87, 1.16) in the intention-to-treat analysis and 0.90 (95% CI: 0.71, 1.12) in the per-protocol analysis. The odds ratios from the case-control sample were identical to the hazard ratios from the cohort. Estimated hazard ratios were identical when additionally adjusting for cancer screening in the past year (data not tabled). Estimated hazard ratios were similar when only adjusting for age (intention-to-treat hazard ratio 1.03, 95% CI: 0.89, 1.19; per-protocol hazard ratio 0.97, 95% CI: 0.80, 1.20) (data not tabled).

We then replicated the approach of the previous case-control study in our data (**Table 3**). The estimated odds ratio for colorectal cancer was 0.55 (95% CI: 0.35, 0.87)

when we imposed the three-month survival requirement and assessed cumulative statin use through the time of selection (diagnosis) for cases and through the time of selection + 3 months for controls. When imposing this survival requirement and instead assessing statin use through the time of selection for both cases and controls, the odds ratio estimate was 0.84 (95% CI: 0.53, 1.34) (data not tabled). Estimates were similar when adjusting for covariates measured at the time of selection instead of at baseline. Cohort analyses mimicking these decisions returned the same estimates. A six-month survival requirement yielded stronger inverse associations (**Table S2**).

Imposing the three-month survival requirement resulted in a loss of 672 eligible cases (18.7%), including 418 who died (11.6%) (similar to the proportions reported in the published study: 19.4% and 8.6%, respectively), and a loss of 298,380 eligible controls (8.3%), including 13,047 who died (0.4%) (**Figure 2**). Among individuals who remained alive and under follow-up three months after selection, a slightly lower proportion of cases was classified as having ≥5 years of statin use, compared with the distribution of exposure at the time of selection (0.6% vs. 1.0%). In addition, 6847 surviving controls were re-classified as having ≥5 years of statin use when statin use was assessed through selection + three months. These small shifts in absolute proportions (slight depletion of cases, and enrichment of controls, for ≥5 years of statin use) are responsible for the large shifts on the odds ratio (multiplicative) scale.

Statins and all-cause mortality

When emulating a target trial of statins and all-cause mortality, we estimated an intention-to-treat hazard ratio of 0.87 (95% CI: 0.79, 0.95), which is close to the estimate

from a meta-analysis of randomized trials (risk ratio 0.86, 95% CI: 0.80, 0.93) (**Table S3**).[17] This estimate progressively decreased when we applied the analytic flaws described above (**Table S3**).

DISCUSSION

After emulating a target trial using the electronic health records of 752,469 adults with up to six years of follow-up, we found little evidence that the risk of colorectal cancer differs between statin users and nonusers. This finding is consistent with meta-analyses of randomized trials.[11,12] As expected, adequate case-control sampling returned the same estimates as the cohort analysis. By contrast, after replicating the analytic approach of a previous case-control study in our data, we found implausibly protective estimates similar to those previously reported.

Case-control studies may have a role to play when conducting causal inference research based on healthcare databases. While such databases provide access to the underlying cohort that gives rise to cases and controls, they may not contain high-quality information on treatment or confounders needed to answer certain causal questions.[3] In these settings, case-control studies allow us to focus limited resources on collecting this information for random samples of cases and controls.

Case-control analyses may seem simple: we compare the treatment status of cases with non-cases. However, a failure to anchor this to an underlying cohort study that explicitly emulates a target trial contributes to two common misconceptions about case-control analyses: (1) that they are immune to many of the biases that afflict cohort analyses, such as time-varying confounding and selection bias due to loss to follow-up and the inclusion of prevalent users, and (2) that they do not require complete treatment and confounder history for cases and controls. While critics of case-control designs within existing databases have largely focused on design flaws leading to confounding

bias,[18] our evaluation showed that other deviations from a target trial in case-control analyses lead to the same biases that affect cohort analyses.

Two deviations from the target trial appeared to drive the biased estimates in this particular application: (1) the requirement for cases and controls to survive for three additional months and (2) the assessment of treatment duration over a longer time period for controls compared with cases. Together, these decisions led to small shifts in treatment classification that depleted cases and enriched controls for ≥5 years of statin use. Importantly, we found that effect estimates on the multiplicative scale, which are generally all that we can obtain from case-control studies, may be particularly susceptible to these biases.

Other deviations from the target trial may, in general, matter. First, comparing cumulative duration of treatment above vs. below a certain threshold (*e.g.,* ≥5 vs. <5 years) does not capture information on the precise timing, duration, or reasons for switching treatment, which may be important for risk. In our analysis, the estimated odds ratio comparing ≥5 vs. <5 years of statin use (0.95, 95% CI: 0.67, 1.34; with no survival requirement, data not tabled) was similar to the intention-to-treat hazard ratio (1.00, 95% CI: 0.86, 1.16), possibly because treatment had no effect on the outcome in this particular application. Second, adjustment for variables measured at (or after) the time of selection will not appropriately adjust for confounding and may induce selection bias. In our analysis, this had little impact possibly because, as suggested by the similarity between age- and fully-adjusted estimates, the adjustment variables were not strong predictors of the outcome no matter when they were measured. Third, failure to adjust for loss to follow-up may result in selection bias if remaining uncensored depends

on treatment history and risk factors. In our analysis, estimates were similar when additionally applying inverse-probability weights for censoring due to loss to follow-up. Lastly, including prevalent users at baseline may contribute to selection bias due to the selection of individuals who received pre-baseline treatment for some time and remained at-risk and under follow-up at baseline. We were unable to explore this deviation in our data. Our approach of explicitly specifying the protocol of the target trial and its observational emulation naturally leads to analytic approaches that prevent these biases.

Our study has several additional strengths. The volume and variety of data in the electronic health records allowed us to evaluate statins and colorectal cancer in a population-based sample with adjustment for many potential confounders. Our analytic approach allowed us to estimate both relative and absolute risks under sustained strategies that realistically depend on dynamic clinical features. Lastly, our analyses of all-cause mortality support that the target trial approach can reproduce effect estimates from trials and that the analytic flaws described above will result in bias for this alternative outcome.

Nevertheless, we were limited by our reliance on diagnosis codes and prescription records, which may contribute to measurement error and residual confounding. However, previous validation studies have confirmed a high proportion of recorded cancers (95%) and other diagnoses in this database.[19,20]

In summary, our findings suggest that flaws in case-control analyses can be mapped to decisions in a cohort analysis that would lead to bias, particularly on the multiplicative scale. Explicitly mapping case-control sampling to the target trial helped

us to reduce bias. Our approach may help to inform the design and analysis of any case-control study where the goal is to assess the benefit-risk of medical treatments.

References

1. Hernán 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.
2. Hernán 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.
3. Schneeweiss S, Suissa S. Discussion of Schuemie et al: "A plea to stop using the case-control design in retrospective database studies". *Stat Med.* 2019;38(22):4209-4212.
4. Miettinen O. Estimability and estimation in case-referent studies. *Am J Epidemiol.* 1976;103(2):226-235.
5. Graaf MR, Beiderbeck AB, Egberts AC, Richel DJ, Guchelaar HJ. The risk of cancer in users of statins. *J Clin Oncol.* 2004;22(12):2388-2394.
6. Poynter JN, Gruber SB, Higgins PD, et al. Statins and the risk of colorectal cancer. *N Engl J Med.* 2005;352(21):2184-2192.
7. Khurana V, Bejjanki HR, Caldito G, Owens MW. Statins reduce the risk of lung cancer in humans: a large case-control study of US veterans. *Chest.* 2007;131(5):1282-1288.
8. Shannon J, Tewoderos S, Garzotto M, et al. Statins and prostate cancer risk: a case-control study. *Am J Epidemiol.* 2005;162(4):318-325.
9. Hoffmeister M, Chang-Claude J, Brenner H. Individual and joint use of statins and low-dose aspirin and risk of colorectal cancer: a population-based case-control study. *Int J Cancer.* 2007;121(6):1325-1330.
10. Boudreau DM, Gardner JS, Malone KE, Heckbert SR, Blough DK, Daling JR. The association between 3-hydroxy-3-methylglutaryl conenzyme A inhibitor use and breast carcinoma risk among postmenopausal women: a case-control study. *Cancer.* 2004;100(11):2308-2316.
11. Dale KM, Coleman CI, Henyan NN, Kluger J, White CM. Statins and cancer risk: a meta-analysis. *Jama.* 2006;295(1):74-80.
12. Cholesterol Treatment Trialists C, Emberson JR, Kearney PM, et al. Lack of effect of lowering LDL cholesterol on cancer: meta-analysis of individual data from 175,000 people in 27 randomised trials of statin therapy. *PLoS One.* 2012;7(1):e29849.
13. Dickerman BA, García-Albéniz X, Logan RW, Denaxas S, Hernán MA. Avoidable flaws in observational analyses: an application to statins and cancer. *Nature Medicine.* 2019;25:1601-1606.
14. Denaxas SC, George J, Herrett E, et al. Data resource profile: cardiovascular disease research using linked bespoke studies and electronic health records (CALIBER). *Int J Epidemiol.* 2012;41(6):1625-1638.
15. Denaxas S, Gonzalez-Izquierdo A, Direk K, et al. UK phenomics platform for developing and validating electronic health record phenotypes: CALIBER. *J Am Med Inform Assoc.* 2019.
16. Rothman K, Greenland S, Lash TL. Case-Control Studies. In: *Modern Epidemiology.* 3rd ed. Philadelphia: Lippincott Williams & Wilkins; 2008.

17. Chou R, Dana T, Blazina I, Daeges M, Jeanne TL. Statins for prevention of cardiovascular disease in adults: evidence report and systematic review for the US Preventive Services Task Force. *Jama.* 2016;316(19):2008-2024.
18. Schuemie MJ, Ryan PB, Man KKC, Wong ICK, Suchard MA, Hripcsak G. A plea to stop using the case-control design in retrospective database studies. *Stat Med.* 2019;38(22):4199-4208.
19. Margulis AV, Fortuny J, Kaye JA, et al. Validation of cancer cases using primary care, cancer registry, and hospitalization data in the United Kingdom. *Epidemiology.* 2018;29(2):308-313.
20. Herrett E, Thomas SL, Schoonen WM, Smeeth L, Hall AJ. Validation and validity of diagnoses in the General Practice Research Database: a systematic review. *Br J Clin Pharmacol.* 2010;69(1):4-14.
21. Hernán MA, Lanoy E, Costagliola D, Robins JM. Comparison of dynamic treatment regimes via inverse probability weighting. *Basic Clin Pharmacol Toxicol.* 2006;98(3):237-242.

Table 1. Specification and emulation of a target trial of statin therapy and colorectal cancer risk using observational data from linked electronic health records accessed through the CALIBER resource.

		Target trial emulation	
Protocol	**Target trial specification**	**Cohort analysis**	**Case-control analysis**
Eligibility criteria	• Aged ≥30 years between 1 January 1998 and 29 February 2016 • No history of cancer (except nonmelanoma skin cancer) • No statin contraindication (hepatic impairment or myopathy) • No statin prescription within the past year • LDL cholesterol <5 mmol/L • At least 1 year of up-to-standard data in a CPRD practice • At least 1 year of potential follow-up Baseline is defined as the first month in which all eligibility criteria are met.	Same as for the target trial. We defined hepatic impairment as a code for hepatic failure or ALT ≥120 IU/L and myopathy as codes for its symptoms: muscle aches, pain, or weakness. We required information on lab values measured during the past year and lifestyle factors during the past four years.	Same as for the cohort analysis. We performed incidence density sampling of the eligible individuals, selecting 1,000 controls per 1 colorectal cancer case.
Treatment strategies	(1) Initiation of any statin therapy at baseline and continuation over follow-up until the development of a contraindication (hepatic impairment or myopathy) (2) No initiation of statin therapy over follow-up until the development of an indication (LDL cholesterol ≥5 mmol/L) Treatment is considered continuous if there is a gap of <30 days between successive prescriptions. When clinically warranted during the follow-up, patients and their physicians will decide whether to start, stop, or switch therapy. Participants must have a primary care consultation at least once every 4 years to assess prognostic factors associated with adherence and loss to follow-up.	Same as for the target trial. We defined the date of medication initiation to be the first date of a prescription. We calculated discontinuation dates using the daily dose and quantity of pills in the prescription.	Same as for the cohort analysis.
Treatment assignment	Individuals are randomly assigned to a strategy at baseline, and individuals and their treating physicians will be aware of the assigned treatment strategy.	We classified individuals according to the strategy that their data were compatible with at baseline and attempted to emulate randomization by adjusting for baseline confounders.	Same as for the cohort analysis.
Outcomes	Colorectal cancer.	Same as for the target trial. Colorectal cancer diagnoses were recorded as Read codes and ICD-10 codes.	Same as for the cohort analysis.
Follow-up	Starts at baseline and ends at the month of colorectal cancer diagnosis, death, loss to follow-up (transfer out of the practice or incomplete follow-up [four years after the last recorded prognostic factors]), six years after baseline, or administrative end of follow-up (end of practice data collection or 29 February 2016), whichever happens first.	Same as for the target trial.	Same as for the cohort analysis.
Causal contrasts	Intention-to-treat effect and per-protocol effect.	Observational analog of intention-to-treat and per-protocol effect.	Same as for the cohort analysis.
Statistical analysis	Intention-to-treat analysis: apply inverse-probability weights to adjust for pre- and post-baseline prognostic factors associated with loss to follow-up. Per-protocol analysis: censor individuals if and when they deviate from their assigned treatment strategy and apply inverse-probability weights to adjust for pre- and post-baseline prognostic factors associated with adherence and loss to follow-up.[21]	Same as for the target trial with adjustment for baseline confounders.	Same as for the cohort analysis.

Abbreviations: ALT, alanine transaminase; CPRD, Clinical Practice Research Database; LDL, low-density lipoprotein.

Table 2. Baseline characteristics of eligible individuals in the cohort analysis and selected individuals in the case-control analysis when emulating a target trial of statin therapy and colorectal cancer risk, CALIBER, 1999-2015*.

	Cohort analysis		Case-control analysis	
Characteristic, mean (SD) or %	**Initiators** (n=25,032)	**Non-initiators** (n=727,437)	**Cases** (n=3596)	**Controls** (n=3,596,000)
Age (years)	62.7 (11.6)	55.9 (13.7)	68.5 (10.7)	56.7 (13.4)
Female, %	42	53	43	52
Body mass index (kg/m^2)	28.8 (5.6)	28.0 (5.7)	27.8 (5.1)	28.2 (5.7)
Smoking status, %				
Never	43	54	49	53
Former	32	27	37	28
Current	25	19	14	19
Low-density lipoprotein cholesterol (mmol/L)	3.7 (0.9)	3.3 (0.8)	3.3 (0.8)	3.3 (0.8)
High-density lipoprotein cholesterol (mmol/L)	1.4 (0.4)	1.5 (0.4)	1.4 (0.4)	1.4 (0.4)
Coronary heart disease, %	9	2	5	3
Hypertension, %	27	17	24	19
Cerebrovascular disease, %	2	1	1	1
Other cardiovascular disease†, %	16	14	19	14
Diabetes, %	18	5	9	7
Antihypertensive use‡, %	54	30	50	34
Aspirin use, %	29	7	17	9
Hormone replacement therapy, % of women	3	4	2	4
Oral contraceptive use, % of women	4	7	2	7
Referrals in the past three months, ≥2, %	4	2	3	2

* Baseline ranges from January 1999 to February 2015.

† Includes acute rheumatic fever, chronic rheumatic heart disease, pulmonary heart disease, and other circulatory disease.

‡ Includes all primary care prescriptions from British National Formulary chapters 2.2.1 thiazides and related diuretics, 2.2.3 potassium-sparing diuretics and aldosterone antagonists, 2.2.4 potassium-sparing diuretics with other diuretics, 2.4 beta-adrenoceptor blocking drugs, 2.5 hypertension and heart failure, 2.6.2 calcium-channel blockers.

Table 3. Estimated risk of colorectal cancer comparing statin therapy with no statin therapy, CALIBER, 1999-2016.

		Case-control analysis		Cohort analysis					
						6-year risk (%)			
	Cases	Odds ratio	95% CI	Hazard ratio	95% CI	Initiator	Non-initiator†	Risk Difference (%)	95% CI
Emulating a target trial*									
Intention-to-treat‡	3596	1.00	0.86, 1.16	1.00	0.87, 1.16	0.8	0.8	0	-0.1, 0.2
Per-protocol§	2735	0.90	0.71, 1.15	0.90	0.71, 1.12	0.8	0.9	-0.1	-0.2, 0.1
Replicating the approach of a previous case-control study‖									
Imposing a 3-month survival requirement from the time of selection¶	2924	1.02	0.86, 1.20	1.02	0.86, 1.20	0.8	0.7	0.1	-0.1, 0.2
+ Comparing ≥5 vs. <5 years of statin use**	2924	0.55	0.35, 0.87	0.55	0.35, 0.87	--	--	--	--
+ Adjusting for covariates instead measured at the time of selection	2924	0.57	0.36, 0.91	0.57	0.36, 0.91	--	--	--	--

Abbreviation: CI, confidence interval.

* Estimates from weighted pooled logistic regression models adjusted for age, sex, BMI, smoking status, LDL cholesterol, HDL cholesterol, months since last measure of LDL cholesterol, months since last measure of HDL cholesterol, coronary heart disease, hypertension, cerebrovascular disease, other cardiovascular disease, diabetes, antihypertensive use, aspirin use, hormone replacement therapy, oral contraceptive use, number of referrals in the past three months. The number of cases is lower in the per-protocol analysis because of the censoring under this approach (see also Appendix 1).

† Refers to statin use for <5 years when replicating the previous case-control approach.

‡ Comparing statin initiation vs. no initiation at baseline.

§ Comparing statin initiation at baseline and continuation over follow-up unless contraindicated with no statin initiation over follow-up unless indicated.

‖ Estimates from unweighted pooled logistic regression models adjusted for the covariates above, assessed at baseline.

¶ Comparing treatment initiation vs. no initiation at baseline. In the case-control sample, the analysis was restricted to individuals alive and under follow-up 3 months after selection. In the full cohort, the analysis excluded monthly records within 3 months of death or censoring.

**In the case-control sample, (1) the analysis was restricted to individuals alive and under follow-up 3 months after selection and (2) cumulative statin use after baseline was assessed through the time of selection (diagnosis) for cases and through the time of selection + 3 months for controls. In the full cohort, (1) the analysis excluded monthly records within 3 months of death or censoring and (2) cumulative statin use was assessed through the current month for event person-months and through the current month + 3 months for non-event person months.

Figure 1. Flowchart for selection of eligible individuals from CALIBER when emulating a target trial of statin therapy and colorectal cancer risk, 1999-2016.

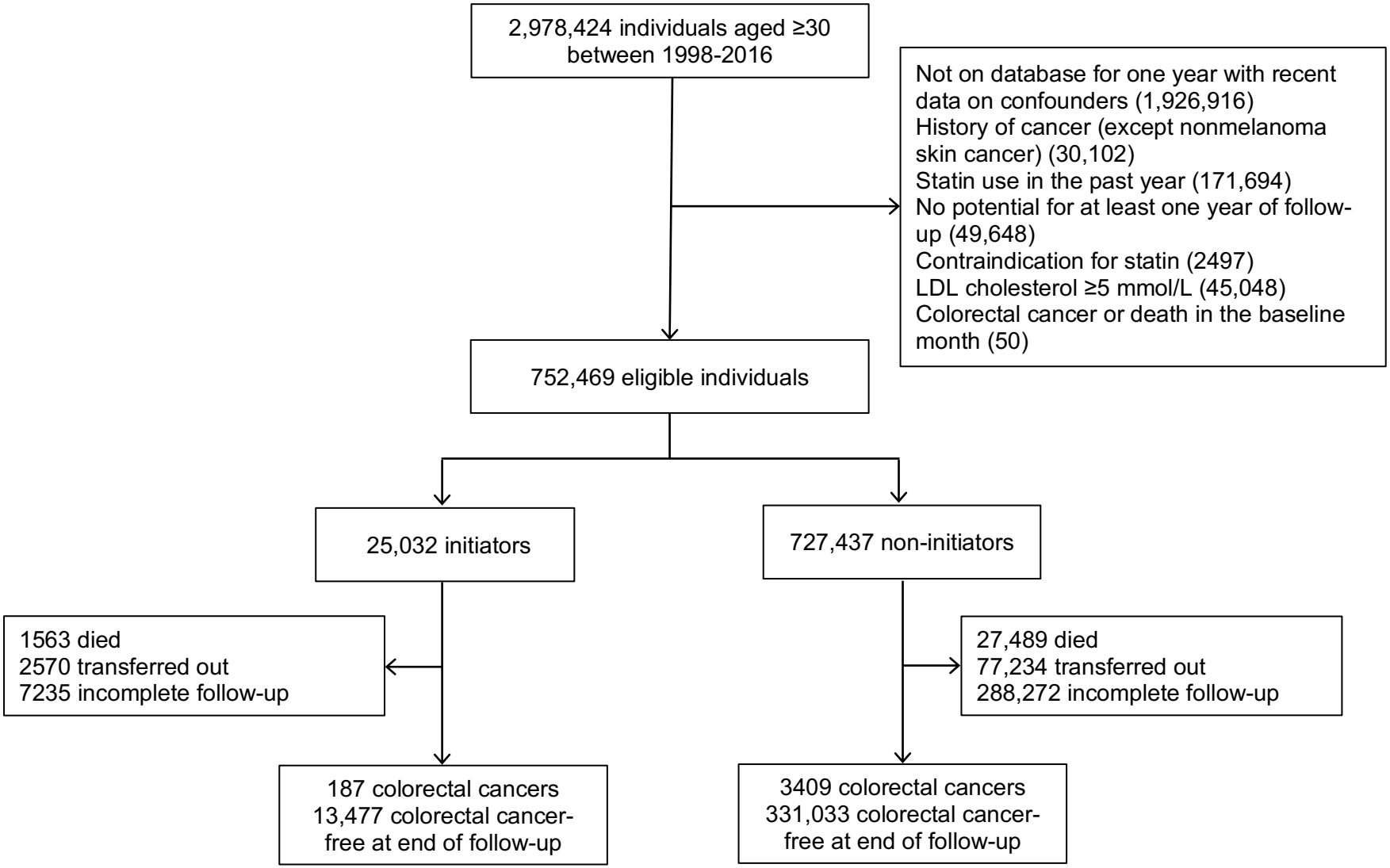


Figure 2. Distribution of statin exposure among cases and controls under no survival requirement (**Panel A**) and a 3-month survival requirement (**Panel B**) from the time of selection, and proportions of individuals lost to various causes (**Panel C**). In addition, 6847 surviving controls who were classified as having <5 years of statin use under no survival requirement were re-classified as having ≥5 years of statin use under the 3-month survival requirement.

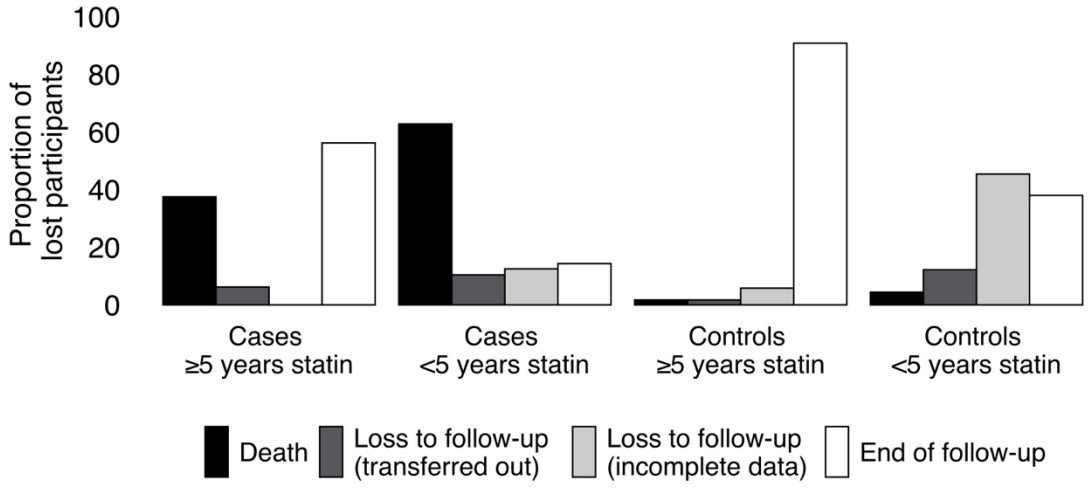
A)

	Cases	**Controls**
≥5 years statin use	35	20,150
<5 years statin use	3561	3,575,850
Total	3596	3,596,000

B)

	Cases	**Controls**
≥5 years statin use	19	20,116
<5 years statin use	2905	3,277,504
Total	2924	3,297,620

C)



Emulating a target trial in case-control designs: an application to statins and colorectal cancer

Barbra A. Dickerman, PhD, Xabier García-Albéniz, MD, PhD, Roger W. Logan, PhD, Spiros Denaxas, PhD, Miguel A. Hernán, MD, DrPH

TABLE OF CONTENTS

Appendix 1

Target trial specification

We have previously described a target trial to investigate the effect of statins on cancer.[1] Briefly, the protocol of the target trial has the following components (see also **Table 1**):

Eligibility criteria: age ≥30 between January 1998 and February 2016, no history of cancer (except nonmelanoma skin cancer), no statin contraindication (hepatic impairment or myopathy), no statin prescription within the past year, LDL cholesterol <5 mmol/L, at least one year of up-to-standard data in a Clinical Practice Research Database (CPRD) practice, and at least one year of potential follow-up. Baseline is defined as the first month in which all eligibility criteria are met.

Treatment strategies. The dynamic strategies to be compared are: (i) initiation of any statin therapy at baseline and continuation over follow-up until the development of a contraindication (hepatic impairment or myopathy), and (ii) no initiation of statin therapy over follow-up unless there is an indication (LDL cholesterol ≥5 mmol/L). Treatment is considered to be continuous if there is a gap of less than 30 days between successive prescriptions. When clinically warranted over follow-up (*i.e.,* upon the development of these indications and contraindications), patients and their physicians will decide whether to start, stop, or switch therapy. Participants must have a primary care consultation at least once every four years to assess prognostic factors associated with adherence and loss to follow-up.

Treatment assignment. Individuals are randomly assigned to a strategy at baseline. Individuals and their treating physicians will be aware of the assigned treatment strategy.

Outcome. Colorectal cancer incidence.

Follow-up. Each individual is followed from baseline until the month of colorectal cancer diagnosis, death, loss to follow-up (transfer out of the practice or incomplete follow-up [four years after the last recorded prognostic factors]), six years after baseline, or administrative end of follow-up (end of practice data collection or February 2016), whichever happens first.

Causal contrasts: the intention-to-treat effect of being assigned to treatment initiation vs. no initiation at baseline and the per-protocol effect of adhering to assigned treatment strategies on colorectal cancer incidence over follow-up.

Statistical analysis. In the intention-to-treat analysis, a pooled logistic regression model is fit to estimate intention-to-treat effects via hazard ratios and standardized risk differences. The model contains an indicator of assigned strategy, a flexible function of months since randomization (linear and quadratic terms) and, if necessary, the baseline covariates. Given a low monthly risk of colorectal cancer, the odds ratio from this model approximates a hazard ratio comparing treatment initiation vs. no initiation.[2] Time-varying

inverse-probability weights are used to adjust for potential selection bias related to loss to follow-up.[3] Because this additional adjustment did not influence our point estimates, we omit this from the primary analysis for simplicity.

In the per-protocol analysis, individuals are censored if and when they deviate from their assigned treatment strategy. Specifically, individuals in the initiator group are censored when they stop statins (in the absence of a contraindication) and individuals in the non-initiator group are censored when they start statins (in the absence of an indication). Time-varying inverse-probability weights are used to adjust for pre- and post-baseline prognostic factors associated with adherence to the assigned treatment strategy and loss to follow-up.[3] The weights for loss to follow-up again had a negligible influence on our point estimates and are omitted for simplicity. Individuals who stop statins because of a contraindication or start statins because of an indication are not censored and their weights for adherence remain constant from the time that one of these conditions develops until the end of follow-up. The pooled logistic model described in the previous paragraph is fit to the uncensored data with each person-month weighted by its corresponding inverse-probability weight. Estimated weights are truncated at their 99th percentile to prevent outliers from affecting the analyses. See **Appendix 2 and Table S1** for details on covariates and models.

Absolute risks under each strategy are estimated by fitting these pooled logistic regression models with an additional product term between treatment and follow-up time. The predicted values from this model are used to estimate the 6-year risk of colorectal cancer under each strategy, which is standardized to the joint distribution of the baseline covariates.

Nonparametric bootstrapping with 500 samples is used to calculate percentile-based 95% confidence intervals for hazard ratio and risk difference estimates.

Target trial emulation

We explicitly emulated this target trial under both a cohort design and a case-control sampling of the cohort using observational data from the Clinical Practice Research Datalink, Hospital Episode Statistics, and Office of National Statistics: population-based datasets comprised of longitudinal UK electronic health records from primary care, hospital, and death registries, accessed through the CALIBER resource.[4,5] Longitudinal primary care data on demographics, lifestyle, symptoms, diagnoses, clinical examination findings, laboratory test results, referrals, and prescriptions are recorded by general practitioners in the CPRD. Hospitalization data are obtained through linkage with Hospital Episode Statistics. Cause-specific mortality data are obtained through linkage with the Office of National Statistics. This linkage extends our previous work to emulate target trials of statins and cancer,[1] which used CPRD primary care electronic health records. Disease phenotypes are derived using algorithms that combine information on diagnoses, symptoms, laboratory values, physiological measures, prescriptions, and procedures, which were created and validated using an established methodology.[6,7]

Cohort analysis

Eligibility criteria. We applied the same eligibility criteria and definition of baseline as for the target trial, requiring laboratory values measured during the past year and lifestyle factors during the past four years.

Treatment strategies. We defined the date of medication initiation to be the first date of a prescription. We calculated discontinuation dates using the daily dose and quantity of pills in the prescription.

Treatment assignment. We classified individuals into one of two groups according to the strategy that their data were compatible with at baseline. We assumed groups were exchangeable at baseline conditional on baseline covariates: *demographics* (age, sex), *lifestyle factors* (body mass index, smoking status), *laboratory measurements* (LDL and HDL cholesterol and time since their last measurement), *diagnoses* (coronary heart disease, hypertension, cerebrovascular disease, other cardiovascular disease, diabetes), *medications* (antihypertensives, aspirin, hormone replacement therapy, oral contraceptives), and *healthcare utilization* (number of any specialist referrals in the past three months).

Outcome. Colorectal cancer diagnoses were recorded as Read codes in primary care and as ICD-10 codes in hospitals.

Follow-up. Same as for the target trial.

Causal contrasts. Observational analogue of the intention-to-treat and per-protocol effects.

Statistical analysis. Same as for the target trial with adjustment for baseline confounders.

Case-control analysis

We sampled cases and controls from the cohort of eligible individuals described above via incidence density sampling (see also **Appendix 2**).[8] Cases were all individuals diagnosed with colorectal cancer over the study period. Controls were individuals who were alive, under follow-up, and free of colorectal cancer at the time of selection. To reduce differences due to random variability when comparing the cohort and case-control estimates, we randomly selected 1,000 controls per case (case-control studies are typically based on a much lower number of controls).

We then fit to the case-control data the same pooled logistic models described above for the cohort analysis. For the per-protocol analysis, only the controls were used to estimate the inverse-probability weights.[9] The odds ratio from the case-control data is an unbiased estimator of the rate ratio obtained from the full cohort.[10] Therefore, if the cohort analysis correctly estimates the hazard ratios from the target trial in **Table 1**, then the case-control analysis does too.

Ethical approval

The CPRD has been granted generic ethical approval for observational studies that make use of only anonymized data and linked anonymized National Health Service healthcare data (Multiple Research Ethics Committee ref. 05/MRE04/87). This study was approved by the Medicines and Healthcare Products Regulatory Agency Independent Scientific Advisory Committee (protocol 16_221) and exempted by the Harvard T.H. Chan School of Public Health Institutional Review Board.

Appendix 2

Estimating the intention-to-treat hazard ratio

In our target trial, the intention-to-treat effect is the effect of being *assigned* to treatment initiation vs. no initiation at baseline on the risk (or rate) of colorectal cancer. Estimating its observational analog requires adjustment for baseline confounders. To do this, we fit a pooled logistic regression model containing an indicator of observed treatment initiation and potential confounders measured in the baseline month. Under the assumptions of no unmeasured confounding given the included covariates and a low monthly risk of the outcome within levels of those covariates, the exponentiated coefficient of the treatment indicator (*i.e.,* $exp(\alpha_1)$) validly estimates the intention-to-treat hazard ratio (averaged over follow-up) that would be seen in a target trial with a similar adherence pattern as in the observational data. Estimates were similar when we additionally applied inverse-probability weights to this model to adjust for potential selection bias due to loss to follow-up.

$$\text{logit}\left(\Pr[Y_{t+1} = 1 | A_0, L_0, \bar{Y}_t = 0]\right) = \alpha_{0,t} + \alpha_1 A_0 + {\alpha_2}^T L_0$$

The overbar indicates the history of a covariate from the start of follow-up. The superscript *T* indicates a transpose of a vector of parameters.

Y_{t+1}	Indicator for the outcome of interest at month *t*+1
$\alpha_{0,t}$	Time-varying intercept, estimated as a constant plus linear and quadratic terms for the follow-up month *t*
A_0	Indicator for treatment group
L_0	Vector of potential confounders at baseline for each individual

Estimating the per-protocol hazard ratio

In our target trial, the per-protocol effect is the effect of *adhering* to the assigned treatment strategies on the risk (or rate) of colorectal cancer. Estimating it or its observational analogue requires adjustment for baseline confounders and time-varying confounders.

First, we censored individuals if and when they deviated from their assigned treatment strategy. That is, we censored individuals in the initiator group when they discontinued statin therapy (unless a contraindication developed) and censored individuals in the non-initiator group when they initiated statin therapy (unless an indication developed). Then, we fit the below pooled logistic regression model to this censored data, additionally applying time-varying nonstabilized inverse-probability weights to adjust for time-varying confounding. We truncated weights at their 99th percentile to prevent outliers with extreme weights from affecting our estimates. Under the same assumptions described in the previous section, the exponentiated coefficient of the treatment indicator (*i.e.,* $exp(\beta_1)$) validly estimates the per-protocol hazard ratio.

$$\text{logit}\left(\Pr[Y_{t+1} = 1 | A_0, L_0, \bar{Y}_t = 0, \bar{C}_{t+1} = 0]\right) = \beta_{0,t} + \beta_1 A_0 + {\beta_2}^T L_0$$

The overbar indicates history of the variable.
The superscript *T* indicates a transpose of a vector of parameters.

Y_{t+1}	Indicator for the outcome of interest at month *t*+1
$\beta_{0,t}$	Time-varying intercept, estimated as a constant plus linear and quadratic terms for the follow-up month *t*
A_0	Indicator for treatment group
L_0	Vector of potential confounders at baseline for each individual

Subject-specific time-varying nonstabilized inverse-probability weights:

Informally, the denominator of this weight at time *t* is the probability that an individual received her observed treatment history given her covariate history by *t*. The application of these weights creates a pseudo-population in which treatment is independent of the measured confounders at all time points.

Weights for censoring due to switching treatment

$$W_t^A = \prod_{k=0}^{t} \frac{1}{f(A_k|\bar{A}_{k-1}, \bar{L}_k, \overline{Y}_{k-1} = 0)}$$

To estimate the denominator, we fit two separate models to allow the probabilities to differ according to prior treatment status.

The first model was fit to person-months who were untreated in the previous month (*i.e.,* $A_{k-1} = 0$):

$$\text{logit}\left(\Pr\left[A_k = 1 \,|A_{k-1} = 0, \bar{L}_k, \overline{Y}_{k-1} = 0\right]\right) = \eta_{0,t} + \eta_1^T L_0 + \eta_2^T L_k$$

The second model was fit to person-months who were treated in the previous month (*i.e.,* $A_{k-1} = 1$):

$$\text{logit}\left(\Pr\left[A_k = 1 \,|A_{k-1} = 1, \bar{L}_k, \overline{Y}_{k-1} = 0\right]\right) = \theta_{0,t} + \theta_1^T L_0 + \theta_2^T L_k$$

Covariate history $\bar{L}_k$ was summarized by baseline L_0 and the most recent measurement of L_k.

We excluded from the weight models the first person-month after treatment initiation, because we allowed a 30-day gap after the end of a treatment prescription and the probability of treatment in that period was therefore 1. We excluded from the first weight model above person-months with a recorded LDL cholesterol ≥5 mmol/L. We excluded from the second weight model above person-months with recorded hepatic impairment or myopathy. The final weight for each individual at each time point was the product of the weights for that individual up until that time.

Appendix 3

Intention-to-treat analysis in the case-control sample

Steps:

1. Select cases and controls from the full cohort of eligible individuals via incidence density sampling.
    a. Select all case person-months in which colorectal cancer was diagnosed.
    b. Randomly select control person-months.
    c. *Note:* Under this sampling procedure, selected controls are eligible to be later selected as a case. We randomly selected 1,000 controls per case to reduce differences due to random variability when comparing the cohort and case-control estimates; case-control studies are typically based on a much lower number of controls.
2. Fit an unweighted pooled logistic regression model for the outcome to the case-control data (see model in **Appendix 2**).
    a. *Note:* The odds ratio from this model is an unbiased estimator of the intention-to-treat hazard ratio obtained from the full cohort.

Per-protocol analysis in the case-control sample

Steps:

1. Select cases and controls from the full cohort of eligible individuals via incidence density sampling.
    a. Select all case person-months in which colorectal cancer was diagnosed.
    b. Randomly select control person-months.
    c. *Note:* Investigators then obtain, for cases and controls, treatment and confounder history from baseline through the month of selection. History is obtained through the latest month of selection if an individual was selected in multiple months.
2. Fit the weight models in **Appendix 2** to the selected control person-months and their history. Use the parameter estimates to generate predicted probabilities for all person-months (cases and controls). As in the cohort analysis, the final weight for each individual at each time point is taken as the product of the weights for that individual up until that time; this requires treatment and confounder history for selected cases and controls.
3. Censor individuals if and when they deviate from their assigned treatment strategy.
    a. *Note:* Weights can be truncated at their 99^{th} percentile after this step.
4. Fit a weighted pooled logistic regression model for the outcome to the selected cases and controls (excluding their non-selected history) (see model in **Appendix 2**). Use robust variances or bootstrapping to calculate 95% confidence intervals.
    a. *Note:* The odds ratio from this model is an unbiased estimator of the per-protocol hazard ratio obtained from the full cohort.

Table S1. Covariates* used when emulating a target trial of statin therapy and colorectal cancer, CALIBER, 1999-2016.

Covariate	Functional form	Categories
Time-fixed		
Age	Linear	N/A
Sex	Indicator	Female/Male
Time-varying		
Month of follow-up	Linear, quadratic	N/A
Body mass index	Linear	N/A
Smoking status	3 categories	Never Former Current
LDL cholesterol (and months since last measure)	Linear	N/A
HDL cholesterol (and months since last measure)	Linear	N/A
Coronary heart disease	Indicator	Yes/No
Hypertension	Indicator	Yes/No
Cerebrovascular disease	Indicator	Yes/No
Other cardiovascular disease	Indicator	Yes/No
Diabetes	Indicator	Yes/No
Antihypertensive use	Indicator	Yes/No
Aspirin use	Indicator	Yes/No
Hormone replacement therapy	Indicator	Yes/No
Oral contraceptive use	Indicator	Yes/No
Number of referrals in the past three months	3 categories	0 1 ≥2

Abbreviations: HDL; high-density lipoprotein; LDL, low-density lipoprotein.

* As described at https://www.caliberresearch.org.

Table S2. Sensitivity analysis imposing a 6-month survival requirement from the time of selection: Estimated risk of colorectal cancer* comparing statin therapy with no statin therapy, CALIBER, 1999-2016.

		Case-control analysis		Cohort analysis	
	Cases	Odds ratio	95% CI	Hazard ratio	95% CI
Replicating the approach of a previous case-control study					
Imposing a 6-month survival requirement from the time of selection†	2,612	1.02	0.86, 1.21	1.02	0.86, 1.22
+ Comparing ≥5 vs. <5 years of statin use‡	2,612	0.35	0.20, 0.62	0.35	0.20, 0.62
+ Adjusting for covariates instead measured at the time of selection	2,612	0.37	0.21, 0.65	0.37	0.21, 0.66

Abbreviation: CI, confidence interval.

* Estimates from unweighted pooled logistic regression models adjusted for baseline covariates: age, sex, BMI, smoking status, LDL cholesterol, HDL cholesterol, months since last measure of LDL cholesterol, months since last measure of HDL cholesterol, coronary heart disease, hypertension, cerebrovascular disease, other cardiovascular disease, diabetes, antihypertensive use, aspirin use, hormone replacement therapy, oral contraceptive use, number of referrals in the past three months.

† Comparing treatment initiation vs. no initiation at baseline. In the case-control sample, the analysis was restricted to individuals alive and under follow-up 6 months after selection. In the full cohort, the analysis excluded monthly records within 6 months of death or censoring.

‡ In the case-control sample: (i) the analysis was restricted to individuals alive and under follow-up 6 months after selection and (ii) cumulative statin use after baseline was assessed through the time of selection (diagnosis) for cases and through the time of selection + 6 months for controls. In the full cohort: (i) the analysis excluded monthly records within 6 months of death or censoring and (ii) cumulative statin use was assessed through the current month for event person-months and through the current month + 6 months for non-event person months.

Table S3. Estimated risk of all-cause mortality comparing statin therapy with no statin therapy,* CALIBER, 1999-2016.

		Case-control analysis		Cohort analysis	
	Cases	**Odds ratio**	**95% CI**	**Hazard ratio**	**95% CI**
Emulating a target trial†					
Intention-to-treat‡	7,072	0.87	0.79, 0.95	0.87	0.79, 0.95
Replicating the approach of a previous case-control study§					
Imposing a 3-month survival requirement from the time of selection||	7,072	0.86	0.79, 0.94	0.86	0.79, 0.94
+ Comparing ≥5 vs. <5 years of statin use¶	7,072	0.82	0.75, 0.91	0.82	0.75, 0.91
+ Adjusting for covariates instead measured at the time of selection	7,072	0.80	0.72, 0.89	0.80	0.72, 0.88

Abbreviation: CI, confidence interval.

*Among individuals with no cardiovascular disease at baseline and at an increased cardiovascular risk (defined as LDL >3.4 mmol/L). Individuals were followed for up to 10 years.

† Estimates from weighted pooled logistic regression models adjusted for age, sex, BMI, smoking status, LDL cholesterol, HDL cholesterol, months since last measure of LDL cholesterol, months since last measure of HDL cholesterol, diabetes, antihypertensive use, aspirin use, hormone replacement therapy, oral contraceptive use, number of referrals in the past three months.

‡ Comparing statin initiation vs. no initiation at baseline.

§ Estimates from unweighted pooled logistic regression models adjusted for the above covariates, assessed at baseline.

|| Comparing treatment initiation vs. no initiation at baseline. In the case-control sample, the analysis was restricted to controls alive and under follow-up 3 months after selection. In the full cohort, the analysis excluded non-event monthly records within 3 months of death or censoring.

¶ In the case-control sample: (i) the analysis was restricted to controls alive and under follow-up 3 months after selection and (ii) cumulative statin use was assessed from the time of selection (diagnosis) for cases and from the time of selection + 3 months for controls. In the full cohort: (i) the analysis excluded non-event monthly records within 3 months of death or censoring and (ii) cumulative statin use was assessed from the current month for event person-months and from the current month + 3 months for non-event person months.

Note: Here, the survival requirement only applied to controls, given that case status was defined by death. Replicating this for colorectal cancer yielded an estimated odds ratio of 0.98 (95% CI: 0.85, 1.14) comparing treatment initiation vs. no initiation, 0.69 (95% CI: 0.49, 0.97) comparing ≥5 vs. <5 years of statin use, and 0.70 (95% CI: 0.49, 0.99) adjusting for covariates measured at the time of selection.

References

1. Dickerman BA, García-Albéniz X, Logan RW, Denaxas S, Hernán MA. Avoidable flaws in observational analyses: an application to statins and cancer. *Nature Medicine.* 2019;25:1601-1606.
2. Thompson WA, Jr. On the treatment of grouped observations in life studies. *Biometrics.* 1977;33(3):463-470.
3. Hernán MA, Lanoy E, Costagliola D, Robins JM. Comparison of dynamic treatment regimes via inverse probability weighting. *Basic Clin Pharmacol Toxicol.* 2006;98(3):237-242.
4. Denaxas SC, George J, Herrett E, et al. Data resource profile: cardiovascular disease research using linked bespoke studies and electronic health records (CALIBER). *Int J Epidemiol.* 2012;41(6):1625-1638.
5. Denaxas S, Gonzalez-Izquierdo A, Direk K, et al. UK phenomics platform for developing and validating electronic health record phenotypes: CALIBER. *J Am Med Inform Assoc.* 2019;26(12):1545-1559.
6. Morley KI, Wallace J, Denaxas SC, et al. Defining disease phenotypes using national linked electronic health records: a case study of atrial fibrillation. *PLoS One.* 2014;9(11):e110900.
7. Kuan V, Denaxas S, Gonzalez-Izquierdo A, et al. A chronological map of 308 physical and mental health conditions from 4 million individuals in the English National Health Service. *The Lancet Digital Health.* 2019;1(2):e63-e77.
8. Rothman KJ, Greenland SL, Lash TL. Case-Control Studies. In: *Modern Epidemiology.* 3rd ed. Philadelphia: Lippincott Williams & Wilkins; 2008.
9. Robins JM. Choice as an alternative to control in observational studies: comment. *Stat Sci.* 1999;14(3):281-293.
10. Miettinen O. Estimability and estimation in case-referent studies. *Am J Epidemiol.* 1976;103(2):226-235.