METHODS FOR CLASSIFYING PATIENT HISTORIES IN SECONDARY HEALTHCARE DATA

Mitchell M. Conover

A dissertation submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Epidemiology in the Gillings School of Global Public Health.

Chapel Hill
2017

Approved by:

Michele Jonsson Funk

Til Stürmer

Charles Poole

Robert J Glynn

Ross J Simpson, Jr.

ABSTRACT

Mitchell M. Conover: Methods for Classifying Patient Histories in Secondary Healthcare Data
(Under the direction of Michele Jonsson Funk)

In clinical safety and effectiveness research using secondary health databases, patient medical histories are typically assessed using fixed look-back approaches. Conventional applications of these approaches exclude patients who are not continuously enrolled in the database for the entire look-back period (e.g. one year), and data occurring outside this period is ignored. An alternate approach has been suggested which assesses all of the available data history, though concerns exist that results may be biased by systematic variation in the amount of available database across important study groups.

We used applied analyses as well as plasmode simulation methods to explore the application of short (1-year) and long (3-year) fixed look-backs and all-available data approaches in analyses of Medicare fee-for-service (FFS) claims data. We assessed the bias and efficiency of effect estimates when we used the different look-backs to 1) assess cohort eligibility and to 2) identify and adjust for confounders. In the applied analysis, we evaluated the effect of statin initiation (vs. non-use) on incidence of 1) cancer within six months (a negative control outcome we expected *a priori* to be null) and 2) all-cause mortality within two years. In the plasmode simulation, exposures (conceptually: statin initiation vs. non-initiation) and outcomes (conceptually: inpatient hospitalization) were simulated as a function of self-reported interview data obtained from the Medicare Current Beneficiary Survey (MCBS, which represented the true underlying confounder of exposure-outcome associations. We evaluated estimates after applying different look-back approaches in the linked claims data.

Compared to short fixed look-back approaches, all-available approaches selected cohorts with superior classification and produced less biased estimates. Compared to long fixed look-back approaches, all-available approaches selected more inclusive cohorts and produced more precise estimates. Though these studies were conducted in a fairly narrow (applied) setting, our findings provide real-world evidence that using all-available look-backs to classify patient histories is superior to fixed look-back approaches. Our findings provide context to investigators seeking to understand the mechanisms through which the different look-backs may produce different estimates.

To my siblings, Alex, Colleen, and J.C.

and

To my parents, John and Marsha, for your support and encouragement.

ACKNOWLEDGEMENTS

I would first like to thank my adviser, Michele Jonsson Funk, for her patient mentorship and guidance. I have learned a tremendous amount from you and am fortunate to have benefited from your expertise and vision. You are a truly gifted mentor and researcher. Thank you for investing in me.

I would like to thank the members of my dissertation committee: Til Stürmer, Charles Poole, Robert J. Glynn, and Ross J. Simpson, Jr. Thank you for your encouragement and the many hours you invested in providing your guidance and expertise. I appreciate all of the opportunities I have had to collaborate with and learn from you.

I would like to thank Virginia Pate for her help and expertise in programming. I would also like to acknowledge Jennifer M. Wu, Susannah M. Bernheim, Jane Brice, and Alison McPherson, for their contributions to my education and professional development. I am very grateful for the friendships and support from my classmates in the pharmacoepidemiology program.

Finally, I would also like to express my sincere gratitude to my parents, family, and friends, whose love and support made this work possible.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ACE	Angiotensin-converting-enzyme inhibitor
ADL	Activity of daily living
ARB	Angiotensin-II receptor blocker
CABG	Coronary artery bypass graft
CI	Confidence interval
CIW	Confidence interval width
CPT	Current Procedural Terminology
DAG	Directed acyclic graph
EHR	Electronic health records
FFS	Fee-for-service
HR	Hazard ratio
ICD-9	International Classification of Diseases, Ninth Revision
IPTW	Inverse probability-of-treatment weighting
IQR	Interquartile range
MCBS	Medicare Current Beneficiary Survey
MSE	Mean-squared-error
NDC	National Drug Code
OR	Odds ratio
RCT	Randomized controlled trial
RD	Rate difference
rMSE	Root mean-squared-error
RR	Rate ratio
SD	Standard deviation
SMRW	Standardized mortality ratio weighting

SNF	Skilled nursing facility

CHAPTER 1

STATEMENT OF SPECIFIC AIMS

Clinical research has increasingly relied on secondary health data to evaluate medical therapies; however to ensure comparable accuracy of information for all subjects, current best practices require investigators to ignore a large portion of information available in these rich data sources.[1] In order to obtain relevant medical information on study subjects, longitudinal studies are routinely restricted to subjects continuously observed within the database for some uniform time period before exposure. The result is that potentially informative data occurring before this time period are typically discarded.[2] For many variables (e.g. clinical conditions, medication use, and procedures), absence of data within this period is typically interpreted as the variable itself being absent (as opposed to it being present but unobserved/missing), even if the information is available in the discarded historical data.[3,4] This may be a dubious assumption in the setting of electronic health records or administrative claims, where clinical conditions and services are only observed within specific contexts (e.g. care occurring within a certain facility or care billed to a certain payer).[5,6]

An alternate approach has been suggested which considers all available database history, regardless of whether that history is available for all patients.[4,7] Existing literature has demonstrated that both differential and non-differential misclassification of study confounders can bias effect estimates away from the null[8] and can produce spurious heterogeneity of effects across levels of the confounder.[9,10] There are many relevant research questions where investigators may expect the completeness and longitudinal breadth of available data to vary systematically between comparator groups (e.g. comparing

a new-to-market therapy to an established therapy or comparing users to non-users), threatening the validity of effect estimates obtained using all-available data approaches. While the information captured in secondary health databases does often vary between subjects, they have a number of key advantages that make them useful for clinical research: they allow large-sample studies, data can be obtained at (relatively) low-costs, and they are considered more representative of patients in routine clinical care than randomized controlled trials.[11]

Despite the widespread use of methods that clearly favor the principal of comparative information-accuracy, methodologists have debated its importance relative to other threats to validity, such as covariate misclassification, which may be reduced by using all of the available data.[4,7,12,13] Only one published study has evaluated use of all-available look-backs in real-world data.[14] It was not designed to evaluate bias and only compared to short fixed look-backs (≤ one year). To date, no research has been published exploring the use of all-available look-backs to define study eligibility criteria. Given the mounting availability of secondary health data which contains rich but heterogeneous information on large patient populations, research exploring how study subjects can be more accurately characterized using all available information is clearly needed and could inform observational clinical research across a wide range of disciplines.

Using Medicare Part A, B, and D administrative claims data, the Medicare Current Beneficiary Survey (MCBS), and hybrid simulation methods, we used all-available and fixed look-back approaches to evaluate the relative effect of statin initiation (vs. non-use) on short-term (6-month) incidence of cancer and 2-year all-cause mortality. These specific aims were intended to evaluate whether using all-available look-back methods to assess study covariates and eligibility criteria can be used to more accurately characterize study covariates and eligibility criteria and obtain less biased effect estimates compared to

conventional (fixed look-back) methods, in time-to-event cohort studies based on secondary health data.

Aim 1: Evaluate different look-back approaches to classify subjects in cohort studies using secondary health data. *Approach*: We identified a cohort of statin users and non-users at elevated cardiovascular risk using Medicare FFS administrative claims data and used fixed, all-available, and missing-data look-backs to evaluate patient histories. We compared the different look-back approaches in terms of the impact on bias and efficiency of hazard ratio estimates. We evaluated the effect of statin initiation (vs. non-use) on 6-month incidence of any cancer among older adults, a negative control outcome we expected a priori to be null. In a parallel analysis, we also evaluated the anticipated protective effect of statin initiation on secondary prevention of 2-year all-cause mortality, using two meta-analyses as alloyed gold standards.[15,16]

Aim 2: Compare performance of alternative look-back approaches for identifying covariates and exclusion criteria in a semi-simulated cohort study where the effect of interest is known. *Approach:* We conducted a plasmode simulation study in which we simulated exposures and outcomes as a function of real covariate data sampled from MCBS and linked Medicare Part A, B, and D claims.[17,18] We layered these two data sources on top of one another to reflect the structure/content of observable secondary data (claims) that is driven by the theoretically unobservable "true" disease states that underlie them (MCBS). We then determined the true bias based on the simulated relationship between exposure and outcome. We estimated rate ratios, and rate differences before and after adjustment using standardized morbidity ratio weighting.[19] To explore factors that have the greatest impact on bias

and precision, we varied simulation parameters and assumptions in a range of scenarios.

This research addresses an important gap in our understanding of widely used observational clinical research methods and seeks to maximize the utility of the wealth of secondary health data becoming available to investigators. Using the outlined approach, we developed and empirically tested novel look-back methods using real-world data. This allowed us to capture the nuances and complex interrelatedness of secondary health data, which may influence the performance of different look-back approaches. The findings of this study seek to inform better practices in clinical research and improve our ability to obtain less biased estimates using secondary health data. The data used to conduct clinical research are constantly evolving. This study represents an effort to provide investigators using secondary health data with needed information on how these data can be leveraged to obtain valid and precise answers to causal questions.

CHAPTER 2

BACKGROUND AND LITERATURE REVIEW

2.1 Secondary healthcare databases and clinical research

Ideally, the benefit and adverse event profiles of drugs, therapies, and devices can be evaluated using experimental designs (i.e. randomized controlled trials [RCTs]). However, observational research has proven to be an important source of information in a variety of cases where experimental data is not available or difficult to obtain. In any design, results may be biased by imbalance in confounding variables, which are associated with both the exposure and the outcome. While RCTs are able to control for the influence of confounding variables experimentally (e.g. by randomly assigning exposures), observational studies control for confounding using statistical adjustment.[20]

In the past few decades, there has been a marked increase in the collection, aggregation, and availability of secondary healthcare data, or data collected for non-research purposes during the routine delivery of care (e.g. administrative billing claims, electronic health records data [EHR]).[1,21] Increasing comprehensiveness and quality of secondary healthcare data is facilitated by technological advances and incentivized by multiple factors (e.g. increased patient interaction with EHR or claims-based evaluation of quality-of-care). In particular, administrative claims data have been shown to be a valid source of information to study a range of conditions and improving integrity of the data over time will likely expand their reach.[3]

Secondary healthcare databases are being increasingly used to conduct clinical research, evaluating the safety and effectiveness of drugs, procedures, and devices.[11,21-24] While RCTs are usually preferable to observational studies using secondary healthcare

data, databases studies have a number of important advantages, including: 1) their large sample size, enabling evaluation of rare outcomes; 2) their ability to reflect real-world clinical practice; 3) the relatively low-cost of data acquisition; 4) the relatively short period of time required to plan, conduct, and complete analyses; 5) the ability to conduct long-term follow-up; and 6) the increased representativeness of patient populations compared to trials.[11,21-24] Furthermore, while actively comparing products and services against existing alternatives on the market is useful to inform clinical decision-making, RCTs frequently conduct comparisons against non-treatment or placebos. Due to the relative ease of implementation, database studies are often used to conduct active (head-to-head) comparisons not addressed in the RCT literature.

However a number of challenges still face those seeking to conduct high quality clinical research using secondary healthcare databases. First, different databases comprised of the same type of data may vary substantially in the quality and completeness of the captured data, resulting in a varying ability to adequately control for confounding covariates. Research has demonstrated that nearly identical studies conducted in different secondary healthcare databases may reach different conclusions.[25] Second, the strength of secondary data sources is greatest when research interests align with the original motivation of data collection. For example, administrative claims are an excellent source of information on medical therapies and drug exposures (i.e. the items being billed) but are less reliable for information on diagnoses, especially those that aren't relevant to the billing of clinical care.[21,26] In most cases, diagnoses are recorded in administrative claims with high specificity but low sensitivity.[10] Finally, compared to studies based on other data sources, secondary healthcare database studies typically consider a wider range of covariates.[11] This can make proper model specification complicated, especially in analyses with rare outcomes.[27] This has led to increasing use of summary confounding measures such as exposure propensity scores and disease risk scores.[11]

2.2 New-user designs and look-back periods

A number of unique methods have been developed for observational studies seeking to evaluate the comparative effectiveness of medical products and services. Under the new-user design, which was popularized by Ray in 2003 and has since become the primary method for conducting comparative analyses using secondary healthcare data, medical treatment groups being compared can only be comprised of new initiators of the medication, since the inclusion of prevalent users has been shown to bias effect estimates.[28] One of the key advantages of the new-user design is that information on potentially confounding covariates is assessed only before initiation. Doing so allows delineation of the temporal relationship between confounders, exposures, and outcome, preventing investigators from adjusting for intermediates between exposure and outcome, which is known to bias estimates.[29]

As a general principal, causal research seeks to standardize measurement and ascertainment of study variables as much as possible.[30] For example, RCTs blind subjects and study personnel to exposures to ensure uniform ascertainment of cases and medical history across the exposure groups. Typically, secondary healthcare database studies evaluate information about relevant covariates in the period preceding exposure using fixed look-back periods. The study population is restricted to subjects who are continuously enrolled in the database for the entire the look-back period and any data preceding the look-back is ignored. For many important study covariates (e.g. prior diagnoses, procedures, drug exposures) absence of affirmative data (e.g. a claim or a diagnosis recorded in the EHR) during the look-back is interpreted as the covariate being absent, despite the possibility that the covariate is present but not recorded (missing). A schematic depicting a simple cohort analysis for a secondary healthcare database study is displayed in Figure 2.1.

Multiple studies have investigated the use of look-backs with varying lengths to characterize patient histories. Data-driven methods for selecting the look-back period have

been explored, however the vast majority of secondary database analyses opt for a look-back of six months, one year, or two years.[31,32] Most investigations of look-back periods focus on the look-backs function in identifying prevalent use of exposures, as opposed to the classification of covariates.[33-35] A recent study evaluated the impact that the length of the fixed look-back period had on classification of subjects as new vs. prevalent users of antibiotics and asthma medications.riis[33] They demonstrated that using short look-backs resulted in severe exposure misclassification. However, the continuous enrollment requirement imposed to observe long fixed look-backs results in highly restricted study populations which has the potential to limit the generalizability of findings.[36]

Some studies have taken the approach of using all available information to classify study covariates.[37,38] While this has the benefit of improved classification of covariates and exposure, the primary concern is that systematic variations in the likelihood of misclassification will bias results (a detailed consideration of the misclassification literature relating to this issue is included below). However, a recent simulation study conducted by Brunelli et al found that using an all-available look-back approach to classify and adjust for a study confounder led to better control of confounding in all scenarios.[4] While this finding was robust to the presence of unmeasured confounding, the results of simulation studies must be interpreted with caution, since only a subset of feasible scenarios can be considered. Recent work by Nakasian, Rassen & Franklin compared adjusted hazard ratios produced by different look-back approaches applied to studies of five different exposure-outcome pairs, conducted within a commercial insurance claims database (Optum/United).[14] They used all-available and short fixed look-backs (180 or 365 days) to ascertain and adjust for confounders. This study did not estimate the bias in estimates produced by the different approaches since it was conducted in an applied setting where the truth was not known. Neither Brunelli et al nor Nakasia, Rassan & Franklin considered the role of all-available look-backs in defining study eligibility criteria.

2.3 Covariate misclassification

Early literature on misclassification in epidemiologic studies focuses on misclassification of an exposure or an outcome. However, unexpectedly large effect estimates reported by a number of observational studies in the late 70's and early 80's raised concern that misclassification of important study confounders may also be an important source of bias.[39-46] A common theme across these studies was that they all sought to evaluate what were expected to be moderate (or null) effects, in the presence of a strong but likely misclassified confounding variable (e.g. the effect of coffee drinking on bladder cancer confounded by smoking status measured by survey self-report[42]). Early work explored the potential for confounding due to misclassification of disease severity[9,47], disease stage[48], smoking status[49], and race[50].

Before proceeding, a basic overview of some ambiguous terminology used in the misclassification literature is necessary. Misclassification refers to the presence of error in the measurement/assignment of exposures, outcomes, or covariates. Misclassification can be further categorized as either differential or non-differential as well as independent or dependent. Differential misclassification occurs when measurement error in one variable is correlated with the true value of another study variable. Dependent misclassification occurs when measurement error in one study variable is correlated with measurement error in another study variable. While truly independent and non-differential misclassification implies separation from all other study variables, these terms are frequently used to describe separation between two variables at a time (e.g. exposure misclassification relative to the outcome) or a subset of study variables (e.g. covariate misclassification relative to the exposure and the outcome). Notably, interpretations vary between literature describing misclassification of outcomes and exposures versus covariates.[9,51,52] In this proposal, use of the terms non-differential and independent misclassification will refer to complete separation from all other study variables, except where explicitly noted.

Misclassification has been represented visually using directed acyclic graphs (DAGs).[53-57] While DAGs cannot fully capture the complex influence of misclassification on causal effect estimates, they can provide some intuition about the effects of adjusting for imperfectly measured proxies of confounders. Following a framework proposed by Hernan & Cole[53] to represent exposure and outcome misclassification, Appendix 2.1 presents DAGs demonstrating non-differential, differential, independent, and dependent misclassification of a study confounder (Panels A-I). Though they do so imperfectly, these DAGs also seek to demonstrate the effects of adjustment under each misclassification scenario.

2.3.1 Non-differential and independent misclassification of a dichotomous covariate

A formal interest in evaluating the theory behind covariate misclassification began after the publication of Greenland's seminal work on non-differential confounding of dichotomous study covariates.[8] Using 2x2 tables, Greenland demonstrated how, unlike non-differential exposure and outcome misclassification, which in most cases conservatively biases estimates towards the null, non-differential covariate misclassification can bias estimates in any direction.[7] As a result, covariate misclassification is generally more likely than exposure or outcome misclassification to lead to a Type 1 error. More specifically, Greenland concluded that adjusting for a dichotomous, independent, non-differentially misclassified confounder (as displayed in Appendix 2.1, Panel A and Panel F) results in partial confounding control, a finding that has been confirmed by later studies.[8,20,58-61] Greenland's partial control finding[8] is important since it describes a scenario where adjustment for the misclassified confounder still yields a less biased estimate and is preferred to not adjusting.

Effect estimates adjusted for the misclassified confounder fall between the crude (unadjusted) estimate and the estimate fully adjusted for the perfectly classified confounder. The worse the classification scheme, the closer the adjusted estimate will be to the crude estimate. Research has demonstrated that the relative bias due to misclassification directly

relates to the misclassified covariate's strength of confounding and inversely to magnitude of the effect of interest (i.e. bias is greatest when the effect of interest is small or null and the misclassified covariate is a strong confounder).[20,49,62] Notably, Marshall & Hastrup[20] demonstrated that if the misclassified covariate is a particularly strong confounder, even a small degree of misclassification can result in highly confounded effect estimates.

There are number of papers which seek to quantify the magnitude of residual confounding described by Greenland and obtain corrected estimates.[8,49,62-66] However, implementing these methods is challenging since they require external information on the degree of misclassification that is often not available to researchers conducting epidemiologic studies using secondary healthcare data. Greenland later expanded on his earlier work, demonstrating that non-differential misclassification of a variable that is not actually a confounder may lead to the spurious appearance of confounding, and that subsequent adjustment for the non-confounder may induce bias away from the null.[67]

Relatively recently, an important condition was added to Greenland's partial-control finding[8]: non-differential misclassification of a dichotomous study confounder results in partial-control for confounding assuming there is no qualitative modification of the effect of the confounder on the outcome by exposure status (i.e. the direction of the confounder's association with the outcome does not reverse between the exposure groups).[68] This finding is important since it challenges the prevailing belief that adjustment for independent, non-differentially misclassified confounders is always preferred to no adjustment at all.

2.3.2 Non-differential and independent misclassification of a dichotomous covariate: stratum-specific estimates

Epidemiologic studies typically use inclusion and exclusion criteria to define study populations or, alternatively phrased, conduct analyses within strata of relevant covariates (e.g. males over 65 with diabetes). Stratification on any misclassified dichotomous covariate (not just confounders) has been shown to result in spurious heterogeneity between stratum-

specific estimates.[8,9,47,62] Stratification on confounding (not modifying) variables is widely considered sufficient to render an unbiased calculation of causal estimates. However, this is not necessarily the case when important study covariates are misclassified, even when misclassification is non-differential and independent.[9,47,67] In a stratified 2x2 table analysis, plausible confounding and misclassification schemes can be specified which lead to stratum-specific estimates that are more biased than the crude (Appendix 2.2). As a result, epidemiology studies that restrict study populations based on even non-differentially misclassified inclusion and exclusion criteria may be biased in any direction. Walker & Lane demonstrated that heterogeneity will be most pronounced in situations where sensitivity is high and specificity is low or where sensitivity is low and specificity is high.[9] In these instances, one strata will be extremely confounded, while the other will be nearly unconfounded.

The magnitude of bias due to spurious heterogeneity in stratum-specific effect estimates among those classified as having the confounder present depends on the specificity of confounder classification and among those without the confounder depends on the sensitivity (Appendix 2.2).[47] If we conceptualize dichotomous study inclusion criteria as the former case and dichotomous study exclusion criteria as the latter case, it follows that we would prefer specific inclusion criteria and sensitive exclusion criteria. Methods have been developed to quantify the magnitude of the bias within strata[58] and to estimate the corrected stratum-specific effect estimates.[47] However these methods require external information on the degree of misclassification that is not available during the conduct of typical studies using secondary healthcare data.

2.3.3 Differential or dependent misclassification of a dichotomous covariate

Despite the plausibility that many study covariates may be misclassified differentially or non-independently, there is relatively little literature exploring the topic. This may be due to finding that, even in the simpler case where misclassification is non-differential and

independent, effect estimates may be biased in any direction. Greenland's 1980 paper is generally interpreted to imply that the partial-control result cannot be extended to non-differential or dependent misclassification.[8] However, given that observational research can rarely assert that misclassification of confounders is fully non-differential or independent, we necessarily assume that within some tolerable degree of differentially or dependence, adjusting for misclassified confounders results in partial control for confounding. Despite this assumption, we found no direct evaluations of when adjustment for differentially or dependently misclassified confounders yields partial control for confounding. Some work, however, has explored differential misclassification. Walker & Lanes used hypothetical data to demonstrate how misclassification of a dichotomous covariate that is differential by exposure status can induce substantial spurious heterogeneity in effect estimates across strata of the covariate.[9]

Using DAGs, misclassified covariates can be conceptualized as descendants (or proxies) of true covariates (see Appendix 2.1).[53-55] When classification of the proxy is differential or dependent relative to exposure or outcome, it acts as a collider on a confounding backdoor path (Panels B-D, G-I). Adjusting for a differentially or dependently classified proxy leads to partial control of the primary confounding path (E←C→D) but un-blocks or opens other confounding paths acting through the classification mechanism.[54,55,57] The DAG clearly demonstrates that the causal estimate will be biased regardless of whether or not we adjust for differentially or dependently classified proxies. However, non-parametric graphical representations cannot establish when adjusting for the proxy will eliminate more bias (via the partially closed primary confounding path) than it introduces (via opened backdoor paths). Still, if the mechanism determining misclassification can be specified, DAGs may suggest other variables that can be adjusted for to close some (if not all) backdoor paths opened by adjusting for the proxy. Unfortunately, these determinants of misclassification are often immeasurable.

By comparing the DAGs presented in Appendix 2.1 Panels A-C, to Panels G-I, we can observe that dependent misclassification may be conceptualized as a form of differential misclassification wherein some proxy effect (e.g. the effect of a proxy of the exposure on a proxy of the outcome) is assumed to be nearly equivalent to the actual effect of interest. In both cases, adjusting for misclassified confounders opens biasing backdoor paths. Given the close relationship between dependent and differential misclassification, the scope of this study will be restricted to differential misclassification. It is likely that patterns of bias control under differential misclassification will be similar to those that would be observed under dependent misclassification. However, this will require that we assume exposures and outcomes are perfectly measured, which may represent a source of residual bias.

2.3.4 Misclassification of a polytomous or continuous covariate

In 1984, a simulation study[63] asserted that Greenland's finding[8] of partial control for confounding when adjusting for an independent, non-differentially misclassified dichotmous covariate could be extended to polytomous covariates. This assertion was not challenged until 1993, when Brenner demonstrated that that the bias induced by adjusting for an independent, non-differential, polytomous confounder can be greater than the bias induced by not adjusting for the confounder (i.e. that the partial control finding does not hold for polytomous confounders).[10]

Misclassification of continuous covariates has been considered by Marshall & Hastrup[20] and Wacholder.[12] Marshall & Hastrup evaluated non-differential, independent misclassification of a continuous confounder and found that Greenland's observation of partial control for confounding holds. However, Wacholder's paper "When measurement errors correlate with truth," explored a unique case where the misclassification of a continuous study covariate depends on the true value of that covariate (e.g. heavy smokers being more likely to under-report cigarette use). Wacholder's conclusion that Greenland's partial control finding[8] does not hold in these situations has direct applications to studies

conducted on secondary healthcare databases. For example, consider a continuous covariate defined as the total number of prescriptions filled in a patient's look-back period. It is plausible that patients with low counts could be more likely to pay for prescriptions out of pocket while patients with high counts would be more likely to rely on insurance. This would result in a greater probability of misclassification among people with low prescription counts than people with high prescription counts. Little research is available on differential or dependent misclassification of continuous covariates. Figure 2.2 displays a flowchart which aggregates literature regarding when Greenland's partial control finding can be expected to hold, for analyses adjusting for a single misclassified study covariate.

In epidemiologic studies, it is common practice to dichotomous polytomous or continuous covariates. While this practice has been shown to result in residual confounding, Greenland's partial control finding holds if the continuous covariate is correctly classified before dichotomization.[69] Interestingly, the dichotmous covariate formed by dichotomizing a non-differentially misclassified continuous variable may be differentially misclassified.[70,71] However, Gustafson & Le observed plausible scenarios where adjusting for the differentially misclassified dichotomous variable produces less biased estimates than adjusting for the non-differentially misclassified continuous parent.[72] This represents an important (though unique) case in the literature where differential misclassification is preferable to non-differential misclassification.

In complex analyses with many misclassified study covariates, some authors have proposed approaches that rely on summary measures of confounding (e.g. propensity scores).[55,66] Just as summary scores can reduce the dimensionality of confounding, they can also be used to reduce the dimensionality of misclassification. This facilitates the implementation of misclassification correction methods that don't easily scale up to analyses with highly dimensional confounding. Summary scores modeled as a function of imperfectly classified covariates can be considered a misclassified continuous covariate. As a

composite measure, differential and dependent misclassification of the component covariates will be reflected in the classification of the summary score. Furthermore, omission of important latent confounders from the model can also result in differential and dependent classification of propensity scores.

Multiple methods have been proposed to reduce propensity score misclassification. Stürmer et al proposed a method wherein the propensity score is calibrated in an external validation set where covariates and exposure are well-classified.[66,73,74] A second approach, proposed by Schneeweiss et al, uses high-dimensional propensity score adjustment and automated variable selection methods to adjust for a wide range of variables.[75] The method seeks to maximize control for true but unobserved study confounders by adjusting for many observable proxy variables.[55] Finally, Pearl proposed using external information (e.g. Bayesian priors or information drawn from external data) on the mechanisms of misclassification to construct many different "pseudo" data sets, which resemble possible manifestations of the true data. With sufficient sample size and proper specification of misclassification mechanisms, propensity scores estimated within these pseudo data can theoretically be used to obtain valid causal estimates.[55,56]

2.3.5 Simultaneous misclassification of multiple covariates

Only one study, conducted by Fewell et al. has evaluated the misclassification of multiple covariates simultaneously, studying the independent, non-differential misclassification of four continuous confounders and the resulting bias on the exposure-outcome effect estimated using logistic regression.[59] Confirming the findings of earlier research, Fewell et al observed that bias was greatest when the degree of misclassification was high and the confounder was strong. However, Fewell demonstrated that the separate biases induced by misclassifying multiple different covariates may act cumulatively on the overall bias of the exposure-outcome effect estimate. This is an important finding since it challenges the assumption that only misclassification of strong confounders can lead to

substantial bias in effect estimates.[76,77] Still, they note that when the confounders (not the error in their measurement) are correlated with one another and the exposure, the biases frequently offset each other. In typical comparative studies conducted using secondary healthcare data, it is likely that all variables are measured with some degree of misclassification, many of which may be correlated.

2.3.6 Misclassification of covariates in secondary healthcare data studies

There are a number of critical gaps in the covariate misclassification literature that limit the applicability of findings to modern clinical studies. No research has been completed evaluating: 1) the misclassification of multiple covariates used as exclusion criteria (stratification variables), or 2) the misclassification of multiple covariates, some of which are used as exclusion criteria and others in as adjustment factors. As described by Cox and Elwood, "the bias from nondifferential misclassification of several variables in a multiple regression analysis, particularly on the stratum-specific odds ratios, may also be difficult to predict."[47]

Most existing studies evaluate misclassification in the context of simplified hypothetical data (e.g. stratified 2x2 tables), which is not representative of the typical methods used in modern clinical research (e.g. propensity score methods, inverse probability of treatment weighting). Some studies have explored covariate misclassification within the context of simple logistic regression analyses, including Kupper[58], Greenland & Robins[67], Marshall & Hastrup[20], Armstrong et al[78], and Fewell et al[59]. All of these studies only considered the misclassification of one covariate, with the exception of Fewell et al[59] which considered the misclassification of four covariates simultaneously.

Using all-available information in secondary data to classify patient histories clearly risks differentially misclassifying secondary covariates and exclusion criteria, which may bias effect estimates away from the null. The decision to use fixed look-backs as opposed to all-available look-backs represents a trade-off: a reduction in the overall sensitivity of covariate

classification in exchange for non-differential misclassification of covariates by exposure and outcome status. However there are a number of plausible reasons to question the value of this trade-off in studies using secondary healthcare databases.

First, database enrollment is not the only determinant of misclassification that may be differential. For many covariates (e.g. a diagnosis or procedure of interest) classification may depend on the frequency of a patient's interaction with the healthcare system, which could plausibly vary by exposure and outcome status. Second, analyses which only require a short fixed look-back could conceivably introduce differential misclassification by outcome status. Research has shown that many potentially relevant diagnoses are unlikely to appear in claims in periods that are proximal to a patient's death.[5,6] Thus, in claims analyses evaluating all-cause mortality, we may expect substantial differential misclassification of covariates by outcome status even using fixed look-backs. In fact, if the fixed look-back only captures a short period proximal to follow-up (i.e. only the recent medical history), we may expect misclassification to be more differential by outcome status (death) using fixed look-backs than all-available look-backs. Similar misclassification patterns are possible in the time periods proximal to hospitalizations for serious clinical conditions. Finally, fixed look-backs cannot assure that misclassification will be independent of the misclassification of other variables. While fully non-differential, independent misclassification of covariates would be ideal, fixed look-backs are unable to attain this goal. Furthermore, biases due to the differential classification using all-available data histories may be offset by gains in sensitivity and specificity of classification.

Multiple methodologists have debated the importance of the principal of comparative information accuracy.[12,13,52,67] There is unanimous agreement that whenever differential and dependent misclassification can be reduced in the design phase, it should be. However, methods which induce non-differential and independent misclassification during the analysis stage (e.g. by ignoring some of the available data that is not available in the entire cohort)

are more heavily debated. This point is discussed by Wacholder: “Strict adherence to the principal of comparable accuracy used to ensure non-differential misclassification in choosing controls for case-control studies may not be advisable when it would require controls with as much error as cases instead of more accurate controls.”[12] While Wacholder was describing case-control studies, similar reasoning may be applicable to the use of fixed look-backs in cohort studies.

Even if fixed look-backs could induce fully independent and non-differential misclassification of covariates, the literature clearly demonstrates that those are not sufficient criteria to ensure bias toward the null.[10,20,49,58,59,62,68] One might argue that it is easier to anticipate and interpret the bias from independent non-differential misclassification of study covariates, since we can expect partial control for confounding and an adjusted estimate that lies between the crude and the truth. However, this interpretability does not scale up when we use multiple potentially misclassified variables, which have a cumulative effect on bias that is difficult to anticipate. Regardless of our choice of look-back, a clear understanding of the individual contributions made to the total bias by the misclassification of each covariate is likely infeasible. Research is needed which evaluates the net effect of multiple interrelated, misclassified covariates on effect estimates.

After exclusion and exclusion criteria have been implemented, observational cohort studies evaluate subject-level covariate information for two different purposes: first, to determine which of the covariates should be adjusted for in analyses, and second, to actually adjust for selected covariates (e.g. regression, weighting, matching). As pointed out by Greenland & Robins, independent, non-differential covariate misclassification can lead to spurious associations between the covariate and the outcome.[67] This may lead investigators to select instrumental variables (i.e. covariates associated with the exposure and not the outcome) for adjustment in analyses, which can substantially bias estimates.[79,80] Given these observations, it would be useful for research to consider the appropriateness of fixed

and all-available look-backs for both the selection of adjustment variables and the implementation of statistical adjustment.

2.3.7 Misclassification of covariates in time-to-event studies

The proposed research does not intend to explore misclassification in time-to-event designs. However, given that many clinical studies use time-to-event designs, we provide a brief overview here. It is important to consider the influence of not only how covariates are classified but when. Some of the earliest wisdom may be drawn from the literature on the Will Rogers Phenomenon and the spurious influence of stage migration on evaluations cancer survival. Feinstein, Sosin & Wells[48] as well as others[81,82], described a bias which results from an important study inclusion criteria (cancer stage) being identified earlier in one exposure group than the other. Increased surveillance drove earlier identification of evolving cancer stages in one comparison group, which meant the other group was more likely to be at a more clinically advanced stage of disease at study entry. This has direct application to time-to-event studies, which may seek to define inclusion and exclusion criteria using covariates that are classified differentially over time between the two comparator groups (e.g. history of acute myocardial infarction).

In a series of papers[83-85], Prentice et al developed formulas to correct for the misclassification of time-varying, normally distributed, continuous covariates in time-to-event analyses. However, implementing the methods described by Prentice require external information on mechanisms of misclassification and distributions of true covariate values. They observed that the parametric form of the function used to estimate the effect estimate influences the magnitude of bias due to covariate misclassification.

2.4 Evidence gaps

Prior research exploring use of different look-back methods has been limited in scope. The published research evaluating use of all-available look-back methods has largely been conducted in purely simulated settings. Only one paper has been published to date

exploring the use of different look-back approaches in real data, where multiple misclassified and interrelated covariates may cumulatively impact the bias of effect estimates.[14] This study did not evaluate the use of long (e.g. multi-year) fixed look-back periods. The impact of different look-backs on the bias and efficiency of effect estimates has only been assessed using simulation, which tends not to represent of the complex analytical strategies that are commonly applied in clinical research (e.g. time-to-event analyses, propensity score adjustment). To date, there is no research evaluating use of all-available look-backs to define study eligibility criteria.

Figure 2.1. Simplified schematic for a secondary healthcare database cohort study

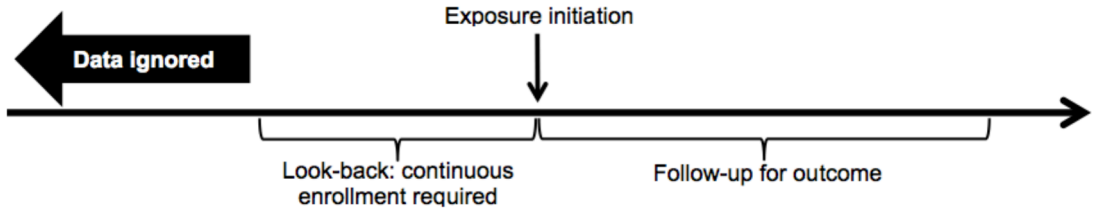
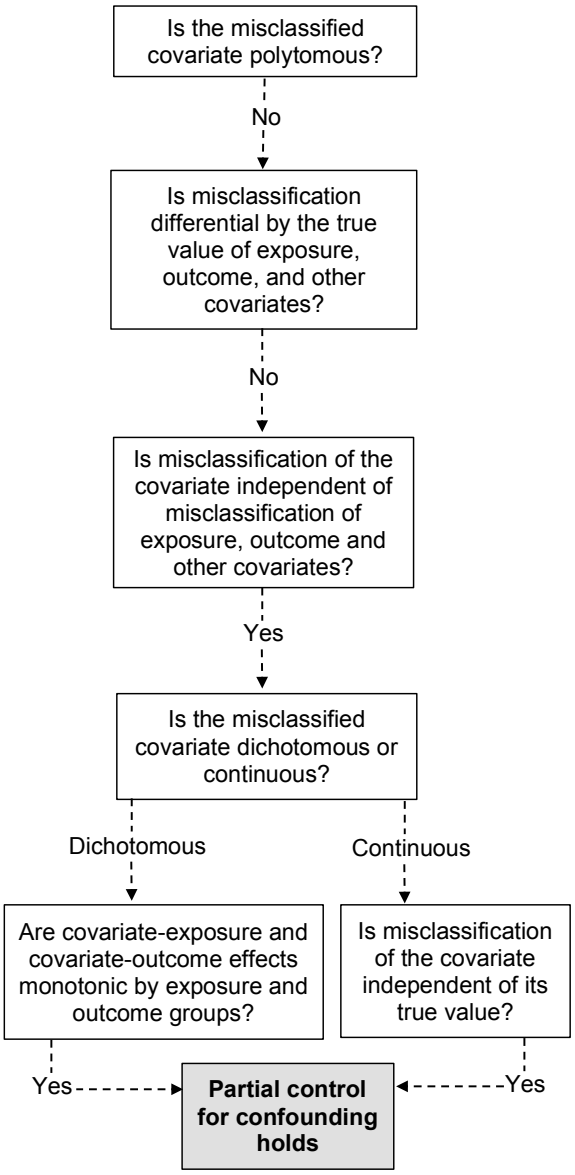


Figure 2.2. Decision flow chart demonstrating criteria for partial control in the context of a single misclassified covariate



CHAPTER 3

METHODS

3.1 Aim 1 methods

3.1.1 Overview

Using Medicare fee-for-service administrative claims data, we selected a cohort of statin users and non-users at elevated cardiovascular risk. Within this cohort, we compared variations of three different approaches for classifying patient medical histories: short and long fixed look-back periods and all-available database history. We compared these multiple approaches in terms of the impact on the bias and efficiency of hazard ratio estimates evaluating the effect of statin initiation (vs. non-initiation) on 1) 6-month incidence of any cancer among older adults, (which we expected *a priori* to be null) and 2) secondary prevention of 2-year all-cause mortality (which we expected *a priori* be protective based on meta-analyses[15,16] conducted in a similar population).

3.1.2 Study population

In Aim 1, our study population was comprised of 20% sample of fee-for-service Medicare Beneficiaries greater than 65 years of age, who were at elevated risk for cardiovascular disease. Medicare beneficiaries are an ideal population to evaluate methods for studying administrative claims data since all U.S. citizens 65 years and older are eligible for the Medicare program and generate observable claims for the routine billing of their healthcare services. An important exception is those Medicare beneficiaries who enroll in managed-care Medicare Advantage plans (Medicare Part C), who do not routinely generate observable claims for their care. As a result, our study population did not include subjects who remain enrolled in Medicare Advantage plans throughout the time-period captured by

our data. Medicare Advantage plan enrollees tend to be healthier than enrollees with traditional Medicare coverage and represent approximately a quarter of all Medicare beneficiaries.[86] However, the number of enrollees who remain continuously enrolled in Medicare Advantage throughout the observed data years (i.e. the beneficiaries who cannot enter our study) represents a smaller proportion. We were unable to observe drug claims for patients before 2007, when Medicare's drug coverage plans (Medicare Part D) were implemented. Thus, we will only evaluated data beginning in 2007.

Initially, we identified all outpatient visits occurring in the observable claims data between 2009 and 2011. Using the Medicare enrollment file, we restricted the population to visits that were preceded by a minimum of six months continuous enrollment in Medicare Parts A, B, and D. This restriction was intended to ensure that we did not include individuals in the study for whom we had too little observable claim history to be informative. Longer periods of continuous enrollment (i.e. one, two, and three years) will be required when applying the fixed and missing-data look-back approaches. We then excluded visits that had pharmaceutical claims indicating statin use in the prior six months. Furthermore, we required at least one observable pharmaceutical claim within six months before the index visit to ensure that we are actually able to Part D pharmaceutical claims. Given that we were studying a cohort of older patients with recent elevated cardiovascular risk and observed interaction with the healthcare system, it seemed reasonable to assume that most patients should have some observable medication use in their history. Failing to apply this requirement would have likely resulted in a large number of statin users being included in the study as non-users simply because we were unable to observe their pharmaceutical claims.

We included individuals at elevated cardiovascular risk by mirroring (to the best of our ability using claims-based proxy variables) eligibility criteria used by the Heart Protection Study.[87] This included patients who in the six months before their index visit have a history

of myocardial infarction, unstable or stable angina, coronary artery bypass graft, or angioplasty; stroke, transient cerebral ischemia, leg artery stenosis, carotid endarterectomy, other arterial surgery or angioplasty; or diabetes mellitus. We excluded those individuals would not be considered candidates for statin therapy, such as those with a recent history of chronic liver disease; acute or chronic kidney disease; inflammatory muscle disease; or severe heart failure. We also excluded subjects with a recent history of using cyclosporine, fibrates, or high-dose niacin. We also excluded visits which were followed by either the cancer or mortality outcome within 14 days, a period which was required to assess exposure (i.e. initiation or non-initiation of statins). The eligibility criteria for the source population are summarized in Table 3.1. The look-back approaches, which were used to further exclude visits preceded by a history of statin use (beyond the 6-month exclusion) and cancer, are described in greater detail in section 3.1.4.

3.1.3 Study design

The short-term (6-month) eligibility criteria described in Section 3.1.2 were applied uniformly for all look-back approaches. For each included outpatient visit, we used various look-back approaches to assess two additional eligibility criteria (i.e. exclusions for patients with history of statin use or the cancer outcome) and to assess covariates (Figure 3.1). After applying exclusions using each look-back, we kept each beneficiary's first eligible index outpatient visit observed in the database, within each exposure group. Thus, beneficiaries could enter the cohort up to two times, once for each exposure group.

For each included visit, we assessed whether a statin claim was observed within the subsequent 14 days. Those with statin claims during this period were classified as statin-initiators and those without as non-initiators. Follow-up for outcomes began at the end of this 14-day exposure assessment period and patients with either cancer or mortality outcomes during this period were excluded. We excluded visits followed by cancer incidence or mortality within 14 days. We evaluated a 6-month follow-up period for the cancer outcome

and a 2-year follow-up period for the mortality outcome. We estimated the time to either first occurrence of an outcome or censoring event. For both analyses, censored people who disenrolled from Medicare Part A for any reason (e.g. when a subject switches to a Medicare Advantage plan). For the cancer outcome, we also censored when patient switched to the opposite exposure group and when a patient died. The methods used to account for the competing risk of mortality are described in section 3.1.8.

3.1.4 Look-back approaches

We applied look-back periods immediately preceding the index outpatient visit to 1) assess baseline covariates and 2) exclude subjects with history of statin use or cancer. The different look-back approaches that we evaluated are described in Table 3.2. For the fixed look-back approaches, we applied 1-, 2-, and 3-year look-back periods and the cohort was restricted to those who are continuously enrolled in Part A, B, and D for this entire look-back. For the all-available look-back approach, we evaluated covariates using any of the available database history preceding the index outpatient visit and only required the 6-month continuous enrollment preceding the index visit. We also conducted a sub-analysis where we independently varied the length of the look-back period used to assess the different study components (i.e. continuous enrollment, prior statin exposure, history of cancer, and covariate assessment).

The study population varied across different look-back approaches we applied since they differed with respect to 1) the period of continuous database enrollment they required and 2) the length of history considered when excluding patients with prior statin exposure and/or cancer history. For the all-available look-back approach, only the minimum 6-month continuous enrollment was required while one, two, or three years was required for the fixed look-back approaches.

3.1.5 Exposures

For Aim 1, we defined two exposure groups: statin initiators and non-initiators. Statin use was assessed using a 14-day period following the index outpatient visit using Medicare Part D pharmaceutical claims. We also conducted a sub-analysis comparing initiators of high potency statins and initiators of low-potency statins. For the analysis assessing the effect of statin initiation on 6-month incidence of cancer, we censored patients who switched from one exposure group to the other. For the non-initiator group, we censored patients on the day they filled any prescription for statin medication. For the statin initiator group, we used the days supplied (which is recorded with each statin claim) to estimate the time period for each patient that they could be feasibly consuming medication. When a 14-day period passed which was not covered by the days supplied, we will assumed the patient stopped taking their statin and censored them (at the end of the 14 days). For analyses assessing the effect of statins on 2-year all-cause mortality, we conducted an intent-to-treat analysis only, since produces a more conservative estimate (i.e. less likely to produce a non-null estimate when none exists).

3.1.6 Outcomes

We evaluated the effect of statin use on two outcomes: one which we knew *a priori* should produce a null finding, and a second where we anticipated a protective effect. For the known null association, we evaluated time to short-term incidence of any cancer occurring within six months of statin initiation. While statins could plausibly impact long-term incidence of cancer, there exist no biologically plausible mechanism whereby statin initiation could meaningfully increase likelihood of developing a clinically diagnosable cancer within a short, 6-month period. For this reason, well-designed cancer studies typically do not classify cancer events occurring during these induction (time from drug initiation to first cancer cell in the body) and latent (time from first cancer cell in the body to a clinically diagnosable cancer) periods as outcome.[88]

While the true causal effect of statins on short-term cancer incidence is null, causal estimates drawn from administrative claims may be biased by multiple factors. Statin users and non-users may differ in their baseline cancer risk and the frequency/intensity of their interaction with the health system. Both of these factors may drive differential surveillance for outcomes between the two exposure groups. Differential surveillance may also result if the two exposure groups vary in their distribution across calendar time, since cancer surveillance typically improves over calendar time. Differential surveillance during follow-up results in outcome misclassification bias, while differential cancer surveillance during baseline (which is an exclusion criteria) results in selection bias. By assessing the effect of statins on short-term cancer incidence, we expected to observe the influence of differential cancer screening and surveillance between the two exposure groups on known null effect estimates. By evaluating the degree to which effect estimates varied from the null, we observed variation in the net impact of these biases when using different look-back approaches.

We also evaluated the effect of statins on time-to all-cause mortality within two years. While we cannot know true effect in our cohort, we *a priori* expected there to be a protective effect. Two meta-analyses have been conducted evaluating the effect of statins on 5-year mortality in an elderly cohort and reported risk ratios of 0.85 (95% CI: 0.78, 0.93)[16] and 0.78 (95% CI: 0.65, 0.89).[15] Null or harmful effect estimates served as strong indicators of biased results. While surveillance for all-cause mortality is non-differential between exposure groups, we expect the statin users to differ from non-users with respect to a number of factors that are also associated with outcomes. Because the ability to control for this confounding depends identification of confounding covariates, the magnitude of this confounding bias may vary across different look-back approaches. We expected similar confounding could possibly influence the results of causal estimates for statins and short-term cancers.

For both the short-term cancer and mortality analyses, we censored follow-up at the end of the database time or when patients disenrolled from the study database. For the cancer analysis, we also censored follow-up if a patient died or switched exposure groups. We censored follow-up among statin-initiators when they spent 14 days without medication coverage and among non-initiators when they filled a statin prescription. We considered mortality as a potential competing risk for the cancer outcome. Thus, in the short-term cancer analysis, we conducted sub-analyses accounting for the competing risk of mortality by fitting the Fine and Grey subdistribution hazards model[89,90], which is described in section 3.1.8.

3.1.7 Covariates

In Table 3.3, we present a list of candidate covariates that were considered for inclusion in adjustment models. Descriptions of the methods used to identify these covariates and rationales are described below.

Baseline covariates included demographics (age, sex, race, geographic region, calendar year of index-date), diagnoses (identified using the International Classification of Disease 9th edition [ICD-9] diagnosis codes), procedures (identified using CPT codes and ICD-9 procedure codes), and medication history (identified using NDC codes). Many of these covariates can only be imperfectly ascertained using claims data. However, it was of direct interest to explore which look-back approach is able to best identify and control for these covariates using the available data. Since imperfect measurement of covariates may impact whether they are identified as confounders and included in adjustment models, we also considered a set of variables that we *a priori* consider to be potential/likely confounders of the relationship between statin use and cancer or statin use and mortality. These include, but are not limited to obesity, hyperlipidemia, hypertension, atherosclerotic disease, cancer history, smoking status, alcohol use, healthcare utilization, and any of the covariates used to identify patients at elevated cardiovascular risk.[91-98] While the cohort will only include

patients who meet Heart Protection Study eligibility criteria[87] in the 6-month baseline, the presence of these criteria in the look-back period preceding this six months will be considered as covariates in analysis.

To evaluate intensity of healthcare utilization, we estimated the rates of outpatient visits, inpatient hospitalizations, skilled nursing facility (SNF) admissions, and unique prescription fills by dividing the measured frequency by the observed person-time in the database before the index-date. For the purposes of this study, using rates to measure these covariates is preferable to using simple counts, since for all-available look-back approaches the frequency of these covariates will depend on the amount of observable database history. Rate estimates for these covariates, on the other hand, should remain relatively stable, regardless of which look-back method is applied.

3.1.8 Analyses

We used multivariate logistic regression models to estimate the propensity score (i.e. the probability of statin initiation), conditional on all baseline covariates selected for adjustment. We then used standardized mortality ratio weighed (SMRW) Cox proportional hazards models to estimate the hazard ratio for each approach.[19] Because we used SMRW, our estimates reflect the effect of statin-use among the population of statin-users, rather than the population in general (assuming that most initiators can be matched). Since a single beneficiary in the database may enter the twice (if they have both an eligible initiation and non-initiation), we produced confidence intervals using robust variance estimators.[99]

We created two propensity score models to estimate standardized mortality ratio weights, one model for the 6-month cancer outcome and the other for the 2-year mortality outcome. For each outcome, we adjusted for the same set of variables across the different look-back approaches. Propensity score models included all variables that were risk factors for the corresponding outcome. We defined risk factors as any variable (among those described in section 3.1.7) that was 1) present in at least 1.5% of all exposed and

unexposed patients across all look-back approaches and also 2) had a significant association with the outcome with magnitude > 1.10 among unexposed patients within at least one of the look-back approaches. The associations between covariates and outcomes were quantified using treatment group specific hazard ratios, estimated using multivariate Cox models accounting for competing risk of mortality.

In the analysis evaluating the effect of statins on short-term cancer, we conducted sub-analyses in which we fit the Fine and Grey subdistribution hazards model (competing risk model), which accounts for the competing risk of mortality.[89,90] Under this approach, we effectively added follow-up time until the end of the study period for those subjects who were censored due to mortality. There are no competing risks in the analysis evaluating the mortality outcome.

In addition to comparing the estimates to one another, we also compared the hazard ratios (and 95% confidence intervals) to the expected true results. For the short-term cancer analysis, we assumed the truth to be null. For each analysis, we evaluated whether the 95% confidence interval of the estimate hazard ratio contains the null. We assessed the relative bias of the different approaches based on the distance between its point estimate of the hazard ratio and the null. We also compared the approaches in terms of variance and mean squared error (MSE). For the analysis evaluating the effect of statins on outcome mortality, we used the results of two meta-analyses as alloyed gold standards.[15,16] These meta-analyses only included elderly patients, and evaluated follow-ups of approximately five years, estimating risk ratios of 0.85 (95% CI: 0.78, 0.93)[15] and 0.78 (95% CI: 0.65, 0.89).[15] Estimated hazard ratios and 95% confidence intervals were compared qualitatively against the results of these meta-analyses. Because the true effect estimate for this analysis cannot be known with certainty, we could not estimate bias quantitatively. However, it is plausible that extremely low, protective effect estimates may be the result of healthy user bias. Furthermore, since we knew statin use has a protective effect on all-cause mortality, any

analysis with a 95% confidence interval that lies entirely above the null can be considered erroneous. Evaluating an anticipated null association (short-term cancer) in parallel provided added context to our interpretation of findings for the mortality analysis. We also compared the variance of the effect estimates produced by each look-back approach.

3.2 Aim 2 methods

3.2.1 Overview of the plasmode simulation design

We used data from the Medicare Current Beneficiary Survey (MCBS) Cost and Use module and linked Medicare fee-for-service (FFS) claims to conduct a plasmode simulation study, exploring the performance of different look-back approaches when applied to longitudinal claims data. In a plasmode simulation study, many cohorts are selected (with replacement) from a source dataset and are then layered with simulated components, in this case simulated exposures and outcomes. As shown in Figure 3.2 and Table 3.4, the exposure (conceptually: statin initiation vs. non-initiation) and outcome (conceptually: mortality) were simulated as a function of the MCBS variables, assumed to represent the true underlying confounder of the exposure-disease association. We evaluated estimates after applying different look-back approaches in the linked claims data to 1) restrict the cohort based on continuous enrollment, 2) exclude baseline statin users, and 3) assess and adjust for imperfect claims-based proxies of the true confounders measured in MCBS.

3.2.2 Selection of the source cohort from MCBS data

We selected the source cohort for the plasmode simulation by identifying all MCBS respondents who completed at least one full year of interviews (MCBS Cost and Use module) in either 2009, 2010, or 2011 (MCBS round 53 to round 61). Figure 3.3 presents the study schematic used to select the select and assess the source data from MCBS and linked claims. The MCBS Cost and Use survey is a nationally representative survey of aged and disabled Medicare beneficiaries which utilizes a rotating panel design, interviewing respondents three times per year (Winter, Summer, Fall) for up to three years. In this study,

the same individual was allowed to be included in the source cohort up to three times, once for each complete year of MCBS data available. For each of these included observations, we set the index date as the date of the Annual Health Status & Functioning (Fall) interview within the survey year. In order to imitate the type of beneficiaries that would be included in a typical claims analysis, we restricted the cohort to respondents were enrolled in Medicare Part A, B, and D at the time of their index Fall interview and who qualified for Medicare due to old age or survivors insurance. Finally, we excluded any subject who refused to answer or responded "don't know" to any survey questions used to assess MCBS covariates (described below).

3.2.3 Assessing MCBS covariates used as underling/true confounders

For each included observation, we used information collected during the Fall interview to assess demographics (age, sex [female or non-]), body-mass-index (BMI), any history of diabetes, receipt of flu-shot (last winter), and a cancer screening in the last year (defined as a mammogram or pap smear in the last year for females and a digital rectal prostate exam or a blood test for rectal cancer for males). We classified subjects as frail if during their Fall interview they reported that a health condition caused them to avoid, have difficulty, or require assistance with the following activities of daily living (ADLs): bathing, dressing, eating, walking, using the toilet, or getting in and out of a bed or chair. Using MCBS data on prescribed medicine events (which is based on a combined assessment of Part D claims and respondent self-report), we assessed use angiotensin converting enzyme (ACE) inhibitors and angiotensin receptor blockers (ARBs) in the two rounds leading up to the Fall interview.

For some variables, MCBS data was sufficiently granular to allow us to separately assess their presence in two distinct time periods: 1) in a proximal period occurring within one year before the index date and 2) in the distal period beginning at least one year before the index date. These variables included: history of high cholesterol, high blood pressure,

severe cancer (sites: lung, stomach, kidney, brain, throat, head, colon, uterus, ovary, cervix), and major adverse cardiovascular event (myocardial infarction/heart attack, congestive heart failure, angina pectoris, or stroke/brain hemorrhage).

3.2.4 Assessing true baseline use of statins

We classified each included observation as having a history of statin use if there was any evidence of statin use identified in any of the available MCBS data or claims before the index Fall interview (Figure 3.3). When applying the imperfect look-back approaches to select a cohort with no baseline statin use, we evaluated prior use within the claims data included in the look-back period.

3.2.5 Simulation

We generated 1,500 plasmode datasets by selecting 10,000 observations with replacement from the source data and then simulating the exposure and outcome.

3.2.5.1 Exposure simulation

For each observation we calculated exposure probability as a logistic function of 1) a pre-specified intercept term calibrated to produce a pre-specified exposure prevalence, 2) all "true confounders" assessed in MCBS.

$$p(exposure = 1) = \frac{1}{1 + e^{-(\beta_0 + \vec{\beta}_i * \vec{x}_i)}}$$

where:

β_0 = intercept term, calibrated to produce a pre-specified exposure prevalence

$\vec{\beta}_i$ = a vector of log-ORs for the effect of each MCBS confounder on the likelihood of exposure

$\vec{x}_i$ = a vector of confounders pulled from the MCBS data.

The log-OR corresponding to each MCBS variable were deliberately selected in order to produce unidirectional upwards confounding (making exposure appear less protective/more harmful). (Note: In scenarios where different variables are allowed to

confound in different directions, residual confounding by different variables tends to offset each other, obscuring our ability to understand the total amount of residual confounding resulting from imperfect confounder ascertainment). To simulate exposure status, we then used the calculated exposure probability of each patient to randomly sample from a Bernoulli distribution. As shown in Table 3.5, we varied the intercept term between those with and those without true baseline statin use such that those with baseline statin use were *more* likely to fill a statin after the Fall index interview.

3.2.5.2 Outcome simulation

Next, we calculated an outcome risk score as an exponential function of 1) a pre-specified intercept term calibrated to produce a pre-specified outcome prevalence, 2) exposure status, and 3) all "true confounders" assessed in MCBS.

$$RR\colon E(Y) = e^{(\beta_0 + \beta_e * exposure + \vec{\beta}_i * \vec{x}_i)}$$

$$or$$

$$RD\colon\ E(Y) = \beta_0\ +\ \beta_e * exposure\ +\ \vec{\beta}_i * \vec{x}_i$$

where:

β_0 = intercept term, calibrated to produce a pre-specified outcome incidence (number of events Y over a fixed time period)

β_e = the log-RR (or RD) for the effect of interest

$\vec{\beta}_i$ = a vector of log-RRs (RDs) for the independent effects of each MCBS confounder (*i*) on the outcome incidence

$\vec{x}_i$ = a vector of values for each confounder (*i*) pulled from the MCBS data.

For each observation, we calculated two expected number of outcomes, one corresponding to each level of exposure. The number of outcomes Y under each exposure was assigned using a random number from a Poisson distribution based on these expected values. The background rate of the outcome was simulated to be the same among true new

users and baseline statin users (Table 3.5). Thus, incorrectly including baseline statin users in the study biases estimates towards the null.

3.2.6 Using look-back approaches (applied in claims data) to assess observable-proxies of MCBS variables

Using the Medicare claims linked to the MCBS survey data, we assessed claims-based variables meant to serve as imperfect proxies for the true variables measured in MCBS (i.e. those that informed the simulation). From each of the 1,500 plasmode datasets, we selected a distinct cohort for each look-back approach which restricted to patients who had 1) the required continuous enrollment (six months for the all-available look-back, one year for the 1-year fixed, three years for the 3-year fixed) and 2) no claims-documented statin use within that look-back.

For each observation, we then assessed covariates using the corresponding look-back approach. We used Part A and B claims to identify diagnoses and care associated with diabetes and cancer, cancer screening (mammograms, pap-smears, rectal exams, and prostate specific antigen testing), and major cardiovascular events, including myocardial infarction, stable/unstable angina, heart failure, and stroke. In addition to Part A and B claims, we used durable medical equipment claims to assess claims-based frailty indicators, which included: heart failure, ambulatory life support, home oxygen, hospital bed, wheelchair, rehabilitation care, weakness or difficulty walking. We assessed Part D prescription claims to identify use of anti-hypertensives (angiotensin-II receptor agonist antiotensin converting enzyme inhibitors), beta-blockers, and anti-diabetics (biguanides, sulfonylurea, insulin, thiazolidinedione, DPP-4 inhibitors).

3.2.7 Analyses and statistics

We estimated adjusted rate ratios and rate differences using inverse-probability-of-treatment weighted (IPTW) logistic and linear regression (respectively). Inverse weights

were based on propensity scores modeling exposure as a logistic function of the covariates measured each of the different look-back approaches.

We present summary estimates using the medians from the 1,500 iterations and estimated 95% confidence intervals using the 2.5th and 97.5th percentiles of the distribution. We used confidence interval width (CIW), calculated as the difference between the log-RR confidence limits, to assess the precision of estimates produced using each approach. We assessed the bias-precision trade-off using the root-mean-squared-error (rMSE), calculated as the square root of the squared bias plus the squared Monte Carlo standard error. For each cohort selected using the different look-back approaches, we assessed the proportion of the 1,500 iterations where Walker's equipoise criterion was met (i.e. where greater than 50% of both exposed and unexposed patients have preference scores between 0.3 and 0.7).[100]

We calculated the bias by contrasting effect estimates with the known/simulated truth, which we estimated by comparing simulated counterfactuals within a cohort formed with perfect selection based on the true MCBS data on baseline statin use. The net bias in IPTW-adjusted estimates represents a summation of two component biases: 1) the residual confounding bias caused by misclassification / inadequate ascertainment of true confounders, and 2) modification bias resulting from the inclusion of subjects with statin use at baseline (who have a diminished effect of interest). We then produced approximate estimates of effects isolating each form of bias (Table 3.6). We estimated the effect isolating residual confounding bias due to misclassification by adjusting for confounding using IPTW models based on the imperfect claims, within a perfectly selected cohort selected using the true MCBS data on baseline statin use. We estimated the effect isolating bias caused by modification / inclusion of prior statin users by comparing simulated counterfactuals within an imperfectly selected cohort, based on claims evidence of baseline statin use. By

comparing each of these estimates to the known/simulated truth, we are able to the estimate magnitude and variance for each component bias.

3.2.8 Sub-analyses

We completed a range of sensitivity analyses that we thought might impact the relative performance of the look-back approaches. We explored scenarios where the effect of exposure on the outcome was protective, representing a case where the outcome is the event the exposure seeks to prevent, and also scenarios where exposure was harmful, representing a case where the outcome is an adverse-event. In the primary analysis, we allowed the impact of MCBS covariates on simulated exposures and outcomes to vary depending on whether they occurred proximally (within the last year) or distally (before the last year) relative to the index interview. We conducted two sub-analyses fixing the effect of these variables on simulated exposures and outcomes: one in which only proximal covariates had an impact and a second in which all covariates (proximal or distal) had the same impact. In the primary analyses, we did not trim the propensity score distribution or truncate weights. However, we explored results with 1% asymmetric trimming of the propensity score distribution (i.e. restricted the cohort to those with propensity scores above the 1st percentile among the exposed and below the 99th percentile among the unexposed.

In order to explore the performance all-available look-backs in scenarios with highly differential information inaccuracy, we conducted sub-analyses in which we intentionally left-truncated data histories differentially by exposure status, outcome status, and both simultaneously. For example, in one analyses, we considered all data history up to one year among initiators while considering all data up to three years of data among non-initiators.

Scenario 1 & 2: Differential information accuracy with respect to exposure status

	Scenario 1: More look-back available among initiators		*Scenario 2:* Less look-back available among initiators	
	Outcome ≥ 1	Outcome = 0	Outcome ≥ 1	Outcome = 0
Initiators	3 years	3 years	1 year	1 year
Non-initiators	1 year	1 year	3 years	3 years

Scenario 3 & 4: Differential information accuracy with respect to outcome status

	Scenario 3: More look-back available among those with outcomes		*Scenario 4:* Less look-back available among those with outcomes	
	Outcome ≥ 1	Outcome = 0	Outcome ≥ 1	Outcome = 0
Initiators	3 years	1 year	1 year	3 years
Non-initiators	3 years	1 year	1 year	3 years

Scenario 5-8: Differential information accuracy with respect to exposure and outcome status

Scenario 5 & 6: More look-back available among those with outcomes

	Scenario 5: More look-back available among initiators		*Scenario 6:* Less look-back available among initiators	
	Outcome ≥ 1	Outcome = 0	Outcome ≥ 1	Outcome = 0
Initiators	3 years	1.5 years	0.5 years	1 year
Non-initiators	1 year	0.5 years	1.5 years	3 years

Scenario 7 & 8: More look-back available among those with no outcomes

	Scenario 7: More look-back available among initiators		*Scenario 8:* Less look-back available among initiators	
	Outcome ≥ 1	Outcome = 0	Outcome ≥ 1	Outcome = 0
Initiators	1.5 years	3 years	1 year	0.5 years
Non-initiators	0.5 years	1 year	3 years	1.5 years

Table 3.1. Aim 1 inclusion and exclusion criteria, by length of look-back used to apply them

Look-back	Inclusion criteria	Exclusion criteria
Index and proximal look-back (six months before index)	1. Observable Fee-for-Service outpatient (Part B) visit between 2007 and 2013 2. ≥ 65 years or older at time of index 3. Continuously enrolled in Part A, B, and D (all-available only) 4. Heart Protection Study inclusion criteria (any of the following): – myocardial infarction, angina, CABG, angioplasty, stroke, transient cerebral ischema, leg artery stenosis, carotid endarterectomy, other arterial surgery, diabetes mellitus	1. No observable Part D pharmaceutical claim (doesn't apply to missing data approaches) 2. Heart Protection Study exclusion criteria (any of the following): – chronic liver disease, acute or chronic kidney disease, inflammatory muscle disease, severe heart failure, cyclosporine use, fibrate use, or high-dose niacin use
Entire look-back (all-available or entire fixed)	5. Continuously enrolled in Part A, B, and D before index (fixed/missing approaches only)	3. Prior statin use 4. Any cancer in history (cancer outcome only)

Table 3.2. Overview of look-back approaches as they pertain to continuous enrollment requirements, covariate assessment, and rules for missing data

	All-available look-back	**Fixed look-back**
Continuous Part A, B, D enrollment required	6 months before index	1 years, 3 years before index
Exclusions for 1) Prior statin exposure 2) Any history of cancer	All-available history before index visit	1 years, 3 years before index
Covariate assessment	All-available history before index visit	1 years, 3 years before index

Table 3.3. Candidate covariates for adjustment in Aim 1 analyses

Covariate category	Specific covariates measured
Demographics	Age, sex, race/ethnicity, geographic region, calendar year of index-date
Diagnoses	Obesity, hyperlipidemia, hypertension, atherosclerotic disease, cancer history, smoking status, alcohol use, acute myocardial infarction, angina, stroke, transient cerebral ischemia, leg artery stenosis, diabetes mellitus, chronic liver disease, kidney disease, inflammatory muscle disease
Procedures	Coronary artery bypass graft (CABG), angioplasty, carotid endarterectomy. arterial surgery, severe heart failure
Prescriptions	Cyclosporine, fibrates, high-dose niacin
Utilization	Rate of inpatient hospitalizations, rate of emergency department visits, rate of prescription fills, and rate of lipid tests

Table 3.4. Overview of the plasmode dataset

Elements of the plasmode dataset		Real data elements assessed using:	Simulated data elements are a function of:
Baseline confounders (pre-exposure)	True confounders	MCBS interview data	-
	Claims-observed proxies	MCBS-linked Medicare Part A/B/D FFS data	-
Exposure initiation (Y/N)		-	Baseline confounders (MCBS) Pre-specified covariate-exposure relationships
Outcome (Y/N)		-	Baseline confounders (MCBS) Pre-specified covariate-outcome relationships[a] Simulated exposure Pre-specified exposure-outcome relationship

[a] We used a multivariate Cox model for 1-year mortality to estimate hazard ratios (HRs) for the relationships between true confounders (MCBS) and the outcome. We then used these HR estimates to simulate the outcome.

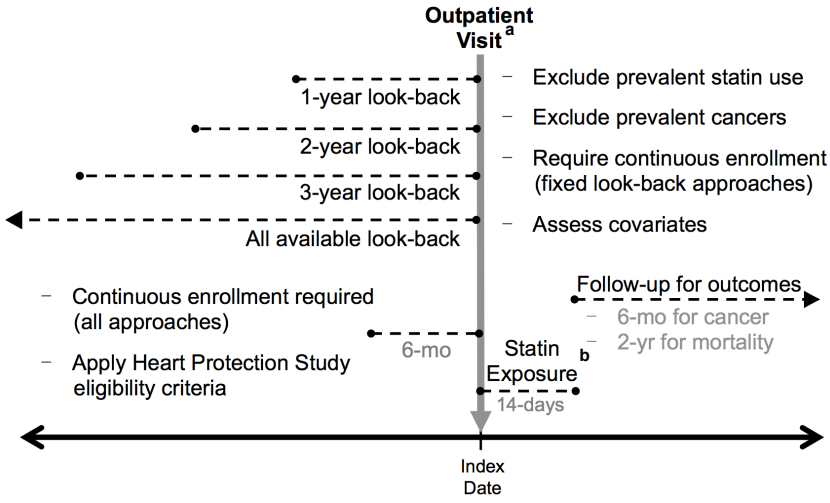
Table 3.5. Variation in the baseline exposure risk, baseline outcome risk score, and effect of interest between those with true baseline statin use and those without

True baseline statin use (assessed in MCBS)	Baseline exposure risk	Baseline outcome risk score	Effect of exposure	
			Protective	Harmful
Baseline statin users	0.33	0.5	0.4	2.5
True new users	0.66		0.8	1.25

Table 3.6. Explanation of plasmode data used for effect estimation and bias calculations

Estimate	Exclusion of prior statin users	Confounding adjustment	Bias calculation
Truth [RR_{truth}]	MCBS data	Contrast counterfactuals	NA (Unbiased)
Net [RR_{net}] Misclassification + modification	Claims-derived proxy variables	IPTW using claims-derived proxy variables	$\ln(RR_{net}) - \ln(RR_{truth})$
Residual confounding due to covariate misclassification [RR_{RC}]	"True" MCBS data	IPTW using claims-derived proxy variables	$\ln(RR_{RC}) - \ln(RR_{truth})$
Modification due to including patients with **prior exposure** [RR_{PE}]	Claims-derived proxy variables	Contrast counterfactuals	$\ln(RR_{PE}) - \ln(RR_{truth})$

Figure 3.1. Aim 1 study schematic



[a] For each person, we kept only the first eligible outpatient visit within each exposure group (i.e. the first eligible initiation visit and the first eligible non-initiation visit).
[b] We excluded any patients who had the cancer outcome or died in the 14-day exposure assessment period.

Figure 3.2. Structure of real and simulated elements in the plasmode dataset (blue arrows denote simulated relationships)

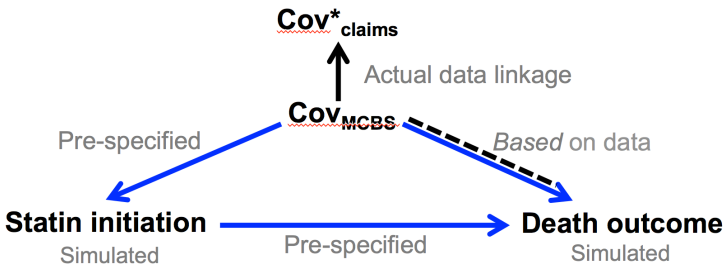
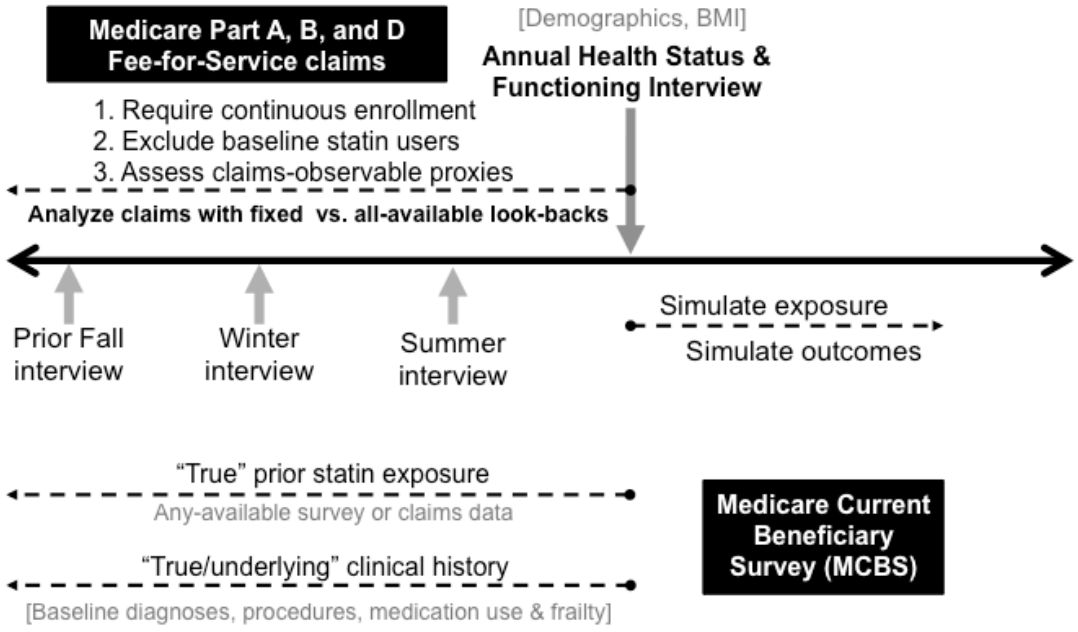


Figure 3.3. Study schematic demonstrating for selection of source data for the plasmode simulation



CHAPTER 4

CLASSIFYING MEDICAL HISTORIES IN U.S. MEDICARE BENEFICIARIES USING FIXED VS. ALL-AVAILABLE LOOK-BACK APPROACHES

4.1 Introduction

Clinical research is increasingly relying on secondary health data to evaluate the safety and effectiveness of medical therapies in real world populations.[101-103] To ensure comparable accuracy of information across comparator groups, longitudinal studies are routinely restricted to those who are continuously observed within the database for some uniform time period before exposure.[1] Potentially informative data occurring before this time period are discarded.[2] These fixed (or uniform) look-back periods are frequently used to define study eligibility criteria (e.g., no observed history of exposures or outcomes, no recent cardiovascular events) and also to capture baseline covariates used to adjust for confounding.

Selecting a fixed look-back period requires investigators to weigh competing priorities. A longer period allows for a more thorough characterization of database enrollees but also selects narrower, smaller cohorts. In many cases, at least in the US, database enrollment depends on a range of complex variables (e.g. employment, socioeconomic status, marital status / family structure, health status, age). It is unclear whether enrollment restrictions, which inadvertently condition on these characteristics, might impact findings. Despite widespread use of methods that clearly favor the principal of comparative information-accuracy in epidemiology, methodologists have debated its importance relative to other threats to validity, such as covariate misclassification or selection bias, which may be reduced by using all of the available data.[4,7,12,13,104] Observing all historical (pre-exposure)

information available in a database while requiring only minimal baseline continuous enrollment has been proposed as a possible compromise which might improve capture of relevant medical history and selection of more inclusive, representative cohorts.[4,7] The common argument against using all-available look-backs is that, for many research questions, we might expect the completeness and longitudinal breadth of available data to vary informatively between exposure (e.g. when comparing users to non-users) or outcome groups, threatening validity of estimates.

To date, there has been limited research exploring the use of all-available data to characterize patient medical histories, primarily using simulations of simplified scenarios.[4,104] Only one paper has been published exploring use of all-available look-backs in actual data with multiple interrelated covariates but not addressing the issue of cohort selection.[14] Thus, we sought to evaluate the application of multiple look-back approaches to select patients and classify covariates in an observational cohort study set in the Medicare claims database. In this study, we estimate the effects of statin initiation (compared to non-initiation) after an outpatient office visit on 1) a null outcome (6-month cancer incidence) and 2) a protective outcome (2-year all-cause mortality).

4.2 Methods

4.2.1 Study population

We used a 20% random sample of Medicare fee-for-service beneficiaries with at least 1 month concomitant parts A, B, and D coverage, to identify all outpatient visits observed between 2007 to 2012 when the patient could have received a new statin prescription. For all look-back approaches, we required a minimum of six months of continuous Part A, B, and D enrollment before the potential index visit (see exposure below) and at least one Part D claim within this period. During the six months preceding the index visit, patients were required to have a diagnosis or procedure code indicative of elevated cardiovascular risk and no medications or diagnosis codes indicative of strong

contraindications for statin therapy. These eligibility criteria were meant to imitate those of the Heart Protection Study.[87]

We identified three cohorts by applying different look-back periods to the set of potential index visits identified using the 6-month period above. For the all-available database history approach, we required no additional continuous enrollment, but excluded all visits preceded by any observable statin claims or cancer (other than non-melanoma skin cancer) diagnosis/treatment. When applying the conventional one- or 3-year fixed look-back periods, we further restricted the cohort to those continuously enrolled throughout the entire look-back and then excluded visits with prevalent statin use or cancer history within these look-back periods. When beneficiaries had multiple eligible outpatient visits, we selected the first eligible visit within each exposure group (i.e. the first eligible initiation visit and the first eligible non-initiation visit). A study schematic illustrating the overall study design is presented in Figure 4.1.

4.2.2 Exposure

We classified each index outpatient visit as either a statin initiation or non-initiation by evaluating whether there was a claim for a statin dispensing at a pharmacy in the subsequent 14 days.

4.2.3 Outcomes and follow-up

In separate analyses, we evaluated the effect of statin initiation on two outcomes 1) incident cancer within six months and 2) all-cause mortality within two years. For both, follow-up began on the day after the 14-day exposure assessment window (15 days after the index outpatient visit). Individuals with either outcome during this 14-day window (≈0.4% of visits) were excluded. For both outcomes, we censored follow-up when individuals disenrolled from the study database or the end of available data, December 31, 2012. For the short-term cancer outcome, we also censored follow-up when patients died or switched

exposures. Exposure switching was defined as a statin fill for non-initiators and 14 days without medication coverage for initiators.

4.2.4 Covariates

We used the index visit claim to assess information on patient demographics (age, sex, race, geographic region, and calendar year). Then, using the various look-back approaches, we assessed historical claims to classify baseline health behaviors, diagnoses and procedures using CPT, HCPCS and ICD-9 codes associated with Part A and B claims and baseline medication use using NDC codes associated with Part D claims. We described utilization variables as rates (e.g. # outpatient visits per month).

4.2.5 Statistical analyses

Within each cohort, we evaluated covariate imbalance between initiators and non-initiators using the average standardized mean difference[105] and then used multivariate logistic regression to estimate a propensity score (i.e. baseline probability of statin initiation conditional on baseline covariates)[106] corresponding to each index visit in the cohort. Propensity score models included all variables that were identified as risk factors for the outcome using any look-back approach. A more detailed description of the approach to variable selection for the propensity score model is available in Appendix 4.2 and the sets of selected variables for each outcome are given in the footnote of Table 4.1.

In each analysis, we estimated crude and adjusted hazard ratios for the effect of interest using Cox proportional hazards models. We used the robust variance to estimate confidence intervals to account for beneficiaries who entered the cohort twice (for an initiation and non-initiation).[99] We adjusted estimates to account for differences in measured baseline covariates using standardized mortality ratio weighting (SMRW) with and without 1% asymmetric trimming of the propensity score.[19,66,107] In a sub-analysis of the cancer outcome, we accounted for competing risk of mortality by fitting the Fine and Grey subdistribution hazards model.[89,90] We used the cumulative hazard function to plot

cumulative incidence curves estimates of the risk difference (i.e. the difference in cumulative incidence at each point in time) over the course of follow-up.

For the 6-month cancer analysis, we anticipated a null effect, since it is implausible for any statin exposure to have a causal effect on the incidence of clinically-detectable cancer within such a short interval after initiation.[108] While this effect should be null, we expected estimates to be biased by uncontrolled differences in selection, baseline cancer risk and cancer surveillance during follow-up. Thus, we estimated mean squared error (MSE) using the equation: $MSE = (1 - \text{log-HR})^2 + (\text{Standard Error}_{\text{log-HR}})^2$. For the analysis evaluating the effect of statins on mortality, the results of two meta-analyses served as alloyed gold standards.[15,16]

4.2.6 Sub-analyses

Unlike the primary analysis, which applied the same look-back uniformly for all study components (e.g. exclusion of prevalent statin users, assessing confounders for adjustment), we conducted a sub-analysis varying each component individually and holding the others fixed. This allowed us a more granular exploration of the mechanisms through which look-backs might alter findings. We also conducted a sub-analysis with an active comparator, i.e. high-potency statins vs. low-potency statins.

This study was reviewed and approved by University of North Carolina's institutional review board (study: 16-1066). All analyses were conducted in SAS 9.4 (SAS Institute, Cary, NC, USA) and figures were produced using SAS 9.4 or R 3.3.1 (R Foundation for Statistical Computing; Vienna, Austria).

4.3 Results

The all-available cohort (71,347 initiators, 476,832 non-initiators) was slightly smaller than the 1-year fixed cohort (86,923 initiators, 559,471 non-initiators) and much larger than the 3-year fixed cohort (18,918 initiators, 204,249 non-initiators) (Table 4.1). As implemented here, the all-available look-back had a far less restrictive continuous

enrollment requirement compared to the 1-year look-back. However, the all-available cohort was smaller than the one year because it excluded more patients with identifiable history of statin use and/or cancer (Figure 4.2). With respect to the proportions of patients excluded for having prior statin use and cancer history, the all-available approach was less restrictive than the 3-year approach, but much more restrictive than the 1-year approach (Figure A4.3.1). Among non-initiators, cancer incidence during follow-up was elevated in cohorts selected using shorter fixed look-backs (1-year: 2.0% vs. 3-year: 1.5%). Cancer incidence in the all-available cohort most closely resembled that of the 3-year fixed cohort. For all look-backs, the inclusion criteria for recently elevated cardiovascular risk was most frequently met by the presence of either diabetes or stroke.

In the all-available cohort, non-initiators had less available Part A/B history (median: 23 months, IQR: 19-38) compared to initiators (median: 31 months, IQR: 21-47) (Figure A4.3.2). The same was also true for Part D database enrollment history among non-initiators (median: 20 months, IQR: 14-30) and initiators (median: 27 months, IQR: 18-41). The amount of available database history was nearly identical across levels of both the cancer and mortality outcomes.

In Figure 4.3 we present the proportion of the cohort with observable history of statin claims (Figure 4.3a) or cancer (Figure 4.3b) when all-available data was considered, stratified by the calendar year of the index visit. The corresponding figure for the active comparator sub-analyses is available in Figure A4.3.3. Compared to non-initiators, initiators in the 1-year look-back cohort were more likely to have identifiable history of statin use; however, in the 3-year look-back cohort, the two groups were similar. In the 1-year look-back cohort, 46% and 30% of initiators and non-initiators (respectively) had identifiable baseline statin use when all available data was considered. For both fixed look-back approaches, non-initiators were more likely to have identifiable cancer history than initiators. Misclassification was less frequent in the cohorts selected using longer look-backs. Due to

the left-truncation of the Medicare data in calendar time (in 2007), the all-available approach was less informative in earlier calendar years (i.e. since in earlier calendar years less data history was available).

It is important to note that most beneficiaries who entered the study twice entered the study as a non-initiator prior to entering as an initiator. The proportion of initiators who had a dual-entry in the cohort as a non-initiator did not vary widely by look-back approach, ranging from 70% of initiators for the 3-year approach to 75% for the 1-year (Table A4.1.1).

Compared to non-initiators, initiators were younger, used more preventive health services / screening, and were more likely to be diabetic (Table A4.1.2). Broadly speaking, the all-available approach tended to identify greater imbalance in measured covariates compared to fixed look-back approaches, although in most cases not by much (Figure 4.4,). For all look-back approaches, covariates were well balanced (standardized difference <5%) after SMR-weighting. Propensity score distributions under each look-back approach are presented in Figure A4.3.4 and Figure A4.3.5.

In analyses of the 6-month cancer outcome, SMRW-adjusted estimates of the hazard ratio generated using fixed look-backs ranged from 0.79 (95% CI: 0.73-0.84, MSE: 1.54) for the 1-year to 1.05 (95% CI: 0.90-1.21, MSE: 0.92) for the 3-year fixed look-back (Table 4.1). The SMRW-adjusted HR estimate for the all-available approach (HR: 0.90, 95% CI: 0.83-0.98, MSE: 1.22) was more biased than the 3-year approach but more precise. In the 6-month cancer analysis, SMRW-adjustment had little impact on estimates, especially in the case of the 1-year look-back.

For the outcome of 2-year all-cause mortality, we observed substantial confounding in the crude estimates (Table 4.1). Crude HR estimates were very similar between the look-backs, spanning from 0.47 to 0.50. Point estimates of the HR were similar for all look-back approaches after applying SMRW adjustment. The adjusted estimate produced by the all-available approach (HR: 0.77, 95% CI: 0.74-0.80) was similar to the estimate produced by

the 3-year fixed look-back (HR: 0.82, 95% CI: 0.76-0.88), but was more precise. All results were consistent after 1% asymmetric propensity score trimming (data not shown). Results from the active comparator sub-analyses are presented in Table A4.1.3.

In the sub-analysis independently varying the look-back to define different study components, estimates were generally insensitive to look-back choice (Table 4.1). An important exception is that in the 6-month cancer analysis, estimates dramatically (and significantly) improved when we excluded patients with prior cancer history using the all-available approach (HR=0.94, 95% CI: 0.84,1.04) or 3-year fixed look-back (HR=0.93, 95% CI: 0.84, 1.03) instead of a short 6-month look-back (HR=0.69, 95% CI: 0.65, 0.74). In the 2-year mortality analysis, estimates were most sensitive to the choice of look-back used to exclude prevalent statin users. Using all-available or longer fixed look-backs moved estimates towards the null and increased the observed mortality in the cohort. Independent variation in the continuous enrollment requirement and assessment of confounders (for adjustment in propensity scores) resulted in negligible movement in estimates.

In Figure 4.5, we present cumulative estimates of the risk difference over the course of the 6-month follow-up for each look-back approach. (The corresponding cumulative incidence curves are available in Figure A4.3.6 and Figure A4.3.7). Risk differences estimated using all-available and 3-year fixed look-backs were generally closer to the presumed truth (null) than the estimates produced using 1-year fixed look-backs. Throughout most of follow-up, the adjusted 3-year look-back estimate is the closest to the true null though, by the end of follow-up, the magnitude of the bias in the all-available estimate was comparable. The results of the short-term cancer analysis accounting for the competing risk of mortality were identical to the primary analysis (data not shown). Figure 4.6 presents the cumulative risk difference estimates for the 2-year mortality analysis. Throughout follow-up, estimates produced by the different look-back approaches overlapped one another nearly perfectly.

4.4 Discussion

For the effects explored in these analyses, differences in estimates produced using all-available and 3-year fixed look-backs were small, with substantial overlap in confidence intervals (Table 4.1). Point estimates produced by the 3-year look-back were slightly less biased than the all-available approach, but less precise. In claims studies, bias is typically of greater concern than precision. However, it is still necessary to understand trade-offs in bias and precision, since their relative importance will depend on the specific study question and population. Generally speaking, the all-available approach tracked closely with the 3-year look-back in sub-analyses where we independently varied specific look-back components (holding the others fixed).

Two meta-analyses evaluating the effect of statin use (vs. non-use) on 5-year mortality among elderly patients with established cardiovascular risk estimated risk ratios of 0.85 (95% CI: 0.78, 0.93)[16] and 0.78 (95% CI: 0.65, 0.89).[15] After SMRW-adjustment and trimming, all look-back approaches produced point estimates for 2-year mortality HR that fell in the plausible range between the point estimates for the risk ratios estimated by these meta-analyses. Two randomized double-blinded trials evaluating effects over shorter follow-up (two[109] and three[110] years) produced estimates of 0.76 (95% CI: 0.51, 1.00) and 0.75 (95% CI: 0.49, 0.99), respectively. Trial estimates may provide a reasonable benchmark. However, we cannot use them to assess the bias of the estimates produced in our study since we are evaluating statin effectiveness, not efficacy, in a broader, more heterogeneous population than was evaluated in the trials.

In the analyses we present, there were four key aspects of the cohort that were affected by the look-back period (Table 4.1 presents results of individually varying each component): the continuous enrollment requirement, exclusion of prevalent statin users, exclusion of patients with a history of the cancer outcome, and assessment of confounders. We discuss the way in which the look-back approaches affected each of these in turn.

4.4.1 Imposing continuous enrollment requirements

We compared statin initiators and non-initiators because it seemed especially plausible that these exposure groups would exhibit striking differences in the accuracy/availability of database information (e.g., as a function of health services utilization and available database history). Indeed, due to our design, we observed less database history among non-initiators, with the median Part A/B look-back being about 8 months shorter among non-initiators. We did not observe meaningful variation in available database history with respect to either the cancer or the mortality outcome. In sub-analyses, independently varying the continuous enrollment requirement had little impact on crude or adjusted effect estimates (Table 4.1).

4.4.2 Excluding prevalent statin users

Proper exclusion of prevalent statin use is necessary to correctly align time at risk after true initiation. A substantial proportion of cohorts selected using short fixed look-backs had identifiable prior statin use when all available data was considered. Unrecognized prior statin exposure appeared non-differential when using a longer fixed look-back but was more common among initiators when using a short fixed look-back. This may indicate that short fixed look-backs are prone to including prevalent users (e.g. patients paying out-of-pocket, recent/short-term discontinuers). Presumably, these patients were identified and excluded by the longer 3-year look-back. Independently varying the look-back for excluding prevalent statin users produced changes in estimate in the 2-year mortality analysis but not the 6-month cancer analysis (since the true effect in the cancer analysis is null) (Table 4.1).

4.4.3 Excluding prevalent cancer cases

Considering all-available data, the short 1-year look-back cohort incorrectly included 18% and 23% of initiators and non-initiators (respectively) who had observable cancer history in the database (Figure 4.3b). A possible explanation for why initiators had less unidentified cancer history might be that they were younger and that approximately 70% of

initiators entered the cohorts as non-initiators prior to entering as initiators. It may also be driven by differential surveillance. Initiators were more likely to have undergone cancer and other health screenings. Initiators' superior cancer surveillance *within* the fixed look-back period may reduce the number of unrecognized cancers in the cohort that can be reclassified using data *outside* the look-back period. Failing to properly exclude patients with observable cancer history in the database is more likely to bias estimates of the effect of statins on short-term cancers, where the truth is known to be null. We observed this in the sub-analysis independently varying exclusion for patients with a history of the cancer outcome, producing meaningful improvements in estimates when using longer look-backs (e.g. 3-year or all-available approaches) to exclude these patients (Table 4.1). This is the most plausible explanation for why the all-available and 3-year fixed analyses of the short-term cancer outcome produced less biased estimates than the 1-year fixed look-back.

4.4.4 Assessment and control for confounding

To informally evaluate the impact of different look-backs on identifying and adjusting for confounding, we can observe change in crude estimates after SMRW adjustment. Unfortunately, in the evaluation of the short-term cancer outcome, the only analysis where we can reasonably estimate bias and MSE, SMRW adjustment had a nearly negligible impact on estimates (Table 4.1). However, in the mortality analysis, where SMRW adjustment produced large changes in estimates indicating a more prominent role of measurable confounding, we observed substantial overlap in estimates before and after adjustment. This may indicate that the information obtained from more distal database history captured by longer look-backs is of limited use. This finding is consistent with the findings of Nakasian et al. who compared short fixed look-backs to all-available approaches in an analysis of a commercial claims database.[14] In the sub-analysis independently varying the look-back used to assess confounders, the all-available (HR=0.79) and 3-year (0.80)

look-back estimates for 2-year mortality were only slightly lower compared to those produce by shorter fixed look-backs (1-year: HR=0.83).

This study has some important limitations. Since this paper explores an applied example in real-world data, it is difficult to know the truth or evaluate true bias as earlier simulation work has. Single empirical examples have, however, previously been successfully used to compare different study designs.[111] Also, it is likely that analyses of the short-term cancer outcome remains confounded by variables that we could not measure in the Medicare data. Minimal change in the cancer estimates before and after adjustment indicates a limited ability to control for confounding when using claims data. However, in analyses of the mortality outcome, where SMRW adjustment resulted in substantial changes in estimates, all look-backs produced similar estimates. Furthermore, we selected a population with recently-observed elevated cardiovascular risk in order to assure that everyone would have a plausible indication for statin therapy. However, it is possible that our estimates remain confounded factors that we measure within the claims data, which may lead a physician to withhold statins from an otherwise indicated patient (e.g. frailty). Our design allowed the same patient to enter as both a statin initiator and non-initiator, and the great majority who did entered first as a non-initiator, i.e., with less available look-back. It is unlikely this impacted the relative performance of the different look-backs since the frequency of repeated patients in the cohort did not vary widely by look-back approach. Furthermore, we adjusted estimates using SMRW (which weights to the treated population), preventing us from double-counting patients who were eligible to enter the cohort in both exposure groups, since they can only appear once as an initiator. Finally, determinants of continuous enrollment, and thus performance of different look-back methods, may vary across different study questions, populations, and databases, which may limit the generalizability of our findings.

Further research exploring these approaches is needed. Formal quantitative bias analysis may be a promising method to explore (and/or bound) the impact that differential database history might have on the performance of different look-backs.[112] Our decision to select each beneficiary's first eligible visit may reduce the benefit of using all-available database information and potentially increases differential information accuracy by exposure status. Our motivation for using this approach was to provide a conservative evaluation of all-available look-backs in a potentially problematic setting. However, further research is needed exploring the performance of different look-back approaches when using alternative cohort selection strategies (e.g. randomly sampling across person-time). Our study design and choice of comparators prevented us from doing so here.

This applied example contributes further evidence to the growing case for using all-available look-backs to characterize patients in longitudinal database studies, particularly when the alternative option is to use a short fixed look-back (e.g. due to the statistical power required to estimate effects or the structure of the database). The case for all-available look-backs is made stronger by the fact that the comparability of information accuracy in study groups being compared can be empirically evaluated (e.g. the amount of available baseline data, or the frequency of healthcare interactions), at least to some degree. The look-backs did not appear to vary substantially with respect to their ability to control for confounding. However, selecting a study population using all-available look-backs produced a cohort with less prevalent exposure and cancer reducing bias in analyses where exclusion of patients with prior cancers was essential. By not requiring long periods of continuous enrollment, cohorts selected using the all-available approach were broader and more clearly defined than cohorts selected using fixed look-backs, enhancing the precision and generalizability of estimates.

Table 4.1. Cohort sizes, outcome frequencies, and hazard ratios (crude and SMRW-adjusted) for primary analyses uniformly applying the same look-back approach for all components and sub-analyses varying the look-back of each component individually

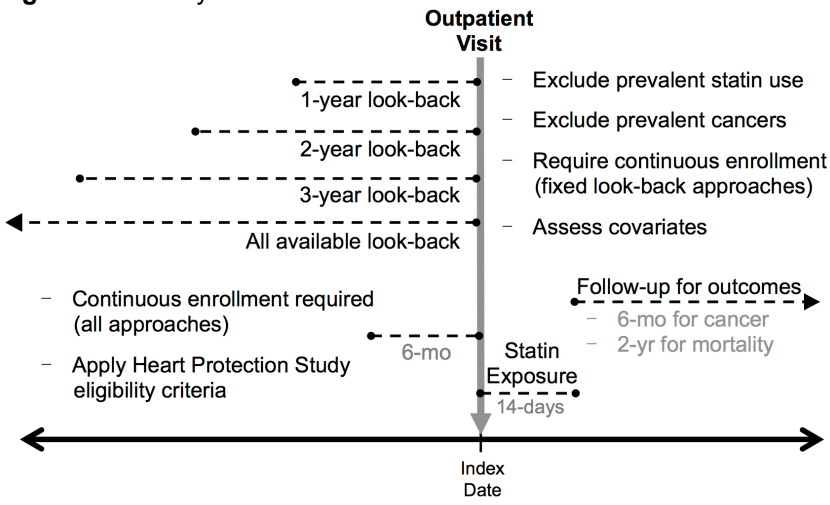
	Look-back parameters				Cohort size (N)		Outcome frequency		Hazard ratio (95% CI)			
	Eligibility criteria			Model					6-month cancer		2-year mortality	
	Con Enr.	BL statin	Can Hist	PS vars[a]	N_{Total}[b]	N_{Statin}[b]	$\%_{Can}$	$\%_{Death}$	Crude	SMRW	Crude	SMRW
Primary results												
1-year fixed	1yr	1yr	1yr	1yr	646,394	86,923	1.8%	8.4%	0.78 (0.73, 0.84)	0.79 (0.73, 0.84)	0.48 (0.47, 0.50)	0.79 (0.76, 0.82)
3-year fixed	3yr	3yr	3yr	3yr	223,167	18,918	1.4%	8.5%	1.00 (0.87, 1.16)	1.05 (0.90, 1.21)	0.50 (0.46, 0.53)	0.82 (0.76, 0.88)
All-available	AA	AA	AA	AA	548,179	71,347	1.5%	8.0%	0.85 (0.79, 0.92)	0.90 (0.83, 0.98)	0.47 (0.45, 0.49)	0.77 (0.74, 0.80)
Varying look-back component												
Continuous enrollment requirement	6mo	6mo	6mo	6mo	952,296	163,184	2.7%	7.8%	0.68 (0.66, 0.71)	0.64 (0.61, 0.67)	0.49 (0.48, 0.50)	0.80 (0.78, 0.82)
	1yr	6mo	6mo	6mo	817,987	137,984	2.7%	7.9%	0.69 (0.66, 0.72)	0.65 (0.62, 0.68)	0.50 (0.48, 0.51)	0.80 (0.78, 0.82)
	3yr	6mo	6mo	6mo	440,427	64,604	2.8%	7.8%	0.72 (0.67, 0.77)	0.69 (0.65, 0.74)	0.51 (0.49, 0.53)	0.82 (0.79, 0.85)
Baseline statin use	3yr	6mo	6mo	6mo	440,427	64,604	2.8%	7.8%	0.72 (0.67, 0.77)	0.69 (0.65, 0.74)	0.51 (0.49, 0.53)	0.82 (0.79, 0.85)
	3yr	1yr	6mo	6mo	372,173	40,687	2.9%	8.3%	0.70 (0.64, 0.76)	0.67 (0.61, 0.72)	0.51 (0.48, 0.53)	0.83 (0.79, 0.87)
	3yr	3yr	6mo	6mo	288,687	23,293	3.1%	8.8%	0.73 (0.66, 0.81)	0.70 (0.63, 0.78)	0.51 (0.48, 0.54)	0.85 (0.79, 0.90)
	3yr	AA	6mo	6mo	255,267	19,779	3.1%	9.1%	0.72 (0.65, 0.80)	0.67 (0.60, 0.75)	0.51 (0.48, 0.54)	0.86 (0.80, 0.92)
Cancer history	3yr	6mo	6mo	6mo	440,427	64,604	2.8%	7.8%	0.72 (0.67, 0.77)	0.69 (0.65, 0.74)	0.51 (0.49, 0.53)	0.82 (0.79, 0.85)
	3yr	6mo	1yr	6mo	404,030	59,987	1.8%	7.8%	0.80 (0.74, 0.86)	0.82 (0.76, 0.89)	0.51 (0.49, 0.53)	0.83 (0.79, 0.86)
	3yr	6mo	3yr	6mo	340,814	51,929	1.3%	7.5%	0.91 (0.83, 1.01)	0.93 (0.84, 1.03)	0.51 (0.48, 0.53)	0.83 (0.79, 0.86)
	3yr	6mo	AA	6mo	300,628	47,042	1.2%	7.5%	0.92 (0.83, 1.02)	0.94 (0.84, 1.04)	0.51 (0.49, 0.54)	0.83 (0.79, 0.87)
Propensity-score variables	3yr	6mo	6mo	6mo	440,427	64,604	2.8%	7.8%	0.72 (0.67, 0.77)	0.69 (0.65, 0.74)	0.51 (0.49, 0.53)	0.82 (0.79, 0.85)
	3yr	6mo	6mo	1yr	440,427	64,604	2.8%	7.8%	0.72 (0.67, 0.77)	0.68 (0.63, 0.72)	0.51 (0.49, 0.53)	0.83 (0.79, 0.86)
	3yr	6mo	6mo	3yr	440,427	64,604	2.8%	7.8%	0.72 (0.67, 0.77)	0.67 (0.63, 0.72)	0.51 (0.49, 0.53)	0.80 (0.77, 0.83)
	3yr	6mo	6mo	AA	440,427	64,604	2.8%	7.8%	0.72 (0.67, 0.77)	0.66 (0.62, 0.71)	0.51 (0.49, 0.53)	0.79 (0.76, 0.83)

[a] *Variables included in propensity score (PS) models for both the 6-month cancer analysis and the 2-year mortality analysis:* sex, age (as a continuous linear term, continuous squared term, categorical term with 5-year categories), calendar year, race, inpatient stays/month (continuous linear term and categorical term divided by quintile), outpatient visits/month, skilled nursing facility admissions/month, unique drugs/month, smoking, substance abuse, anemia, COPD, dementia, hyperlipidemia, venous thromboembolism, cancer screening, cardiac stress test, colonoscopy, hs-CRP, sulfonylurea, insulin, home oxygen. *Variables only*

included in PS models for the 6-month cancer analysis: inclusion for diabetes (≤6-months), diabetes (>6-months), stroke (>6-months), chronic liver disease (>six months), arthritis, rheumatoid arthritis, gastrointestinal bleed, PSA testing, creatinine. *Variables only included in PS models for the 2-year mortality analysis:* inclusion for stroke (≤6-months), chronic kidney disease (> 6-months), obesity, angiography, pulmonary circulation disorders, peripheral vascular disease, osteoarthritis, asthma, atrial fibrillation, psychiatric disorder, inflammatory bowel, paralysis, sepsis, vertigo, lipid panel, echocardiograph, fecal occult blood testing, ARB, diuretics, thiazide, ambulatory life support, weakness, wheelchair.

[b] These counts denote unique observations in the dataset. Patients who enter the cohort twice for eligible initiations and non-initiations are counted twice in the N_{total} statistic (one for each exposure). Since they cannot appear twice in the same exposure group, the N_{statin} statistic denotes counts of unique patients.

Figure 4.1. Study schematic



[a] For each person, we kept only the first eligible outpatient visit within each exposure group (i.e. the first eligible initiation visit and the first eligible non-initiation visit).
[b] We excluded any patients who had the cancer outcome or died in the 14-day exposure assessment period.

Figure 4.2. Bar chart showing the proportion of distinct beneficiaries excluded for each of three eligibility criteria applied using different look-back approaches and the final proportion included in final cohorts

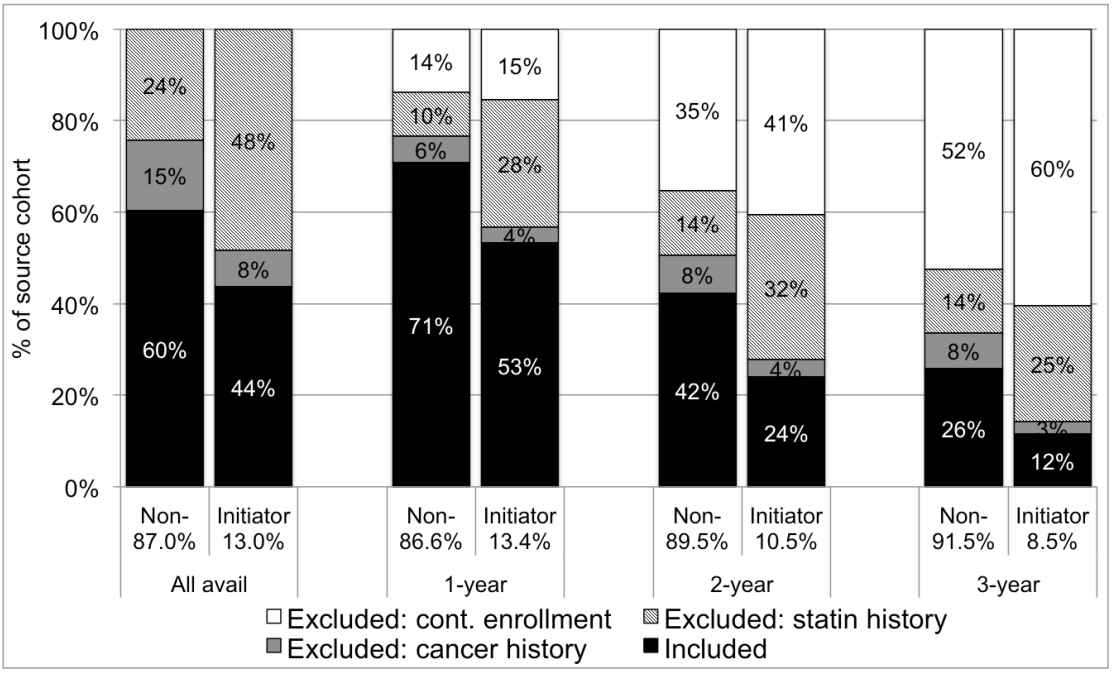
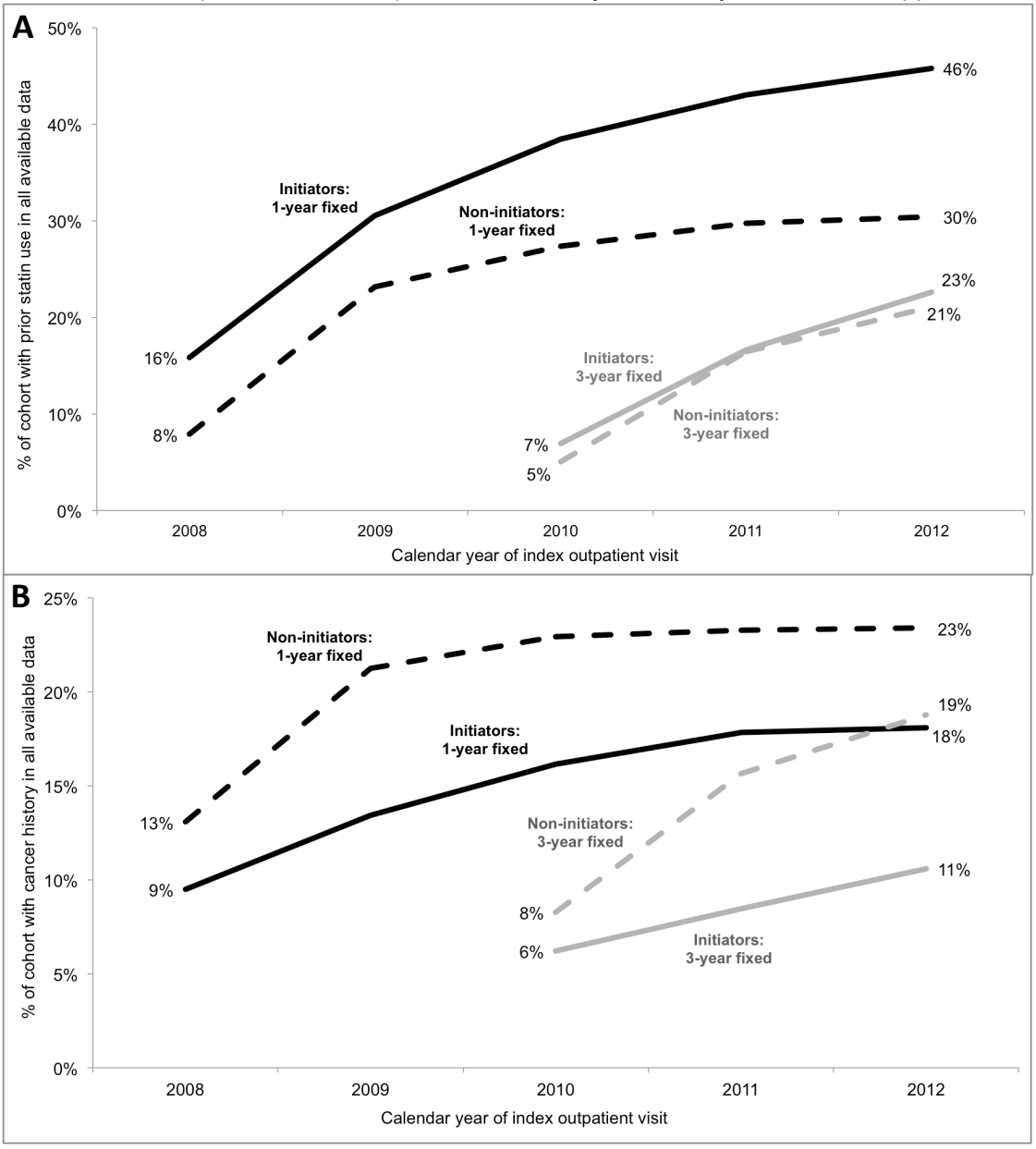
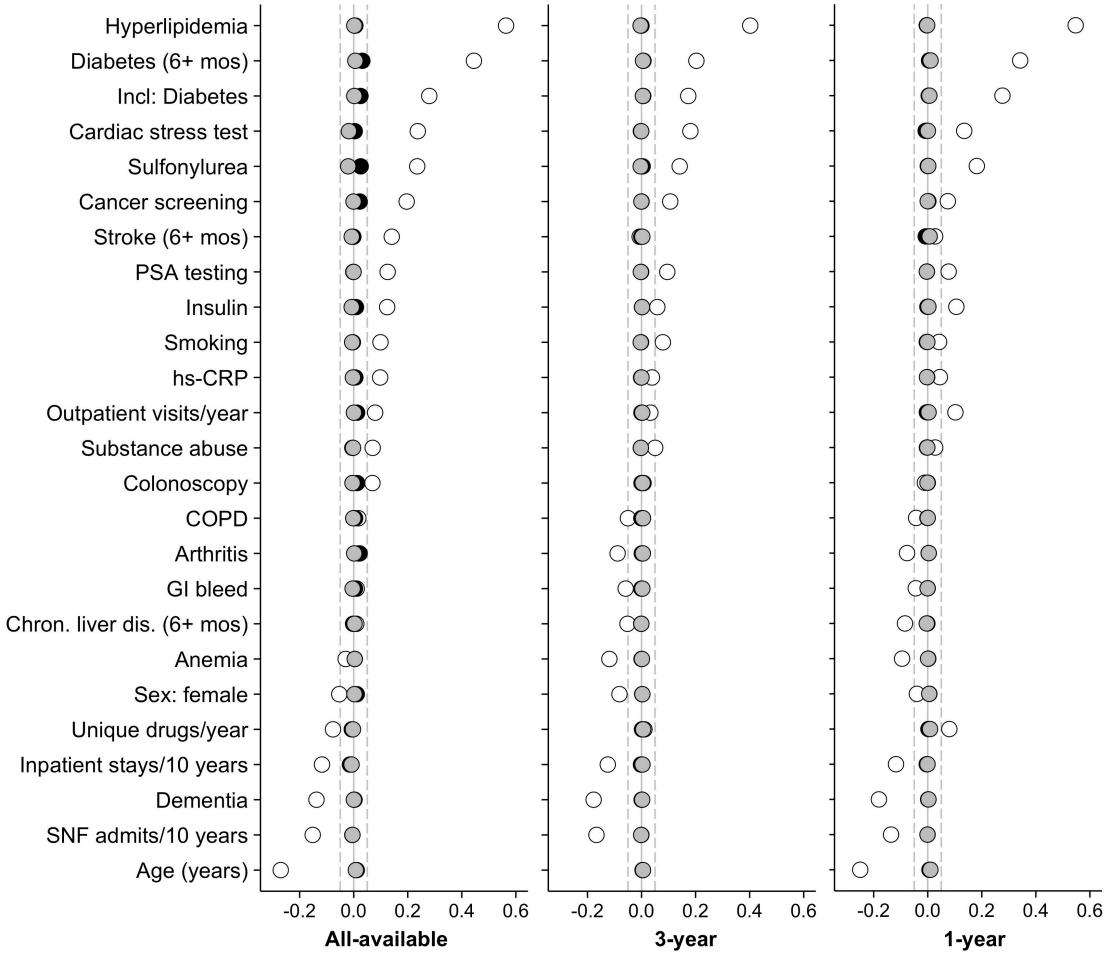


Figure 4.3. Proportion of the 1-year fixed and 3-year fixed cohort with observable history in the database of A) statin use and B) cancer for the 1-year and 3-year look-back approaches



[a] When using the fixed look-back approaches, actual classification is constant over time. The upward slope of the curves shown in this figure reflect the diminished power of using all-available database history in earlier calendar years, when available historical data in the database is sparse.

Figure 4.4. Average standardized mean difference for selected variables in the analysis of 6-month cancer, for the crude (white) analysis and SMRW analysis before (black) and after (grey) 1% asymmetric trimming

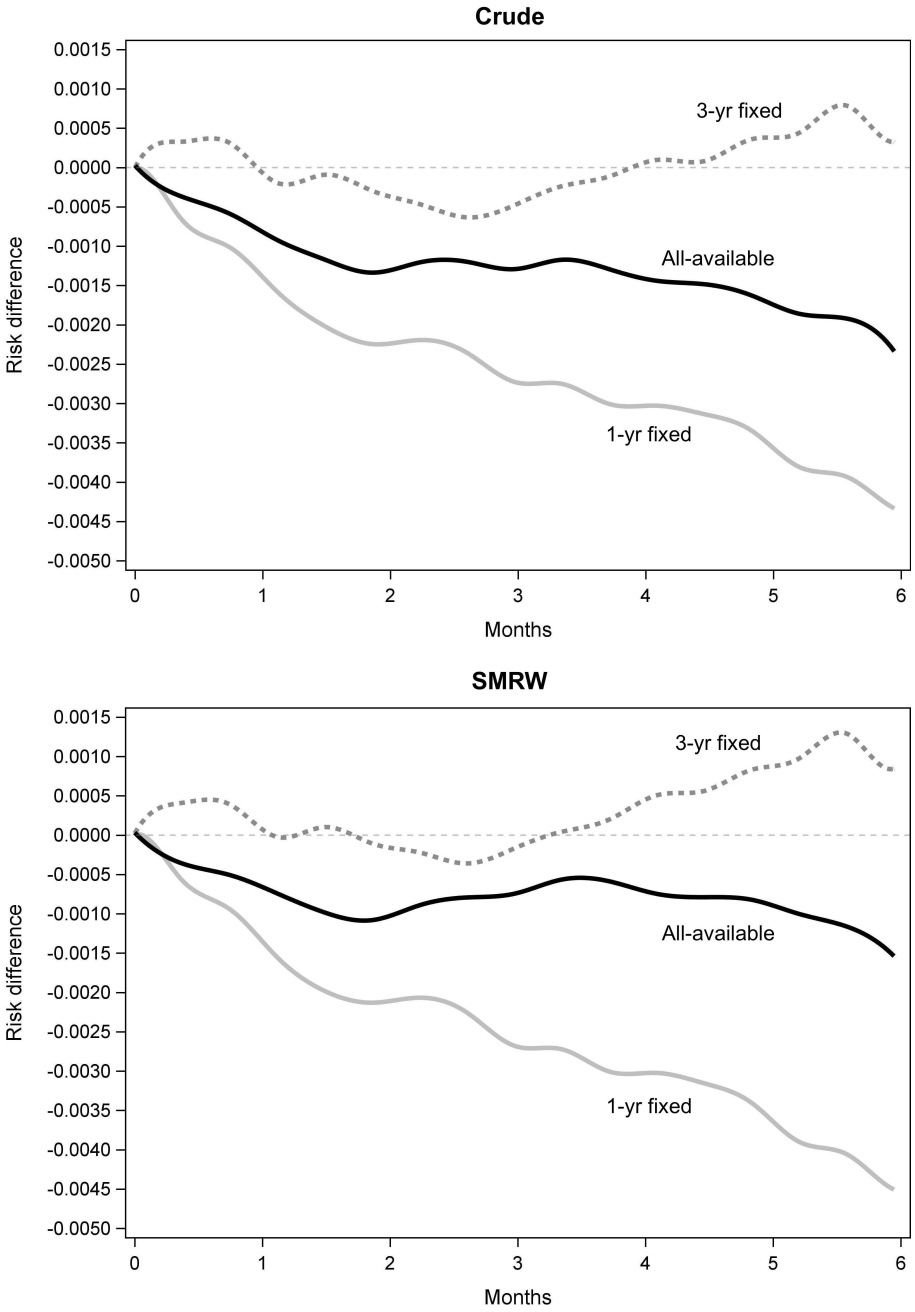


[a] Positive standardized differences indicate greater mean or proportion observed among initiators

[b] Dashed grey lines mark standardized differences of -0.05 and 0.05

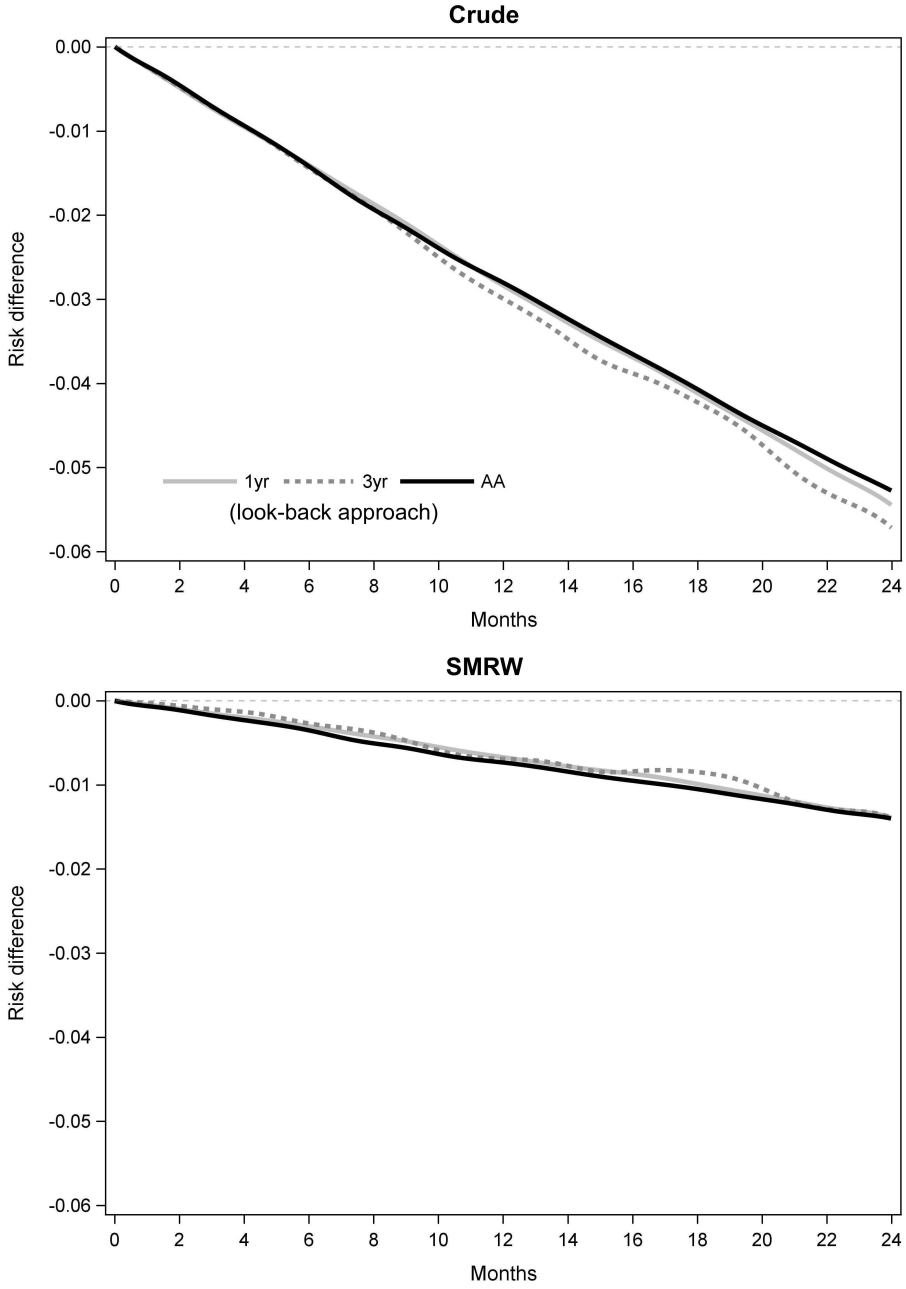
[c] This figure presents all variables which 1) were included in the propensity score model for the 6-month cancer analysis, 2) had a crude standardized difference > 0.05 for any look-back approach, and 3) was prevalent in at least 5% of users or non-users, for any look-back approach.

Figure 4.5. Crude and SMRW-adjusted cumulative risk differences in the 6-month cancer analysis using the all-available, 3-year, and 1-year look-back approaches



[a] We smoothed the curves using penalized B-splines with 15 knots.

Figure 4.6. Crude and SMRW-adjusted cumulative risk differences in the 2-year mortality analysis using the all-available, 3-year and 1-year look-back approaches



[a] We smoothed the curves using penalized B-splines with 15 knots.

CHAPTER 5

PERFORMANCE OF FIXED AND ALL-AVAILABLE LOOK-BACK APPROACHES IN LONGITUDINAL DATABASE STUDIES: A PLASMODE SIMULATION

5.1 Introduction

In the last 30 years, there has been a dramatic increase in the utilization of secondary health data for research, primarily administrative billing and electronic health records databases. However, the quality of information available in secondary healthcare databases varies widely and both missing and misclassified data are commonplace.

In longitudinal database studies, analysts typically characterize patient histories using look-back periods, which begin at some date of interest (e.g. the index or exposure date in a cohort study) and extends backward through time. In order to assure that all histories are classified with comparable accuracy, most studies apply fixed look-back periods that are the same length for every patient, ignoring the data outside of this period.[1,2] This approach is motivated by a desire to make covariate misclassification non-differential across important study variables (e.g. exposure, outcome, confounders). By doing so, the investigator seeks to assert that at least partial control for the confounder has been achieved, defined by Greenland as the scenario in which the adjusted estimate is closer to the truth than the unadjusted estimate.[8]

However, using fixed-look back periods in database studies is an imperfect solution to the problem of differential misclassification, since variable classification is affected by a range of factors beyond just database enrollment. Plausible mechanisms exist whereby using short fixed look-backs may increase differential misclassification (e.g. by outcome, if coding and billing practices differ among sicker patients who are more likely to experience

the outcome).[5,6] Even if fixed look-backs were able to produce non-differential misclassification, they are still unable to guarantee partial control for confounding unless a range of other criteria are also satisfied.[8-10,12,47,67] Furthermore, look-backs require study populations be restricted to patients who are continuously enrolled in the database throughout the entire look-back period. Such restrictions may compromise the external validity of findings since they inadvertently condition on a range of factors that determine database enrollment. An alternate approach has been suggested which considers all available database history, regardless of whether that history is available for all patients.[4,7]

Simulation studies have been conducted which indicate that all-available look-backs may be superior to fixed look-backs when classifying a confounder and cohort eligibility criterion.[4,104] A study conducted in a commercial claims database found that using all-available look-backs to classify confounders produces similar effects as short (i.e. 180 or 365-day) fixed look-back approaches across five exposure-outcome pairs.[14] However, this study was not designed to estimate bias. Furthermore, no literature exists evaluating the real-world (i.e. not simulated) performance of different look-back approaches when classifying eligibility criteria. The utility of using fixed look-backs depends on how informative the data is that they discard; if most predictors of exposure and outcome risk can be obtained from data proximal to the index date, fixed look-back periods may be preferable.

The purpose of this study was to assess bias and precision of estimates produced using various look-back approaches applied within administrative claims data. We used plasmode simulation, which provided a controlled environment in which bias could be estimated, that still reflected the complex, interrelated structure of real-world claims data.[17,18] We produced adjusted rate ratios and rate differences using all available and fixed look-back approaches to classify 1) a study eligibility criterion and 2) confounders adjusted for in analysis.

5.2 Methods

5.2.1 Overview

Plasmode simulation is a method that combines sampling (with replacement) from real-world source data with simulated components.[17,18] The source data sampled to produce plasmode datasets was the Medicare Current Beneficiary Survey (MCBS) and linked Medicare fee-for-service claims.[113-115] In the simulation, the MCBS interview data on treatment and disease history represent the true underlying health status, acting as independent predictors of the simulated exposure (conceptually: statin initiation) and outcome (conceptually: inpatient hospitalizations). We analyzed these plasmode datasets by applying the look-back approaches to the data available in the linked Medicare claims and then estimated bias by comparing to the true (simulated) treatment effect.

5.2.2 Selection of the source cohort from MCBS data

The MCBS Cost and Use survey is a nationally representative survey of Medicare beneficiaries with a rotating panel design (interviewing respondents three times per year for up to three years.[113-115] The source cohort for the plasmode simulation included MCBS respondents who completed at least one full year of interviews (Cost and Use module) between 2009 and 2011. We created an observation for each full year of data, setting the index date as the annual Health Status and Functioning (Fall) interview; thus, one beneficiary generated could generate up to three observations in the source data. We restricted the cohort to respondents who were enrolled in Medicare Part A, B and D at the time of their index interview and who had complete MCBS data for all study variables. Figure 5.1 presents the study schematic used to select and assess the source data from MCBS and linked claims.

5.2.3 MCBS covariates used as underling/true confounders

For each observation in the source cohort, we used information collected during the Fall interview to assess demographics (age, sex), body-mass-index (BMI), any history of

diabetes, and routine cancer screening in the last year. We classified subjects as frail if they reported that a health condition caused them to avoid, have difficulty, or require assistance with any activity of daily living.[116-120] Using MCBS data on prescribed medicine events (which is based on a combined assessment of Part D claims and respondent self-report), we assessed antihypertensive use in the two rounds leading up to the Fall interview. For certain variables, MCBS respondents provide information regarding two distinct time periods: 1) in a proximal period occurring within one year before the index date and 2) in the distal period beginning at least one year before the index date. These variables include: serious cancer (sites: lung, stomach, kidney, brain, throat, head, colon, uterus, ovary, cervix), and major adverse cardiovascular events (myocardial infarction/heart attack, congestive heart failure, angina pectoris, or stroke/brain hemorrhage). A table displaying detailed definitions for all MCBS variables is included in Table A5.2.1. The procedure we used to select which MCBS variables were included in the simulation is described in Appendix 5.1.

5.2.4 Assessing true prior use of statins

For each observation, we defined prior statin use as any reported exposure in any available MCBS or claims data before the Fall interview.

5.2.5 Simulation

We generated 1,500 plasmode datasets by selecting 10,000 observations with replacement from the source data, then simulating exposures and outcomes (Figure 5.2).

5.2.5.1 Exposure simulation

For each observation we calculated exposure probability as a function of 1) an intercept term calibrated to produce a pre-specified exposure prevalence, 2) all "true confounders" as reflected in MCBS.

$$p(exposure = 1) = \frac{1}{1 + e^{-(\beta_0 + \vec{\beta}_i * \vec{x}_i)}}$$

where:

β_0 = intercept term, calibrated to produce a pre-specified exposure prevalence

$\vec{\beta}_i$ = a vector of log-ORs for the independent relative effect of each MCBS confounder (*i*) on the likelihood of exposure (Table A5.2.2)

$\vec{x}_i$ = a vector of values for each confounder (*i*) pulled from the MCBS data.

The log-ORs corresponding to each MCBS variable ($\vec{\beta}_i$) were deliberately specified to produce unidirectional *upwards* confounding, making exposure appear less protective/more harmful (Table A5.2.2). This prevents multiple biases from balancing one another out, which may obscure results. To simulate a binary exposure (statin initiation), we drew from a Bernoulli distribution using the exposure probabilities produced by the above equation. We varied the intercept term between those with and those without true prior statin use such that those with prior statin use were *more* likely to fill a statin after the index interview.

5.2.5.2 Outcome simulation

We calculated the expected number of outcomes over a fixed time period as a function of 1) an intercept term calibrated to produce a pre-specified outcome incidence, 2) exposure status, and 3) all "true confounders" as reflected in MCBS.

$$RR\text{: } E(Y) = e^{(\beta_0 + \beta_e * exposure + \vec{\beta}_i * \vec{x}_i)}$$

$$or$$

$$RD\text{: } E(Y) = \beta_0 + \beta_e * exposure + \vec{\beta}_i * \vec{x}_i$$

where:

β_0 = intercept term, calibrated to produce a pre-specified outcome incidence (number of events Y over a fixed time period)

β_e = the log-RR (or RD) for the effect of interest

$\vec{\beta}_i$ = a vector of log-RRs (RDs) for the independent effects of each MCBS confounder (*i*) on the outcome incidence (Table A5.2.2)

$\vec{x}_i$ = a vector of values for each confounder (*i*) pulled from the MCBS data.

For each observation, we calculated two expected number of outcomes, one corresponding to each level of exposure. The number of outcomes Y under each exposure

was assigned using a random number from a Poisson distribution based on these expected values. The background outcome incidence rate (in absence of exposure) was simulated to be the same among those without prior statin exposure as well as prior statin users. However, the effect of exposure on outcome incidence was less pronounced among prior statin users (RR=0.8 or RD=-0.04) than among true new users (RR=0.4 or RD=-0.08). Thus, incorrectly including prior statin users in the study would bias statin treatment effect estimates towards the null. A more detailed description of the procedure used to calibrate intercept terms and specify $\vec{\beta}_i$ are included in Appendix 5.1.

5.2.6 Using look-back approaches to assess observable-proxies in claims data

Next, we used the linked claims data to assess claims-derived proxies for the true confounders/modifiers measured in MCBS. Within each of the 1,500 plasmode datasets, we selected a distinct cohort for each look-back approach restricting to respondents who had 1) the required continuous enrollment (all-available look-back: 6-months, 1-year fixed: one-year, and 3-year fixed: three years), and 2) no claims-documented statin use in the look-back.

We used Part A and B claims to identify diagnoses and care associated with diabetes and cancer, cancer screening (mammograms, pap-smears, rectal exams, and prostate specific antigen testing), and major cardiovascular events, including myocardial infarction, stable/unstable angina, heart failure, and stroke. In addition to Part A and B claims, we used durable medical equipment (e.g. home oxygen, hospital bed or wheelchair) to capture proxies for frailty.[116] We assessed Part D prescription claims to identify use of anti-hypertensives (angiotensin converting enzyme (ACE) inhibitors, angiotensin-II receptor blockers (ARB), and anti-diabetics (biguanides, sulfonylurea, insulin, thiazolidinedione, DPP-4 inhibitors).

5.2.7 Analyses and statistics

Within the cohorts selected with each look-back approach, we used logistic regression to model the probability of the simulated treatment conditional on the observed claims data (i.e. propensity scores). We used stabilized inverse probability of treatment weighting (IPTW) to estimate adjusted rate ratios and rate differences.

We calculated bias by contrasting effect estimates with the known/simulated truth, which we calculated by comparing simulated counterfactuals within a cohort formed with perfect selection based on the true MCBS data on prior statin use. We present the median bias from the 1,500 plasmode datasets and 95% confidence intervals estimated by drawing the 2.5^{th} and 97.5^{th} percentiles of the distribution of the statin effect estimates. We assessed the precision of estimates using confidence interval width (CIW), which we calculated as the difference between the log-RR confidence limits. We assessed the precision-bias trade-off using the root-mean-squared-error (rMSE), which is equal to the square root of the squared bias plus the squared Monte Carlo standard error. We also evaluated the proportion of the 1,500 iterations in which Walker's equipoise criterion was met (i.e. where greater than 50% of both exposed and unexposed patients have preference scores between 0.3 and 0.7).[100]

The net bias in IPTW-adjusted estimates represents a summation of two component biases: 1) the residual confounding bias caused by misclassification / inadequate ascertainment of true confounders (hereafter referred to as residual confounding), and 2) modification bias resulting from the inclusion of subjects with prior statin use who have a diminished effect of interest (hereafter referred to as prior user bias). We estimated effects isolating each form of bias (detailed explanation provided in Table 5.1). We estimated the effect isolating residual confounding bias due to misclassification by adjusting for confounding using IPTW models based on the imperfect claims, within a perfectly selected cohort based on the true MCBS data on prior statin use. We estimated the effect isolating bias caused by modification / inclusion of prior statin users by comparing simulated

counterfactuals within an imperfectly selected cohort, based on claims evidence of prior statin use. Appendix 5.1 highlights some additional considerations for interpreting estimates of component bias.

Sub-analyses

In order to explore the performance all-available look-backs in scenarios with extreme differential information inaccuracy, we conducted sub-analyses in which we intentionally left-truncated data histories differentially by exposure status, outcome status, and both simultaneously. For example, in one analyses, we considered all data history up to one year among initiators while considering all data up to three years of data among non-initiators. A more detailed description of these and additional sub-analyses we conducted is available in Appendix 5.1.

5.3 Results

5.3.1 Characteristics of the source (MCBS) cohort

The source cohort included 5,176 total observations, 3,025 (58.4%) with identifiable prior statin use and 2,151 (41.6%) with no prior statin use (Table 5.2). Among patients without prior statin use, 71% were female, the mean age was 79.6 years, and the mean BMI was 26.3 kg/m^2. Approximately 13% had difficulty with at least one activity of daily living and 10% had a history of a serious (non-skin) cancer. Statin users were more likely to have a history of diabetes (30% vs. 17%) and major cardiovascular events (40% vs. 21%), compared to patients with no prior statin use. For most variables, the claims-derived proxies assessed using any look-back approach led to over-ascertainment of the “true” MCBS confounders, except for obesity.

5.3.2 Characteristics of the plasmode cohorts selected using different look-back approaches

Plasmode cohorts (N=10,000) included a median of 4,155 observations with truly no prior statin use (Table A5.2.3). The median sample size of cohorts selected using the all-

available approach was similar (4,125 people). Cohorts selected using the 1-year look-back were 13% larger, including a median 4,661 observations, while the cohorts selected using the 3-year look-back were 20% smaller, including a median 3,314 observations. Variation in sample size was primarily driven by the continuous enrollment requirement. The all-available approach's 6-month requirement excluded only 1% of subjects while the 3-year requirement excluded 22%.

The proportion of patients excluded due to prior statin use did not vary widely between the look-back approaches (58% for the all-available and 3-year look-backs, 50% for the 1-year) (Table 5.3). However, this translated to dramatic differences in the proportion of patients in the cohorts who were prior statin users. In cohorts selected using 1-year look-backs, 26% of initiators and 10% of non-initiators were prior users. In cohorts selected using all-available and 3-year look-backs, prior users comprised 14% and 13% of initiators, respectively, and 5% and 4% of non-initiators, respectively. It follows logic that in all cohorts, inclusion of prior users is more likely among initiators.

As shown in Figure 5.3, in cohorts selected using the all-available approach, initiators had more observable database enrollment for Medicare Parts A and B than non-initiators but similar Part D enrollment. Differences were slightly more pronounced by outcome status, with those with outcomes having more available database history. Regardless of exposure and outcome status, the majority had greater than 3-years of cumulative Part A, B, and D enrollment in the database.

In a propensity-score model fit using the true MCBS confounders, Walker's equipoise criteria was met for ≈60% of both simulated exposure groups. The all-available approach which used time-stratified adjustment produced equipoise estimates closest to the truth (76% of initiators and 81% of non-initiators).

5.3.3 Bias in effect estimates produced using different look-back approaches

The crude (unadjusted) estimates produced by the different look-back approaches ranged from 0.73 to 0.80, compared to the true effect (RR=0.40). IPTW-adjusted estimates, meanwhile, ranged from 0.59 to 0.66, indicating that substantial bias remained for all of the look-back approaches, even after adjustment. The precision of estimates produced by each look-back approach was a function of how many people were selected for inclusion. Thus, the more inclusive 1-year look-back consistently produced more precise estimates (log-RR CI width = 0.21) than the all-available (0.24) and 3-year (0.25) look-back approaches. For all look-back approaches, confidence intervals widened only slightly after incorporating time-stratified adjustment.

Net bias in IPTW-adjusted rate ratio estimates (Figure 5.4) were similar across all look-back approaches and had substantially overlapping confidence intervals. However, the all-available approach using time-stratified adjustment consistently produced the least biased point estimate (bias=0.39 [0.28, 0.50]) and the best rMSE (0.399). The 1-year look-back performed worst in terms of both bias (0.50 [0.39, 0.60]) and rMSE (0.499).

Isolating the two individual sources of bias (i.e. inadequate confounding adjustment and inclusion of prior users) provides a more informative picture. When isolating the residual confounding bias, the all-available approach using time-stratified adjustment again outperformed the other methods (bias=0.33 [0.21, 0.45]). However, the point estimate for the 3-year look-back was, in this case, slightly more biased than the 1-year look-back. The fact that the 3-year look-back estimate produced less net bias was due to its superior performance eliminating prior users. Isolating the prior user bias, the 1-year look-back was significantly more biased (0.13 [0.10, 0.16]) than all other approaches (e.g. time-stratified all-available: 0.06 [0.04, 0.09]).

5.3.4 Sub-analyses

All findings were nearly identical in parallel simulations of a rate difference effect, homogenous on the absolute scale (Figure A5.3.1, Figure A5.3.2). In sub-analyses where observable database history was left-truncated informatively by exposure and/or outcome status (Figure 5.5), the all-available approach always generated at least partial control for confounding. The only exception was that in one sub-analysis with informative left-truncation (3-yr[E=1|O=0], 2-yr[E=0|O=0], 1-yr[E=1|O>0], 6-mo[E=1|O=0]) and a harmful effect of the simulated exposure, the IPTW adjustment failed to reduce any confounding compared the crude (Figure A5.3.3). Results from additional sub-analyses are available in Appendix 5.1.

5.4 Discussion

The findings of this plasmode simulation provide a detailed picture of how different look-back approaches perform, albeit within a narrow setting. Estimates produced by the different look-back approaches did not vary widely; however the all-available approaches repeatedly produced the least biased point estimates. Approaches employing time-stratified adjustment for confounders were marginally superior to time-fixed adjustment. These findings were robust to: 1) variation in the effect of exposure (protective and harmful), 2) estimation of multiplicative effects using the rate ratio and absolute effects using the rate difference, 3) multiple sources of bias (residual confounding bias due to misclassification, prior user bias), 3) simulation of time-stratified and time-fixed confounding, and 4) adjustment for time-stratified and time-fixed confounders. All-available approaches produced larger and more inclusive cohorts than long-fixed look-backs, and included substantially fewer prior statin users than short fixed look-backs.

We used plasmode simulation, an innovative approach which leverages desirable features of both simulation methods (i.e. a known truth and the ability to estimate bias) and applied analyses (i.e. complex, interrelated data structures). By design, the MCBS survey data represented the true underlying confounders in our simulation. However, the survey

data (much of which relies on self-report) is an imperfect indicator of true underlying health status which confounds associations. It is likely that for certain covariates the claims data may be a more accurate measure of true underlying health status than self-reported survey data. It is possible that the higher covariate prevalences we observed in the claims data may be the result of under-reporting in the MCBS survey data.

We conducted a range of sub-analyses in which we informatively left-truncated the amount of database enrollment such that information accuracy was extremely differential with respect to exposure and/or outcome status. Even though these sub-analyses were intentionally extreme, it is notable that the all-available approach never led to confounding-adjusted estimates that were more biased than the crude, which is one of the primary concerns that discourages researchers from using all-available look-backs. It is difficult to say whether we can expect this finding to hold in other study settings with different variable frequencies and confounding structures or in different databases where enrollment history may vary more widely. Regardless, a thorough investigator can assess whether using all-available look-backs will produce differential information accuracy by considering the amount of database history available within relevant sub-groups.

Our findings are consistent with recently published research demonstrating that in real-world settings, using all-available look-back approaches produces similar estimates to conventional fixed look-backs. Recent work by Nakasian, Rassen & Franklin compared adjusted hazard ratios produced by different look-back approaches applied to five different studies conducted within a commercial insurance claims database (Optum/United).[14] They used all-available and short fixed look-backs (180 or 365 days) to ascertain and adjust for confounders. Our study expands on the findings of Nakasian et al. by 1) calculating actual bias using simulated counterfactuals (thus allowing us to better evaluate the performance of each approach), 2) evaluating longer fixed look-backs, 3) evaluating the use of look-backs to

implement eligibility criteria and 4) evaluating the use of time-stratified adjustment for confounding.

In the cohort selected with a one-year fixed look-back, we assessed that 26% of initiators actually had prior (unidentified) use of a statin at any point before the index date. This proportion is compatible with the results of another study (conducted using MCBS data) which found that 20% of new users identified in claims data actually had prior statin use within the last year.[121] Our findings indicate that even long look-backs may not be completely able to eliminate prior users from the cohort. In cohorts selected using the all-available approach, 14% of assumed "new-users" still had unidentified prior statin exposures. In many studies, prior use of the exposure will be much more rare. In these cases, the benefit of using an all-available look-back to reduce prior user bias may be less pronounced.

The marginally superior performance we observed in estimates produced by all-available look-backs is consistent with results generated from pure simulation studies. Brunelli et al. found that all-available look-backs were superior to fixed look-backs when ascertaining covariates for adjustment.[4] A similar simulation demonstrated the same finding when applying all-available look-backs to implement study eligibility criteria.[104] While informative, the findings produced by these simulations are limited since they aren't designed to capture the complex structure of real-world data or realities encountered in actual study settings. For example, our study and others conducted in applied settings indicate that the amount of database history does not appear to vary widely between exposure groups (especially when using well-chosen active comparators). This may (to some degree) alleviate some of the theoretical concerns about differential information accuracy biasing estimates produced by the all-available look-back, at least within this study setting.

Fixed look-backs reduce differential misclassification by selectively ignoring information. Sholom Wacholder warned against such approaches in a series of papers outlining principals for control selection in case-control designs, "…study designs that tolerate errors in one group so that errors are not differential, should be examined carefully. Strict adherence to the principle of comparable accuracy used to ensure non-differential misclassification in choosing controls for case-control studies may not be advisable when it would require controls with as much error as cases instead of more accurate controls."[12,13,122]

There are some important limitations to our findings which should be highlighted. Using a simulation design affords a certain degree of control over study parameters (e.g. the magnitude and direction of the effect-of-interest); however it is unclear whether or not these findings can be generalized to other study settings (e.g. other databases, study populations). We evaluated a fairly narrow source population (i.e. MCBS respondents who completed a full year of interviews), and findings could vary in other study populations. For example, only 1.3% of the source cohort lacked the required 6-month continuous Part A/B/D enrollment (Table 5.3, Figure 5.3). It is also important to note that studies have been published in which all available and fixed look-back approaches produce meaningfully different estimates.[123] The fact that this study and the Nakasian et al. study both failed to produce meaningfully different estimates may indicate further interrogation is needed to identify settings in which all-available approaches are appropriate.

Finally, the limited pool of respondents in the source MCBS data limited us to simulating cohorts of N=10,000, which is substantially smaller than typical administrative claims analyses. This prevented us from simulating rare outcomes and exposures and also reduced our power to detect meaningful differences in estimates produced using the different approaches. In order to assure collapsibility of simulated effects, we simulated a recurring outcome (conceptually, inpatient hospitalizations) and estimate rate ratios.

However, our simulation assumed a fixed complete follow-up for all subjects; since rate ratios are approximately equal the risk ratio in this setting, these findings should be applicable to analyses estimating risk ratios.[124]

5.5 Conclusions

Our findings indicate that all-available look-back approaches may outperform fixed look-back approaches in pharmacoepidemiologic studies based on claims data. Our results also provide context for investigators seeking to understand differences between estimates produced by all-available and fixed look-back approaches. Investigators employing all-available look-backs should check to ensure the amount of database history does not vary dramatically across levels of exposures, outcomes, and confounders. The fact that point estimates produced by the all-available look-back approach were marginally less biased than fixed look-backs may not necessarily indicate meaningful differences. However, they should encourage us to continue to explore and test the performance of all-available approaches applied in different databases, study populations, and comparisons. When designing a study, the investigator should always take into consideration the relevant time-periods for different confounding variables and eligibility criteria. When selecting a look-back approach, a well-designed study should always consider how much look-back is needed, whether differential classification can be identified in the data, and how continuous enrollment restrictions will impact external validity.

Table 5.1. Explanation of plasmode data used for effect estimation and bias calculations

Estimate	Exclusion of prior statin users	Confounding adjustment	Bias calculation
Truth [RR_{truth}]	MCBS data	Contrast counterfactuals	NA (Unbiased)
Net [RR_{net}] Misclassification + modification	Claims-derived proxy variables	IPTW using claims-derived proxy variables	$\ln(RR_{net}) - \ln(RR_{truth})$
Residual confounding due to covariate misclassification [RR_{RC}]	"True" MCBS data	IPTW using claims-derived proxy variables	$\ln(RR_{RC}) - \ln(RR_{truth})$
Modification due to including patients with **prior exposure** [RR_{PE}]	Claims-derived proxy variables	Contrast counterfactuals	$\ln(RR_{PE}) - \ln(RR_{truth})$

Table 5.2. Characteristics of the 2009-2011 Medicare Current Beneficiary Survey respondents included in the plasmode simulation source cohort

MCBS survey variables (true confounders)	Simulation parameters: Exposure odds ratio	Simulation parameters: Outcome rate ratio	No prior statin exposure (N=2,151)[a]: Mean	No prior statin exposure (N=2,151)[a]: S.D.	Prior exposure (N=3,025)[a]: Mean	Prior exposure (N=3,025)[a]: S.D.	Claims-derived proxy variables	All-avail. (N=2,188): Mean	1-year (N=2,622): Mean	3-year (N=2,268): Mean
Age (years)[b]	0.996	1.06	79.6	7.62	78.3	6.98	Age (years)[c] [mean]	79.4	79.2	79.4
Age (squared term)[b]	1.001	-								
BMI (kg/m^2)[b]	1.015	0.72	26.3	5.28	27.7	5.26				
BMI (squared term)[b]	1.002	1.005								
			N	**%**	**N**	**%**		**%**	**%**	**%**
Overweight (BMI: 25-30)	-	-	762	35.4%	1,238	40.9%	Obesity[c]	6.1%	3.8%	6.0%
Obese (BMI≥30)	-	-	455	21.2%	844	27.9%				
Sex: female[b]	0.50	0.56	1,531	71.2%	1,914	63.3%	Sex: female[c]	69.7%	68.9%	69.4%
Routine cancer screening[b,d]	0.50	0.61	1,192	55.4%	1,920	63.5%	Any cancer screening[c,d]	72.3%	47.0%	66.3%
							Mammogram	43.5%	27.6%	40.4%
							Pap smear	23.7%	7.6%	17.5%
							Prostate check	26.1%	17.9%	24.3%
Serious (non-skin) cancer[b]	-	-	204	9.5%	239	7.9%	Cancer[c]	28.9%	21.7%	27.6%
Within prior year[b]	4.00	3.75	37	1.7%	33	1.1%				
Before prior year[b]	2.00	1.96	167	7.8%	206	6.8%				
Major CV event[b]	-	-	456	21.2%	1,205	39.8%	Any major CV event	36.4%	25.5%	35.9%
Within prior year[b]	4.00	2.22	111	5.2%	317	10.5%	Acute myocardial infarction[c]	2.4%	1.2%	2.4%
Before prior year[b]	2.00	1.77	345	16.0%	888	29.4%	Angina[c]	2.5%	1.3%	2.6%
							Stable angina[c]	6.3%	3.7%	6.0%
							Heart failure (acute)[c]	17.4%	12.1%	17.2%
							Stroke[c]	23.2%	14.0%	22.8%
Ace inhibitor or ARB (prior two interviews)[b]	0.50	0.5	1,031	47.9%	1,860	61.5%	Any ACE inhibitor or ARB	52.0%	47.0%	51.0%
							ACE Inhibitor[c]	37.7%	31.4%	36.3%
							ARB[c]	22.6%	18.2%	21.5%
Diabetes (ever)[b]	2.00	2.00	359	16.7%	908	30.0%	Any diabetes indicator	29.6%	26.6%	29.4%
							Diabetes (diagnosis)[c]	29.3%	26.4%	29.2%
							Diabetes (complications)[c]	7.8%	6.4%	7.8%
							Biguanide[c]	9.4%	8.1%	9.0%
							Sulfonylurea[c]	7.6%	7.1%	7.7%
							Insulin[c]	3.2%	3.2%	3.3%
							Thiazolidinedione[c]	3.6%	2.3%	3.0%
							DPP-4 Inhibitors[c]	1.9%	1.5%	1.9%
Frailty (difficulty with any activity of daily living)[b,e]	2.00	3.02	272	12.6%	321	10.6%	Any frailty indicator (broad)	55.7%	37.7%	52.7%
							Ambulatory life support[c]	28.5%	11.4%	22.7%
							Difficulty walking[c]	23.7%	14.1%	23.4%
							HF Frailty[c]	23.4%	16.0%	23.1%
							Rehab[c]	14.5%	7.8%	14.2%
							Weakness[c]	14.4%	8.9%	14.4%

MCBS survey variables (true confounders)	Simulation parameters		No prior statin exposure (N=2,151)[a]		Prior exposure (N=3,025)[a]		Claims-derived proxy variables	All-avail. (N=2,188)	1-year (N=2,622)	3-year (N=2,268)
	Exposure odds ratio	Outcome rate ratio	**Mean**	**S.D.**	**Mean**	**S.D.**		**Mean**	**Mean**	**Mean**
							Home oxygen[c]	8.6%	7.1%	8.2%
							Wheelchair	5.4%	2.4%	4.2%
							Home hospital bed	2.4%	1.4%	2.3%

[a] In total, the source data included 5,176 observations from 3,334 beneficiaries. The 2,151 observations with no recorded statin use before the index date included 1,456 unique MCBS respondents (an average of 1.48 observations per respondent). The 3,025 observations with recorded statin use before the index date included 1954 unique respondents (an average of 1.55 observations per respondent).

[b] These "true" confounding variables were used to simulate exposures and outcomes

[c] These variables were included in propensity score models.

[d] Routine cancer screening (within the last year) was defined as either a mammogram or pap smear among women or a digital rectal prostate exam or blood test for rectal cancer among men.

[e] Activities of daily living: bathing, dressing, eating, walking, using the toilet, or getting in and out of a bed/chair.

Table 5.3. Simulation characteristics and overview of results for each look-back approach (median results among 1,500 iterations) in the primary simulation with a protective effect of exposure and time-stratified confounding (+/- 1-yr)

	% of cohort in equipoise[a]		Exclusions (%)[a]		Prior statin use (% of final cohort)[a]		Rate ratio (true effect = 0.40)			Rate difference (true effect = -0.08)		
	Initiator	Non-initiator	Cont. enroll.	Prior statin use	Initiator	Non-initiator	Crude	IPTW	C.I. width	Crude	IPTW	C.I. width
MCBS	58%	61%	0%	58%	0%	0%	0.70	0.40 (0.36, 0.45)	0.24	0.54	-0.08 (-0.21, 0.03)	0.24
Time-stratified adjustment (+/- 6-mo)												
All avail.	76%	80%	1%	58%	14%	5%	0.74	0.59 (0.53, 0.67)	0.24	0.50	0.26 (0.14, 0.39)	0.24
1-yr	77%	81%	7%	50%	26%	10%	0.80	0.65 (0.58, 0.72)	0.22	0.53	0.30 (0.19, 0.42)	0.23
3-yr	78%	82%	22%	58%	13%	4%	0.73	0.60 (0.53, 0.68)	0.26	0.55	0.32 (0.17, 0.47)	0.30
Time-fixed adjustment												
All avail.	80%	84%	1%	58%	14%	5%	0.74	0.61 (0.54, 0.69)	0.24	0.50	0.30 (0.18, 0.43)	0.25
1-yr	79%	83%	7%	50%	26%	10%	0.80	0.66 (0.59, 0.73)	0.21	0.53	0.32 (0.20, 0.43)	0.23
3-yr	81%	85%	22%	58%	13%	4%	0.73	0.63 (0.55, 0.71)	0.25	0.55	0.37 (0.22, 0.53)	0.31
Sub-analyses[b]												
3yr(E=0)|1yr(E=1)	60%	61%	1%	53%	27%	5%	0.78	0.65 (0.57, 0.73)	0.25	0.49	0.29 (0.16, 0.44)	0.27
3yr(E=1)|1yr(E=0)	58%	65%	1%	54%	15%	11%	0.76	0.61 (0.54, 0.69)	0.24	0.52	0.29 (0.16, 0.42)	0.25
3yr(O=0)|1yr(O>0)	78%	82%	1%	54%	21%	7%	0.84	0.70 (0.63, 0.77)	0.20	0.59	0.35 (0.23, 0.46)	0.23
3yr(O>0)/1yr(O=0)	81%	84%	1%	53%	22%	9%	0.70	0.61 (0.54, 0.68)	0.24	0.42	0.25 (0.14, 0.37)	0.23
Differential info #1[c]	54%	55%	1%	53%	16%	11%	0.73	0.56 (0.49, 0.63)	0.25	0.47	0.20 (0.06, 0.34)	0.27
Differential info #2[d]	47%	46%	1%	52%	27%	6%	0.81	0.73 (0.64, 0.84)	0.28	0.52	0.41 (0.26, 0.58)	0.32
Differential info #3[e]	52%	57%	1%	53%	16%	11%	0.80	0.71 (0.62, 0.80)	0.25	0.56	0.40 (0.28, 0.54)	0.26
Differential info #4[f]	55%	58%	1%	52%	28%	6%	0.85	0.65 (0.56, 0.74)	0.28	0.57	0.27 (0.13, 0.43)	0.30

[a] These results reflect the results from the simulation of a homogenous rate ratio. However, the values of these statistics were nearly identical in simulations of a homogenous rate difference.

[b] In the sub-analyses listed below, we restricted the cohort to respondents with at least 3-years continuous enrollment then intentionally censored the look-back selectively by exposure and outcome status.

[c] Differential Info #1: 3-yr(E=1|O>0), 2-yr(E=1|O=0), 1-yr(E=0|O>0), 6-mo(E=0|O=0)

[d] Differential Info #2: 3-yr(E=1|O=0), 2-yr(E=0|O=0), 1-yr(E=1|O>0), 6-mo(E=1|O=0)

[e] Differential Info #3: 3-yr(E=0|O>0), 2-yr(E=1|O>0), 1-yr(E=0|O=0), 6-mo(E=0|O>0)

[f] Differential Info #4: 3-yr(E=0|O=0), 2-yr(E=0|O>0), 1-yr(E=1|O=0), 6-mo(E=1|O>0)

Figure 5.1. A schematic illustrating the structure of the plasmode dataset layering real-world data (the Medicare Current Beneficiary Survey [MCBS] and linked Medicare claims) with simulated exposures and outcomes.
The source cohort included MCBS respondents from 2009-2011. We simulated exposures and outcomes as a function of covariates ascertained from the self-reported MCBS interview data (bottom). We then applied each look-back approach using the linked claims data (top) to produce adjusted effect estimates.

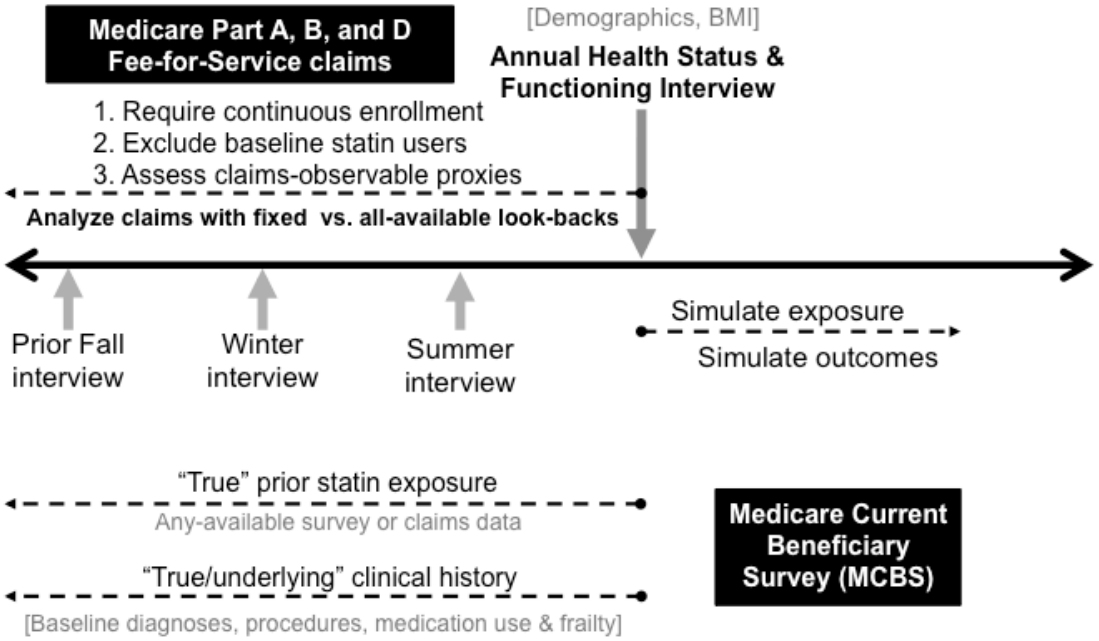
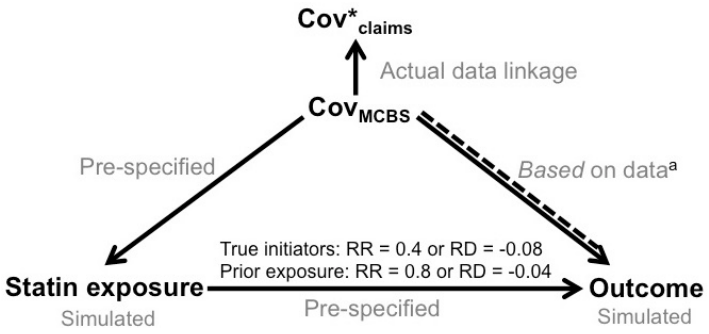


Figure 5.2. A directed acyclic graph demonstrating the relationships between real and simulated data components in the plasmode simulation.
Exposures and outcomes were simulated as a function of covariates assessed in the MCBS survey data. The effect of exposure on the outcome was diminished among patients with prior statin exposures. We applied look-backs within the MCBS-linked Medicare Part A/B/D claims data.



[a] The relationships between the covariates assessed in the MCBS data and the simulated outcome were based on the observed relationship between those same covariates and 1-year mortality, evaluated using a multivariate Cox model (these methods are described in more detail in Appendix 5.1).

Figure 5.3. Box-and whisker plots displaying distribution of available Medicare Part A, B, and D data history (in months) using the all-available look-back approach, stratified by A) exposure and B) outcome status.
Regardless of exposure or outcome status, a majority of people have greater than three years of Part A/B/D history. A) The amount of available database history does not vary strongly by exposure status, although initiators have slightly more Part A/B history. B) Compared to those without outcomes, those with outcomes appeared to have more Part A/B history and slightly more Part D history.

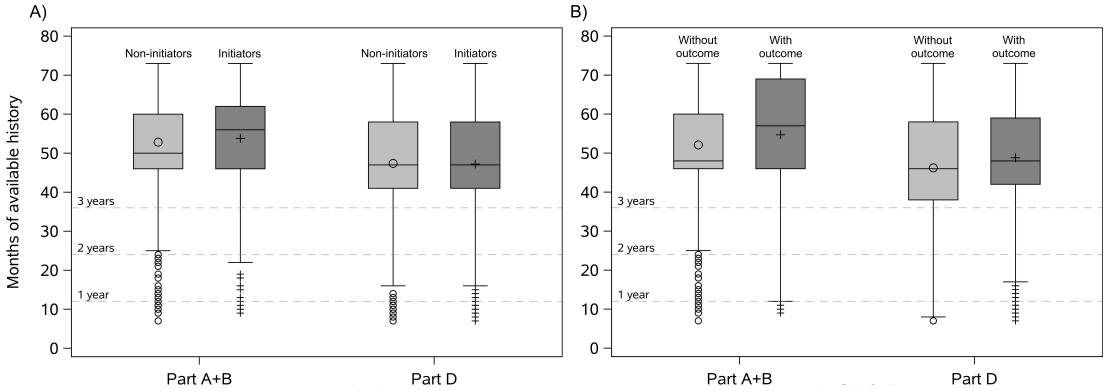


Figure 5.4. Forest plot displaying performance of various look-back approaches when estimating IPTW-adjusted estimates of the rate ratio in simulations of a multiplicative exposure effect.

Bias estimates are accompanied by 95% confidence intervals (CIs) and corresponding estimates of the root-mean-squared-error (rMSE). For each look-back approach, the plot includes estimates of the net (total) bias, residual confounding bias due to misclassification of covariates, and bias due to inclusion of prior statin users. *(Note: the corresponding figure for the rate difference analysis is presented in Figure A5.3.1.)*

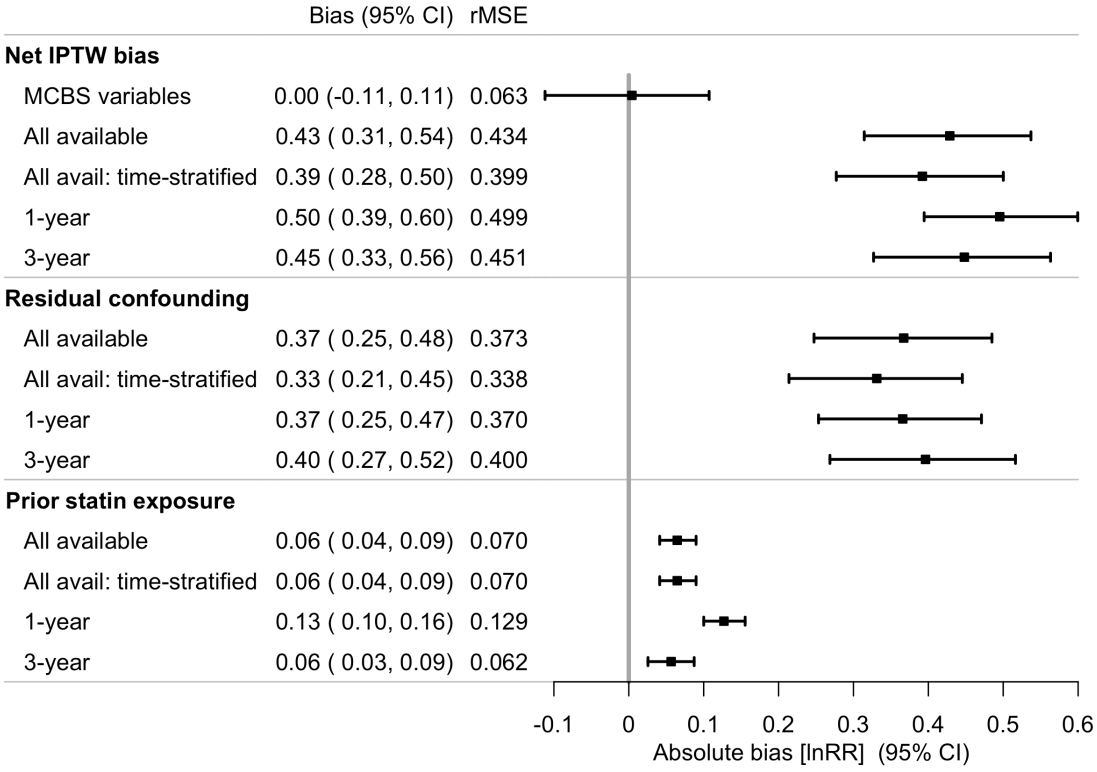
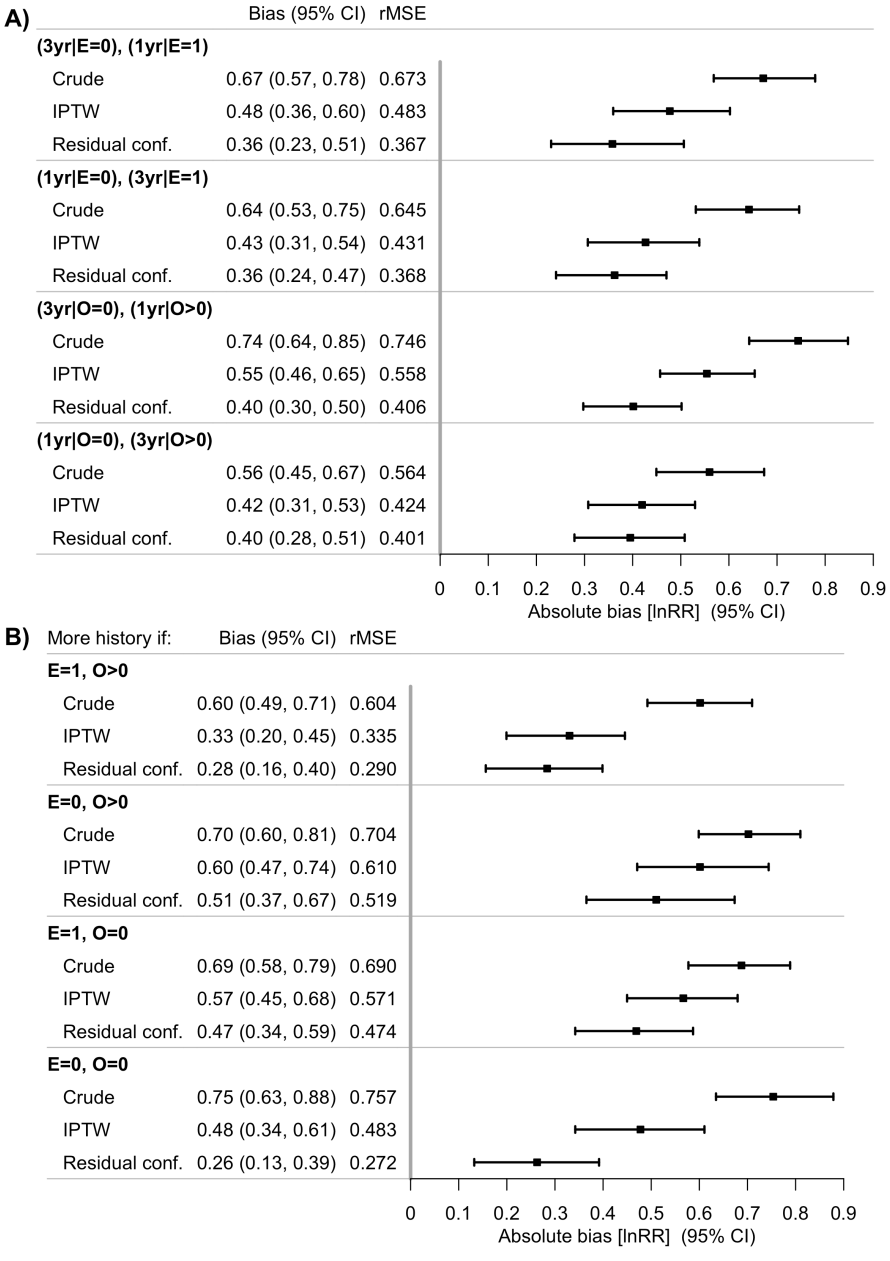


Figure 5.5. Forest plot displaying performance of the all-available look-back approach in sub-analyses with highly differential information accuracy A) by exposure or outcome status, and B) by exposure and outcome status (simultaneously) when estimating IPTW-adjusted estimates of the rate ratio in simulations of a multiplicative exposure effect.

E=exposure, O=outcome. Bias estimates are accompanied by 95% confidence intervals (CIs) and corresponding estimates of the root-mean-squared-error (rMSE). For each sub-analysis, the plot includes estimates of the net (total) bias and residual confounding bias due to misclassification of covariates. Partial control for confounding holds in all analyses since the residual confounding bias due to misclassification of covariates is closer to zero than the crude (unadjusted) bias. *(Note: the corresponding figure for the rate ratio analysis is presented in Figure A5.3.2.)*



CHAPTER 6

CONCLUSIONS AND PUBLIC HEALTH SIGNIFICANCE

6.1 Summary of specific aims

In recent decades, observational analyses of secondary health data have played an increasingly prominent role in evaluating the safety and efficacy of medical products and procedures. Research conducted using secondary health data has a number of advantages over prospective randomized clinical trials: they are less costly, can be completed in less time, and include more representative patient populations. However, misclassification of study variables remains a key concern for studies of secondary health data. Furthermore, in analyses conducted using secondary health data, accuracy of data elements may vary systematically with respect to relevant study groups (e.g. by exposure or outcome status). Despite these limitations, the growing importance of secondary data analysis is undeniable, aided by ongoing emergence of new technologies and methods intended to make it easier to collect[125,126], securely store/access[127-129], link[1,130,131], and efficiently conduct analyses[75,101,102] using aggregated data. It is important that this growth is accompanied by rigorous evaluations of the methods used for conducting health research within large databases. These evaluations have far-reaching significance, as they can inform the design of studies conducted within a wide range of clinical disciplines and data sources.

In an effort to reduce systematic variation in information accuracy, analyses conducted within secondary health databases typically assess patient histories using fixed look-back periods. Fixed look-back approaches restrict study populations to people with some minimum required period of continuous database enrollment and ignore any data outside this period. An alternate approach has been suggested in which all available

database history is considered, regardless of whether that history is available for all patients.[4,7] The primary concern with all-available data approaches is that estimates may be biased by the differential misclassification of study variables (e.g. confounders adjusted for in analysis, or eligibility criteria used to include/exclude subjects from the study population).

The current paradigm favoring use of fixed look-backs in large database research is more heavily influenced by misclassification theory than by empirical evaluations of real data. However, even in theory using fixed-look backs in database studies is an imperfect solution to the problem of differential misclassification, since variable classification is affected by a range of factors beyond just database enrollment. Furthermore, it is important that best practices in database research be informed by evaluations in real-world data. Theory alone cannot address, for example, the impact of conditioning study populations on continuous database enrollment. Furthermore, the utility of using fixed look-backs instead of all-available approaches depends on how informative the data is that they discard. If this information is especially informative, it may be worth using an all-available approach and tolerating some degree of differential information accuracy. Sholom Wacholder described such a trade-off in a series of papers[12,13,122] outlining principals for control selection in case-control studies:

> "…study designs that tolerate errors in one group so that errors are not differential should be examined carefully. Strict adherence to the principle of comparable accuracy used to ensure non-differential misclassification in choosing controls for case-control studies may not be advisable when it would require controls with as much error as cases instead of more accurate controls."[12]

6.1.1 Summary of specific aim 1

In an applied example assessing the impact of statin initiation on short-term cancer incidence within six months (a negative control) and 2-year mortality, we were able to evaluate the impact of fixed and all-available data approaches on both cohort selection and confounding adjustment. Existing evaluations of look-back approaches have focused primarily on the latter; no existing research explores how different look-back approach affect

study populations via continuous enrollment restrictions and the classification of eligibility criteria.

Our findings agree with those of other studies conducted within real-world data which show that the all-available approach does not produce superior control for confounding compared to fixed look-back approaches. However, we did observe meaningful changes in estimates when different look-backs were used to define exclusion criteria (i.e. having prior exposures or the cancer outcome). The all-available approach and the 3-year fixed approach dramatically out-performed the shorter 1-year fixed approach. This finding should motivate further research considering the role that look-backs play in selecting cohorts.

We found that the all-available approach was superior to applying a short (1-year) fixed look-back, which failed to exclude a large number of people who had unrecognized prior exposures and/or history of the cancer outcome. Point estimates produced by the three-year look-back approach were slightly less biased than the all-available approach; however, it was substantially less precise as a majority of patients did not meet the 3-year continuous enrollment requirement and were excluded. The impact of continuous enrollment restrictions and the cost of the resulting loss in precision likely varies across different database settings and study populations. Outside of the Medicare database evaluated in this study, lengthy continuous enrollment requirements may be even more restrictive, particularly when churn in database enrollment is very common (e.g. electronic health record databases or administrative claims databases for commercial or employer-supplemented insurance). By foregoing lengthy continuous enrollment requirements, the all-available approach selected broader and more clearly defined cohorts, enhancing the precision and generalizability of estimates.

6.1.2 Summary of specific aim 2

We implemented an innovative method, plasmode simulation[17,18], to study the performance of different look-back approaches in a setting which 1) reflected the complexity

real-world data structures, and 2) allowed the true effect (and thus bias in estimates) to be known. We sought to impersonate the relationships between real-world health status that confound effect estimates and the proxy variables that can be observed in an administrative claims study. To do so, we simulated exposures and outcomes as a function of real-world self-report data from the Medicare Current Beneficiary Survey (MCBS), and then proceeded to conduct cohort analyses within the linked claims using the various look-back approaches. Layering multiple real-world data sources and simulated data elements provided a detailed view of how different look-back approaches affect the validity of findings, under a range of study conditions. This approach is promising and could be useful to evaluate the impact of many study design decisions, particularly those that depend on complex data structures that are difficult to imitate in a purely simulated setting.

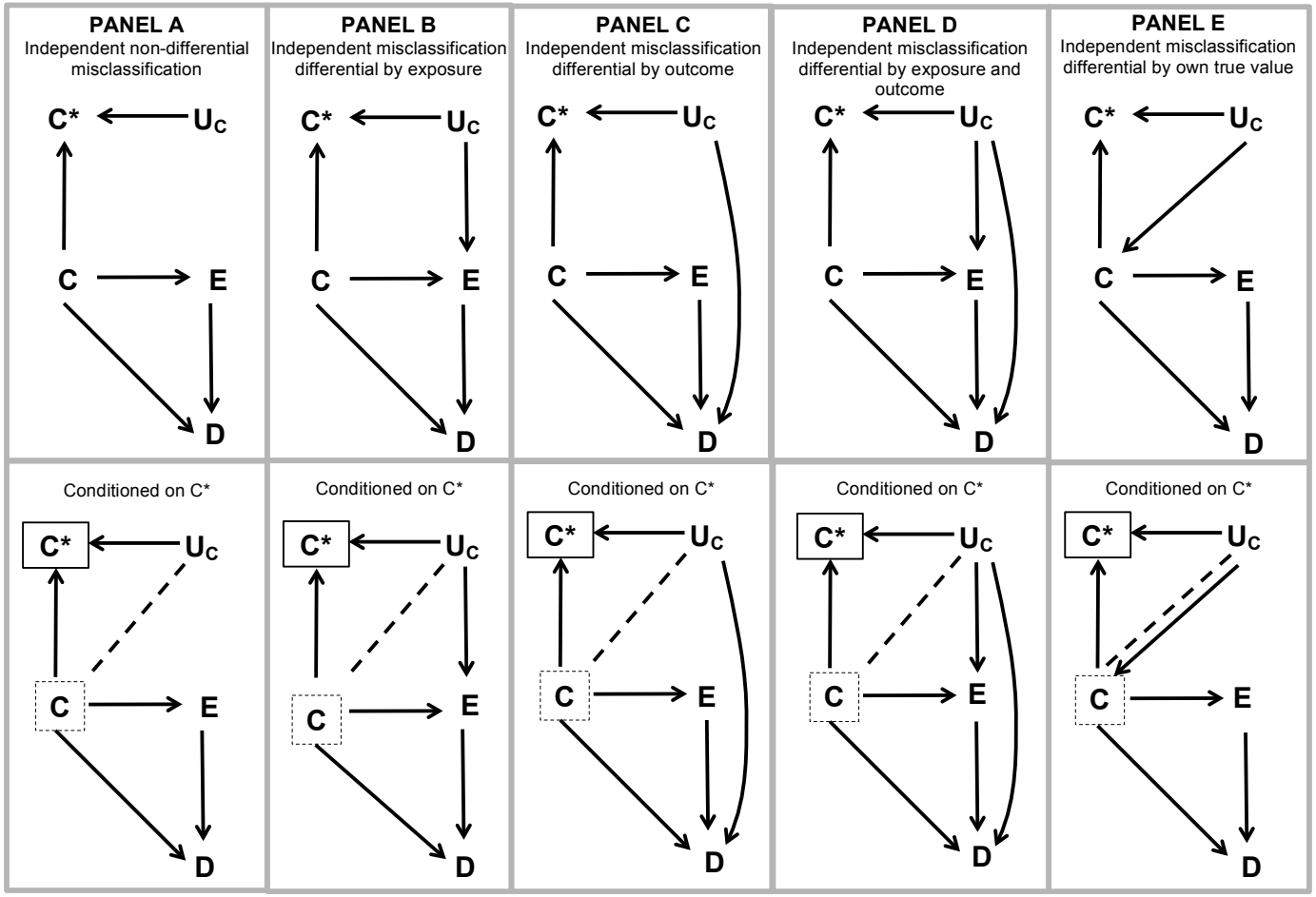
In simulations, estimates produced by the different look-back approaches were comparable in terms of bias. The all-available approach produced the least biased point-estimates. For all look-back approaches, including the all-available, estimates were marginally improved when covariates definitions were stratified by time (i.e. within the last six months and before the last six months). Cohorts selected by the all-available approach were larger and more inclusive than cohorts selected by long fixed look-backs. Compared to short fixed look-backs, the all-available approach led to cohorts that were better classified, including fewer patients who had prior exposures to statins. Findings were consistent across a range of simulated scenarios, including 1) variation in the effect of exposure (protective and harmful), 2) estimation of multiplicative effects using the rate ratio and absolute effects using the rate difference, 3) multiple sources of bias (residual confounding bias due to misclassification, prior user bias), and 3) simulation of time-stratified and time-fixed confounding.

The fact that estimates produced by the different look-back approaches largely overlapped may be partially due to the limited sample size available in the MCBS data.

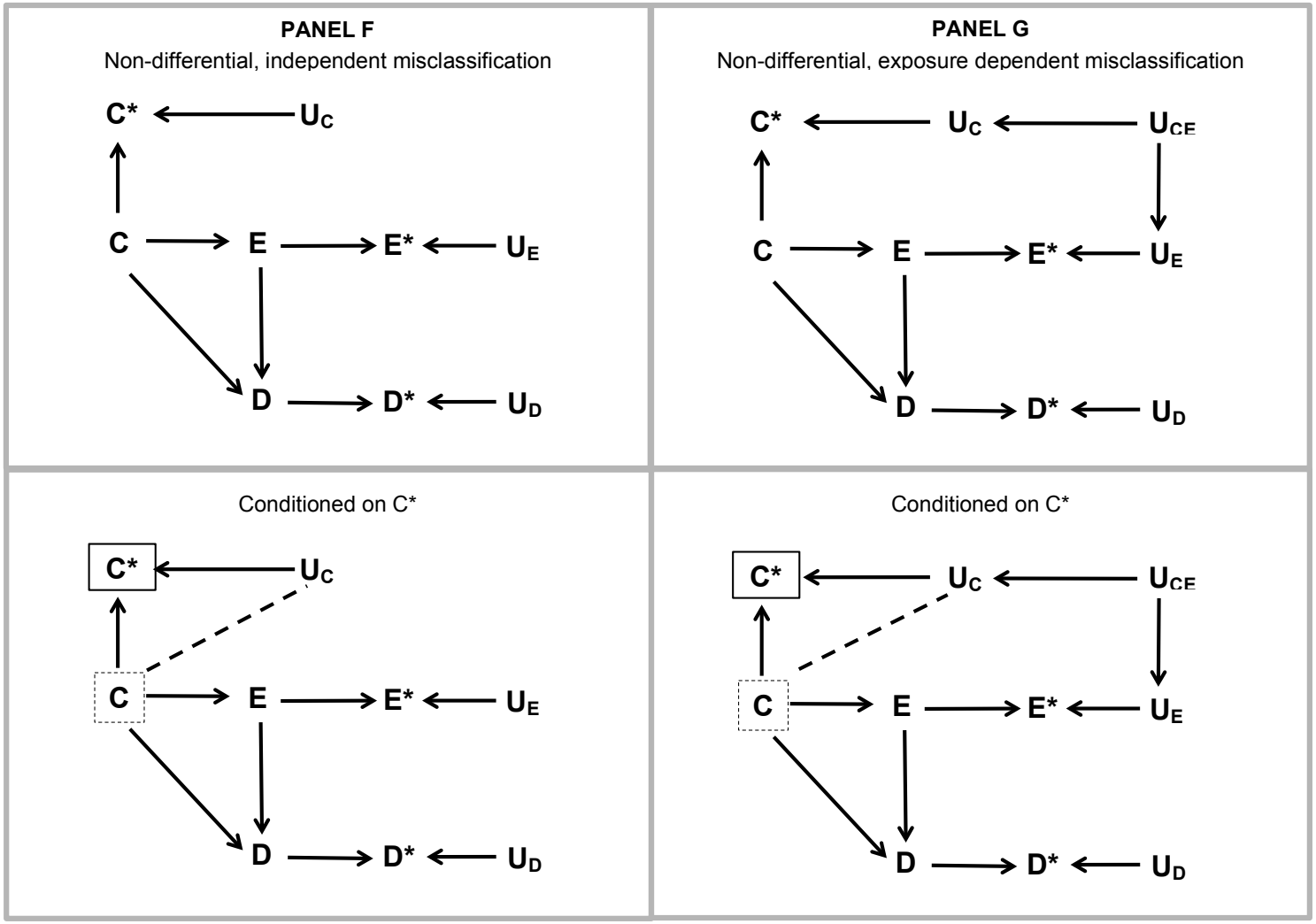
Furthermore, moderately superior point estimates produced by the all-available approach does not necessarily imply meaningfully superior performance. Regardless, the fact that all-available approaches produce similar estimates to the 3-year look-back is promising. It indicates that when we use all-available approaches, the bias caused by systematic differences in available database history is either negligible or offset by improvement in cohort selection and confounder classification. In the absence of this concern, all-available look-backs may be inherently preferable to fixed look-back approaches, since they select broader, more representative study populations.

Our findings may indicate that using all-available look-backs to classify patient histories is superior to the widely used fixed look-back approach. However, our evaluation of these look-back approaches was conducted within a narrow study setting (i.e. a Medicare claims analysis of statin exposures). Further evaluation of look-back approaches within other study populations and databases is needed before all-available approaches are needed before wide adoption can be justified. However, these findings provide strong real-world evidence that it is worthwhile for investigators to at least consider estimates produced by all-available look-backs in sub-analyses. Furthermore, we provide context to investigators seeking to understand the mechanisms through which the different look-backs may produce different estimates.

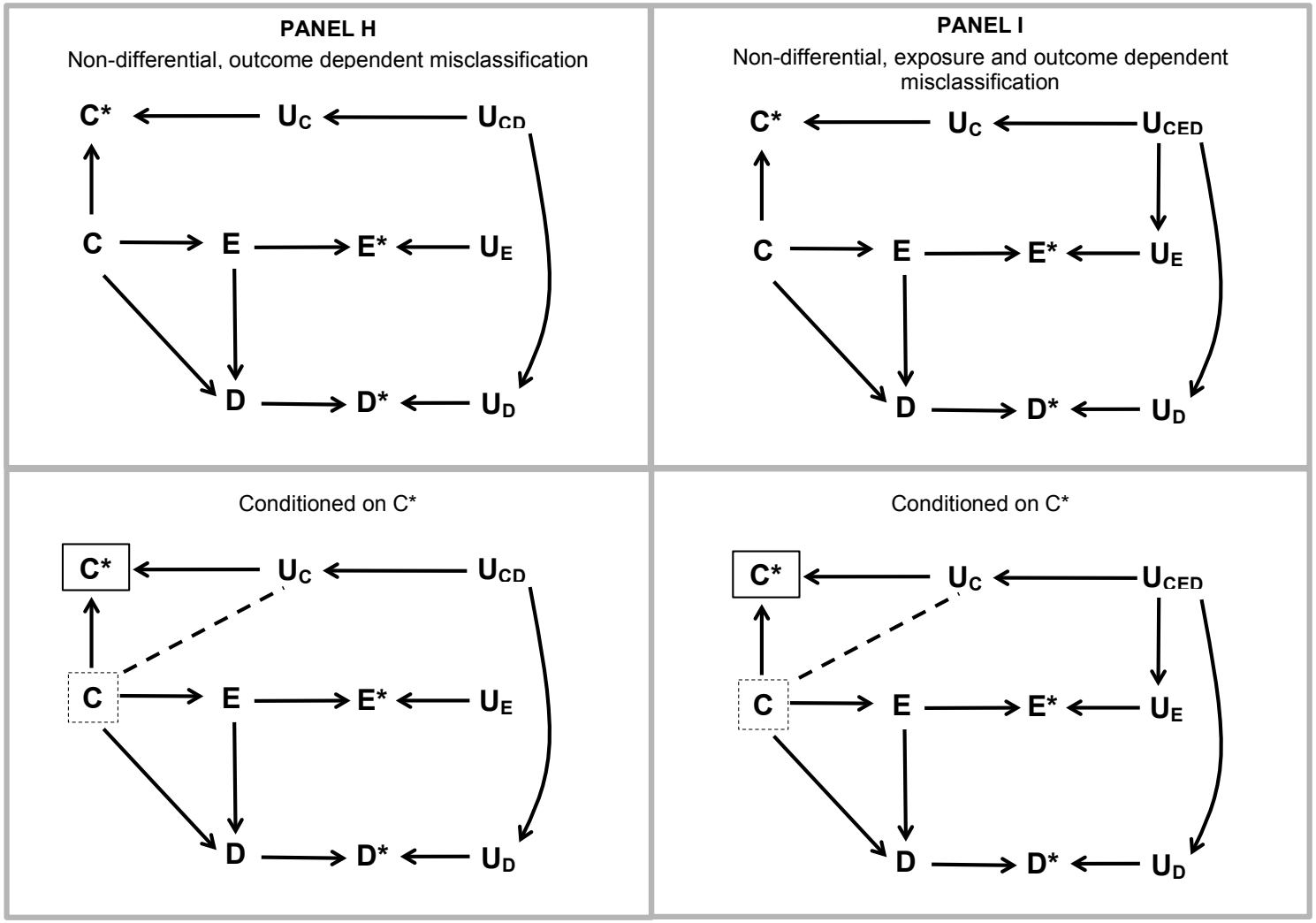
APPENDIX 2.1: DIRECTED ACYCLIC GRAPHS DEMONSTRATING INDEPENDENT AND NON-DIFFERENTIAL MISCLASSIFICATION



E = exposure, D = disease, C = confounder, C* = measured C, U_C = determinants of measurement error in C



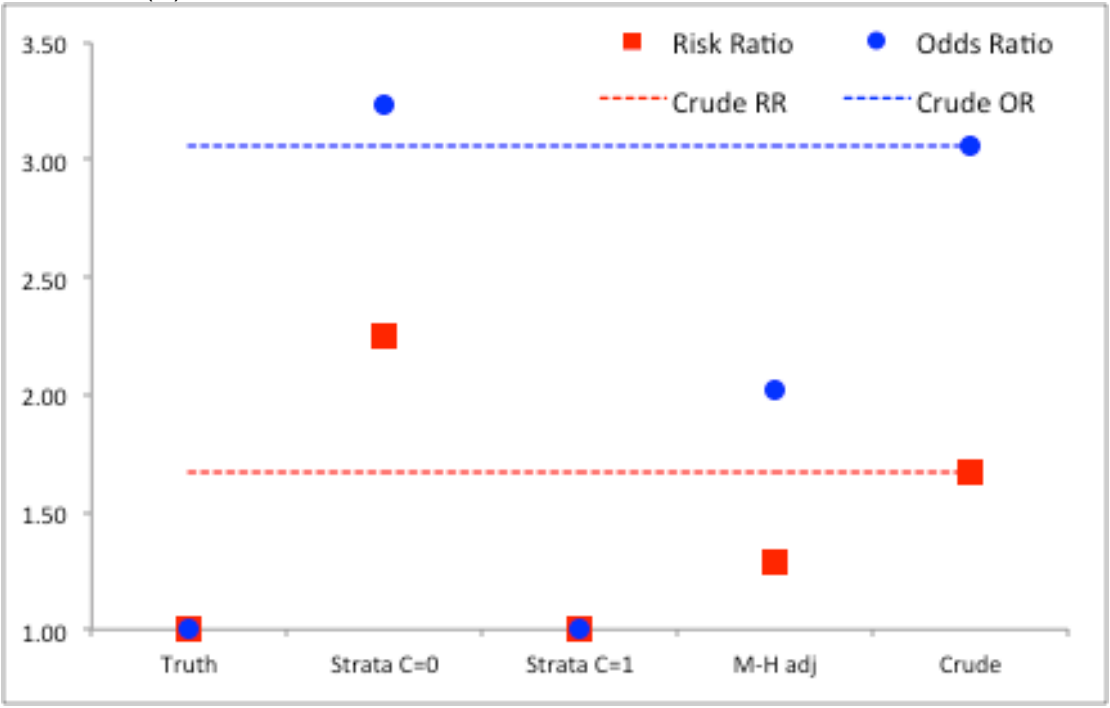
E = exposure, E* = measured E, D = disease, D* = measured D, C = confounder, C* = measured C, U_C = determinants of measurement error in C, U_E = determinants of measurement error in E, U_D = determinants of measurement error in D



E = exposure, E* = measured E, D = disease, D* = measured D, C = confounder, C* = measured C, U_C = determinants of measurement error in C, U_E = determinants of measurement error in E, U_D = determinants of measurement error in D

APPENDIX 2.2: FIGURE DISPLAYING BIAS IN EFFECT ESTIMATES AFTER STRATIFICATION ON A MISCLASSIFIED STUDY COVARIATE

Figure A2.2.1. Example stratum-specific estimates and adjusted parameters from a scenario with non-differential and independent misclassification of a dichotomous study confounder (C), relative to the crude and the true effect



[a] Parameters for the scenario displayed:
Pr(C)=70%
Specificity=1
Sensitivity=0.8
Pr(E=1|C=0)=40%
Pr(E=1|C=1)=80%
Pr(D=1|E=1,C=0) = 10%
Pr(D=1|E=0,C=0) = 10%
Pr(D=1|E=1,C=1) = 80%
Pr(D=1|E=0,C=1) = 80%

APPENDIX 4.1: AIM 1 SUPPLEMENTAL TABLES

Table A4.1.1. Description of beneficiaries who enter as both an initiator and non-initiator in cohorts formed by multiple look-back approaches.

	All-available	1-year	3-year
Total beneficiaries	497,203	581,186	210,018
Non-initiation only	425,856	494,263	191,100
Initiation only	20,371	21,715	5,769
Initiation & non-initiation	50,976	65,208	13,149
% of non-initiations	11%	12%	6%
% of initiations	71%	75%	70%
Average date gap (days)[a]	417	416	296
% of initiators with date gap:			
≤ 7 days	1%	1%	1%
8-14 days	1%	1%	2%
15-30 days	5%	5%	5%
31-90 days	11%	11%	11%
91-180 days	11%	12%	13%
181-365 days	14%	13%	15%
> 365 days	29%	32%	23%
No dual entry	29%	25%	30%

[a] The date gap describes, for a subject who entered the cohort with both an eligible statin initiation visit and an eligible non-visit, the absolute value of the time difference between the two index visit dates.

Table A4.1.2. Unadjusted distribution of baseline covariates by exposure group, for each look-back approach

	All-available			1-year fixed			3-year fixed		
Covariate description	**Initiators %**	**Non- %**	**Stdz Diff**	**Initiators %**	**Non- %**	**Stdz Diff**	**Initiators %**	**Non- %**	**Stdz Diff**
* In 6-month cancer PS model † In 2-year mortality PS model									
Continuous variables	**Mean (SD)**	**Mean (SD)**	**Stdz Diff**	**Mean (SD)**	**Mean (SD)**	**Stdz Diff**	**Mean (SD)**	**Mean (SD)**	**Stdz Diff**
Age (years) *†	74.1 (6.7)	76.2 (8.2)	-0.27	75.4 (6.6)	77.2 (8.0)	-0.25	76.7 (6.2)	78.8 (7.6)	-0.30
Inpatient stays/10 years *†	2.2 (4.5)	2.8 (5.9)	-0.12	2.7 (6.4)	3.5 (7.7)	-0.12	2.1 (3.9)	2.7 (4.9)	-0.12
Outpatient visits/year *†	7.5 (5.8)	7.1 (5.9)	0.08	9.4 (6.8)	8.7 (6.6)	0.10	8.0 (5.7)	7.8 (5.7)	0.03
SNF admits/10 years *†	0.3 (1.5)	0.6 (2.6)	-0.15	0.4 (2.5)	0.9 (3.7)	-0.14	0.3 (1.5)	0.7 (2.2)	-0.17
Unique drugs/year *†	7.1 (5.3)	7.5 (6.2)	-0.08	9.0 (5.2)	8.6 (5.4)	0.08	4.9 (2.8)	4.9 (2.9)	0.01
Demographics									
Sex: female *†	65.9%	68.4%	-0.05	66.7%	68.5%	-0.04	67.6%	71.3%	-0.08
<70 years *†	34.5%	30.0%	0.10	25.3%	23.4%	0.04	13.6%	13.5%	0.00
70 to 74 years *†	26.0%	20.3%	0.13	28.2%	21.6%	0.15	33.0%	24.2%	0.20
75 to 79 years *†	19.1%	17.6%	0.04	21.8%	19.0%	0.07	25.1%	20.7%	0.10
80 to 84 years *†	12.7%	15.2%	-0.07	15.1%	17.0%	-0.05	16.9%	18.8%	-0.05
85 to 89 years *†	6.0%	10.5%	-0.16	7.5%	11.9%	-0.15	8.5%	13.7%	-0.17
90 to 94 years *†	1.5%	4.9%	-0.19	1.9%	5.5%	-0.19	2.5%	6.8%	-0.20
95 to 99 years *†	0.2%	1.3%	-0.13	0.3%	1.5%	-0.13	0.4%	1.9%	-0.15
100+ years *†		0.2%	-0.05			-0.05			-0.06
Year: 2007 *†	16.3%	35.4%	-0.45						
Year: 2008 *†	23.2%	19.6%	0.09	22.2%	40.4%	-0.40			
Year: 2009 *†	18.6%	12.9%	0.16	21.6%	17.9%	0.09			
Year: 2010 *†	14.9%	11.0%	0.12	19.6%	15.0%	0.12	35.7%	54.8%	-0.39
Year: 2011 *†	13.8%	10.5%	0.10	18.7%	13.7%	0.14	33.3%	24.4%	0.20
Year: 2012 *†	13.2%	10.6%	0.08	17.9%	13.1%	0.13	31.0%	20.8%	0.23
Race: White *†	76.2%	80.6%	-0.11	77.1%	81.9%	-0.12	78.4%	81.6%	-0.08

	All-available			1-year fixed			3-year fixed		
Covariate description	**Initiators %**	**Non- %**	**Stdz Diff**	**Initiators %**	**Non- %**	**Stdz Diff**	**Initiators %**	**Non- %**	**Stdz Diff**
Race: Black *†	10.4%	9.6%	0.03	9.7%	8.7%	0.04	9.1%	8.6%	0.02
Race: Hispanic *†	5.4%	3.6%	0.08	5.4%	3.6%	0.09	4.9%	3.5%	0.07
Race: Asian *†	4.7%	3.4%	0.06	4.6%	3.3%	0.07	4.4%	3.7%	0.04
Race: North American Native *†	0.7%	0.5%	0.02	0.6%	0.5%	0.01	0.7%	0.6%	0.01
Race: Other *†	2.4%	1.9%	0.03	2.3%	1.8%	0.04	2.4%	1.9%	0.03
Race: Unknown *†	0.3%	0.3%	0.00	0.2%	0.2%	0.00	0.1%	0.1%	-0.01
Inclusion criteria									
Incl: AMI	2.7%	1.5%	0.08	2.4%	1.5%	0.06	3.2%	1.4%	0.12
AMI (6+ mos)	1.8%	1.4%	0.03	0.6%	0.7%	-0.01	1.7%	1.9%	-0.02
Incl: Angioplasty	0.4%	0.3%	0.03	0.4%	0.2%	0.03	0.4%	0.2%	0.03
Angioplasty (6+ mos)	0.5%	0.3%	0.04	0.1%	0.1%	0.01	0.4%	0.3%	0.02
Incl: CABG	0.5%	0.1%	0.07	0.4%	0.1%	0.07	0.5%	0.1%	0.08
CABG (6+ mos)	0.4%	0.2%	0.04	0.1%	0.1%	0.01	0.2%	0.2%	0.01
Incl: Diabetes *	71.3%	58.1%	0.28	71.3%	58.3%	0.28	66.1%	57.7%	0.17
Diabetes (6+ mos) *	59.7%	38.0%	0.44	55.0%	38.2%	0.34	59.4%	49.3%	0.20
Incl: Endarterectomy †	3.1%	7.2%	-0.19	3.2%	6.9%	-0.17	3.6%	7.2%	-0.16
Endarterectomy (6+ mos)	4.9%	2.9%	0.10	1.8%	1.5%	0.02	5.6%	5.6%	0.00
Incl: PTCA	1.5%	0.4%	0.12	1.4%	0.4%	0.11	1.5%	0.2%	0.13
PTCA (6+ mos)	1.0%	0.5%	0.06	0.3%	0.2%	0.01	0.7%	0.6%	0.02
Incl: Stable angina	7.7%	5.9%	0.07	7.7%	5.9%	0.07	8.1%	5.2%	0.12
Stable angina (6+ mos)	6.9%	4.0%	0.13	2.9%	2.2%	0.04	7.1%	6.1%	0.04
Incl: Stent	2.0%	0.5%	0.14	1.9%	0.5%	0.13	2.3%	0.4%	0.17
Stent (6+ mos)	1.2%	0.6%	0.07	0.2%	0.2%	0.02	0.7%	0.6%	0.02
Incl: Stroke †	29.7%	35.4%	-0.12	30.8%	36.4%	-0.12	34.8%	36.9%	-0.04
Stroke (6+ mos) *	23.8%	18.0%	0.14	13.3%	12.4%	0.03	29.0%	29.3%	-0.01
Exclusion criteria									

	All-available			1-year fixed			3-year fixed		
	Initiators	Non-		Initiators	Non-		Initiators	Non-	
Covariate description	**%**	**%**	**Stdz Diff**	**%**	**%**	**Stdz Diff**	**%**	**%**	**Stdz Diff**
Acute kidney injury (6+ mos)	2.0%	2.6%	-0.04	0.8%	1.5%	-0.07	2.6%	4.2%	-0.09
Chronic kidney disease (6+ mos) †	3.8%	4.6%	-0.04	1.9%	3.6%	-0.10	5.3%	7.9%	-0.11
Chronic liver disease (6+ mos) *	5.4%	5.2%	0.01	1.8%	3.1%	-0.08	6.3%	7.6%	-0.05
Cyclosporine (6+ mos)	0.8%	0.8%	-0.01	0.4%	0.7%	-0.04	1.6%	1.9%	-0.03
Fibrate (6+ mos)	1.8%	1.3%	0.04	1.1%	1.4%	-0.03	3.0%	2.6%	0.02
Heart failure (6+ mos)	9.1%	11.7%	-0.09	3.7%	8.0%	-0.19	10.4%	15.8%	-0.16
Inflammatory muscle (6+ mos)	0.7%	0.9%	-0.02	0.2%	0.4%	-0.04	0.7%	1.2%	-0.05
Niacin (6+ mos)	0.6%	0.4%	0.04	0.3%	0.4%	-0.02	1.0%	0.9%	0.01
Health behaviors									
Obesity †	12.6%	8.2%	0.14	7.9%	6.0%	0.07	13.4%	10.8%	0.08
Smoking *†	14.1%	10.8%	0.10	9.0%	7.8%	0.04	15.8%	13.0%	0.08
Alcohol use	1.9%	2.2%	-0.02	0.9%	1.4%	-0.05	2.0%	2.5%	-0.04
Substance abuse *†	9.8%	7.8%	0.07	6.1%	5.4%	0.03	10.1%	8.6%	0.05
Comorbidities									
Anemia *†	31.9%	33.3%	-0.03	23.8%	28.0%	-0.09	37.0%	42.8%	-0.12
Angina	6.2%	3.5%	0.13	3.8%	2.2%	0.09	6.2%	4.0%	0.10
Angiography †	8.7%	4.0%	0.19	5.3%	2.6%	0.14	8.8%	4.7%	0.16
Arterial embolism and thrombosis	2.0%	1.4%	0.04	1.1%	0.9%	0.02	2.2%	1.7%	0.03
Pulmonary circulation disorders †	4.3%	4.3%	0.00	2.8%	3.2%	-0.03	5.5%	6.2%	-0.03
Peripheral vascular disease †	28.2%	25.1%	0.07	21.5%	21.3%	0.01	32.5%	32.8%	-0.01
Arthritis *	64.6%	63.7%	0.02	51.4%	55.2%	-0.08	72.0%	75.8%	-0.09
Osteoarthritis †	46.6%	45.7%	0.02	34.8%	37.6%	-0.06	53.4%	57.3%	-0.08
Osteoporosis	31.7%	32.2%	-0.01	22.5%	25.3%	-0.07	38.3%	42.6%	-0.09
Rheumatoid arthritis *	10.0%	9.7%	0.01	6.8%	7.7%	-0.03	11.5%	12.9%	-0.04
Asthma †	12.7%	11.8%	0.03	8.9%	9.3%	-0.01	14.0%	14.6%	-0.02

	All-available			1-year fixed			3-year fixed		
	Initiators	Non-		Initiators	Non-		Initiators	Non-	
Covariate description	%	%	Stdz Diff	%	%	Stdz Diff	%	%	Stdz Diff
COPD *†	30.6%	29.8%	0.02	22.8%	24.7%	-0.04	33.1%	35.5%	-0.05
Atrial fibrillation †	14.8%	16.4%	-0.04	12.2%	14.8%	-0.08	17.5%	20.9%	-0.09
Dementia *†	15.1%	20.3%	-0.14	11.7%	18.1%	-0.18	18.2%	25.5%	-0.18
Depression	7.8%	9.9%	-0.07	5.5%	7.8%	-0.09	8.8%	12.2%	-0.11
Psychiatric disorder †	27.5%	30.8%	-0.07	21.1%	26.9%	-0.14	31.9%	38.2%	-0.13
Diabetes (complications)	19.7%	11.9%	0.21	16.1%	11.4%	0.14	19.8%	16.1%	0.10
Dialysis	0.1%	0.1%	0.00	0.0%	0.0%	-0.01	0.1%	0.1%	0.00
ESRD	0.2%	0.2%	-0.02	0.0%	0.1%	-0.03	0.2%	0.3%	-0.02
GI bleed *	7.3%	7.0%	0.01	3.2%	4.1%	-0.04	7.5%	9.1%	-0.06
Gout	6.3%	5.4%	0.03	4.4%	4.3%	0.01	7.6%	7.6%	0.00
HIV	0.1%	0.1%	0.00	0.1%	0.1%	0.00	0.1%	0.1%	0.00
Hyperlipidemia *†	88.1%	65.2%	0.56	83.6%	59.8%	0.55	88.1%	72.5%	0.40
Lipid abnormality	88.9%	66.3%	0.56	84.5%	60.8%	0.55	88.8%	73.4%	0.40
Hypertension	90.9%	84.9%	0.19	88.6%	82.7%	0.17	92.1%	89.9%	0.08
Hyperthyroidism	3.9%	3.5%	0.02	2.2%	2.3%	-0.01	4.3%	4.5%	-0.01
Inflammatory bowel †	4.1%	4.1%	0.00	2.2%	2.6%	-0.03	4.6%	5.5%	-0.04
Ischemic heart disease	37.4%	29.9%	0.16	33.1%	27.4%	0.12	39.7%	34.4%	0.11
Paralysis †	4.1%	4.2%	-0.01	3.1%	3.4%	-0.02	4.7%	5.3%	-0.03
Parkinsons	1.5%	2.8%	-0.09	1.4%	2.7%	-0.09	1.9%	3.4%	-0.10
Peptic ulcer	5.1%	4.7%	0.02	2.5%	2.9%	-0.02	5.3%	6.0%	-0.03
Lupus	0.7%	0.7%	-0.01	0.4%	0.5%	-0.02	0.7%	1.0%	-0.03
Podiatric	14.9%	15.1%	-0.01	10.9%	12.5%	-0.05	17.4%	20.8%	-0.09
Psoriasis	2.0%	1.6%	0.03	1.2%	1.2%	0.00	2.2%	2.1%	0.01
Rheumatic heart disease	6.8%	5.9%	0.04	4.2%	4.0%	0.01	8.0%	7.8%	0.01
Sepsis †	20.4%	20.2%	0.00	11.4%	14.2%	-0.08	23.3%	27.7%	-0.10
Vertigo †	27.0%	23.5%	0.08	17.3%	17.3%	0.00	31.6%	31.1%	0.01

	All-available			1-year fixed			3-year fixed		
	Initiators	Non-		Initiators	Non-		Initiators	Non-	
Covariate description	**%**	**%**	**Stdz Diff**	**%**	**%**	**Stdz Diff**	**%**	**%**	**Stdz Diff**
VTE *†	3.4%	3.7%	-0.02	2.2%	2.8%	-0.04	4.1%	5.0%	-0.04
Health screening & preventive									
Apolipoprotein assay	1.4%	0.7%	0.07	1.2%	0.6%	0.06	1.8%	1.1%	0.07
Lipid panel †	92.6%	76.0%	0.47	86.5%	66.7%	0.48	93.9%	83.8%	0.32
Bone density	0.5%	0.5%	0.01	0.1%	0.2%	-0.01	0.4%	0.4%	0.00
Cancer screening *†	55.8%	46.0%	0.20	37.9%	34.4%	0.07	59.2%	53.9%	0.11
Cardiac stress test *†	27.7%	17.8%	0.24	16.0%	11.3%	0.14	29.7%	21.8%	0.18
Colonoscopy *†	9.1%	7.2%	0.07	3.1%	3.3%	-0.01	9.0%	8.8%	0.01
Echocardiograph †	40.6%	32.6%	0.17	28.2%	24.1%	0.09	47.3%	41.7%	0.11
Fecal occult blood testing †	7.4%	6.4%	0.04	3.0%	2.6%	0.03	6.8%	5.8%	0.04
hs-CRP *†	15.7%	12.4%	0.10	10.7%	9.3%	0.04	18.6%	17.2%	0.04
PSA testing *	25.8%	20.5%	0.13	20.0%	17.0%	0.08	26.2%	22.1%	0.10
Flu vaccination	57.3%	51.5%	0.12	51.1%	51.2%	0.00	65.8%	67.3%	-0.03
Pnemonia vaccination	19.2%	14.0%	0.14	9.1%	8.0%	0.04	20.7%	18.8%	0.05
Medications									
ARB †	23.4%	18.6%	0.12	22.5%	19.0%	0.09	26.3%	23.3%	0.07
Beta blockers	41.2%	35.8%	0.11	41.4%	37.8%	0.07	45.9%	43.5%	0.05
Biguanide	35.0%	17.4%	0.41	32.9%	18.3%	0.34	33.7%	21.2%	0.28
Calcium channel blockers	31.3%	27.4%	0.09	31.2%	28.6%	0.06	36.9%	35.3%	0.03
Creatinine *	3.5%	2.9%	0.04	1.9%	1.9%	0.00	3.6%	3.9%	-0.02
Diuretics †	41.2%	37.2%	0.08	38.0%	37.2%	0.02	46.7%	46.7%	0.00
ACE Inhibitor	45.9%	33.6%	0.25	41.9%	33.4%	0.18	48.1%	40.6%	0.15
Thiazide †	30.7%	24.5%	0.14	26.9%	23.4%	0.08	35.3%	31.3%	0.09
Sulfonylurea *†	21.2%	12.5%	0.23	19.7%	13.0%	0.18	20.2%	14.9%	0.14
Thiazolidinedione	11.9%	6.0%	0.21	8.9%	5.9%	0.12	9.2%	7.0%	0.08
Insulin *†	7.4%	4.4%	0.12	7.8%	5.2%	0.11	7.1%	5.7%	0.06

	All-available			1-year fixed			3-year fixed		
Covariate description	**Initiators %**	**Non- %**	**Stdz Diff**	**Initiators %**	**Non- %**	**Stdz Diff**	**Initiators %**	**Non- %**	**Stdz Diff**
DPP-4 Inhibitors	3.8%	1.4%	0.15	4.2%	2.0%	0.13	5.3%	3.0%	0.12
Frailty indicators									
Ambulatory life support †	19.6%	22.8%	-0.08	12.4%	17.6%	-0.15	23.0%	30.0%	-0.16
Difficulty walking	20.5%	23.5%	-0.07	14.5%	19.4%	-0.13	25.4%	32.2%	-0.15
HF Frailty	18.5%	19.5%	-0.02	10.5%	14.2%	-0.11	21.2%	25.3%	-0.10
Home hospital bed	1.5%	2.4%	-0.07	1.0%	2.0%	-0.08	1.6%	3.1%	-0.10
Home oxygen *†	3.7%	4.2%	-0.02	3.4%	4.2%	-0.04	4.3%	5.3%	-0.05
Weakness †	11.6%	14.3%	-0.08	7.9%	11.8%	-0.13	15.5%	21.1%	-0.15
Wheelchair †	3.3%	4.5%	-0.06	2.1%	3.5%	-0.09	3.4%	5.7%	-0.11
Rehab	13.3%	14.3%	-0.03	7.6%	9.9%	-0.08	15.1%	18.9%	-0.10
Utilization variables (categorical)									
HS = 0 *†	66.9%	66.6%	0.01	80.7%	76.3%	0.11	63.8%	59.1%	0.10
HS: Non-0 P0-P20 *†	9.0%	6.0%	0.11						
HS: Non-0 P20-P40 *†	8.5%	6.5%	0.08						
HS: Non-0 P40-P60 *†	3.9%	4.6%	-0.04				21.7%	21.9%	0.00
HS: Non-0 P60-P80 *†	7.5%	8.8%	-0.05	14.1%	16.4%	-0.06	8.4%	10.0%	-0.06
HS: Non-0 P80-P100 *†	4.3%	7.4%	-0.13	5.2%	7.3%	-0.09	6.1%	9.1%	-0.11
OutptVisits: Non-0 P0-P20 *†	15.4%	20.7%	-0.14	15.8%	20.2%	-0.12	16.8%	18.1%	-0.04
OutptVisits: Non-0 P20-P40 *†	19.9%	20.0%	0.00	16.6%	17.3%	-0.02	19.0%	19.4%	-0.01
OutptVisits: Non-0 P40-P60 *†	20.9%	19.1%	0.04	22.8%	22.2%	0.02	20.8%	20.4%	0.01
OutptVisits: Non-0 P60-P80 *†	22.3%	20.1%	0.05	20.6%	19.0%	0.04	22.7%	22.1%	0.01
OutptVisits: Non-0 P80-P100 *†	21.5%	20.1%	0.04	24.2%	21.3%	0.07	20.7%	20.0%	0.02
SNF = 0 *†	93.9%	90.6%	0.13	96.7%	93.6%	0.14	92.7%	87.6%	0.17
SNF: Non-0 P0-P20 *†	1.8%	1.7%	0.01						
SNF: Non-0 P20-P40 *†	1.6%	1.8%	-0.02						

	All-available			1-year fixed			3-year fixed		
Covariate description	Initiators %	Non- %	Stdz Diff	Initiators %	Non- %	Stdz Diff	Initiators %	Non- %	Stdz Diff
SNF: Non-0 P40-P60 *†	1.2%	1.9%	-0.06						
SNF: Non-0 P60-P80 *†	0.8%	2.0%	-0.10	2.7%	4.8%	-0.11	5.3%	8.2%	-0.12
SNF: Non-0 P80-P100 *†	0.7%	2.0%	-0.11	0.7%	1.6%	-0.08	2.0%	4.3%	-0.13
UniqueRx: Non-0 P0-P20 *†	18.5%	20.2%	-0.04	11.4%	15.8%	-0.13	15.5%	16.2%	-0.02
UniqueRx: Non-0 P20-P40 *†	20.8%	18.6%	0.05	16.1%	16.4%	-0.01	20.4%	20.7%	-0.01
UniqueRx: Non-0 P40-P60 *†	22.3%	20.7%	0.04	25.6%	24.5%	0.02	21.0%	20.1%	0.02
UniqueRx: Non-0 P60-P80 *†	20.8%	19.9%	0.02	25.3%	23.0%	0.06	21.6%	21.7%	0.00
UniqueRx: Non-0 P80-P100 *†	17.6%	20.6%	-0.08	21.6%	20.2%	0.03	21.5%	21.1%	0.01

Table A4.1.3. Cohort sizes, outcome frequencies, and hazard ratios (crude and SMRW-adjusted) for active-comparator sub-analyses

Look-back approach	Cohort size (N)		Outcome frequency		Hazard ratio (95% CI)			
					6-month cancer		2-year mortality	
	N_{Total}[a]	$N_{HiPotent}$[a]	$\%_{Can}$	$\%_{Death}$	Crude	SMRW	Crude	SMRW
1-year fixed	86,923	18,182	1.1%	4.2%	0.90 (0.76-1.06)	0.89 (0.75-1.05)	0.90 (0.82-0.97)	0.93 (0.85-1.01)
3-year fixed	18,918	3,426	1.0%	3.9%	1.07 (0.74-1.54)	1.09 (0.75-1.59)	1.02 (0.84-1.23)	1.06 (0.87-1.30)
All-available	71,347	13,607	1.0%	4.0%	1.00 (0.82-1.21)	0.99 (0.81-1.20)	0.94 (0.85-1.03)	0.95 (0.86-1.05)

[a] These counts denote unique observations in the dataset. Patients who enter the cohort twice for eligible initiations of both high- and low-potency statins are counted twice in the N_{total} statistic (one for each exposure). Since they cannot appear twice in the same exposure group, the $N_{HiPotent}$ statistic denotes counts of unique patients.

[b] The propensity score models in the active comparator analyses adjusted for the same variables as the primary analysis (see footnote in Table 4.1).

APPENDIX 4.2: PROPENSITY SCORE VARIABLE SELECTION

A4.2.1 Methods

We created two propensity score models, one for the 6-month cancer outcome and the other for the 2-year mortality outcome. We adjusted for the same set of variables across the different look-back approaches. Propensity score models included all variables that were risk factors for the corresponding outcome. We defined risk factors as any variable that was 1) present in at least 1.5% of all exposed and unexposed patients across all look-back approaches and also 2) had a significant association with the outcome with magnitude > 1.10 among unexposed patients within at least one of the look-back approaches. The associations between covariates and outcomes were quantified using treatment group specific hazard ratios, estimated using multivariate Cox models accounting for competing risk of mortality.

A4.2.2 Results

Table A4.1.2 presents the prevalence (by exposure group) of all variables considered for inclusion in propensity score models (for each look-back approach) and also describes which variables were ultimately included in models. As shown in Figure 4.4, for all look-back approaches, non-initiators tended to be older with greater baseline utilization of medications and skilled nursing facilities (SNF), more frequent hospital admissions, and greater prevalence of dementia. Initiators, meanwhile, were more likely to have claims for routine screenings and surveillance. Also, initiators were more likely to be included in the cohort due to a recent history of diabetes (the most chronic / least acute of all the Heart Protection Study eligibility criteria). Finally, as expected, initiators are dramatically more likely to have a recorded diagnosis of hyperlipidemia. Figure A4.3.4 and Figure A4.3.5 display propensity score distributions produced using each look-back approach for the cancer and mortality outcomes, respectively.

APPENDIX 4.3: AIM 1 SUPPLEMENTAL FIGURES

Figure A4.3.1. Flow chart describing the selection of the source data and cohorts using each look-back approach

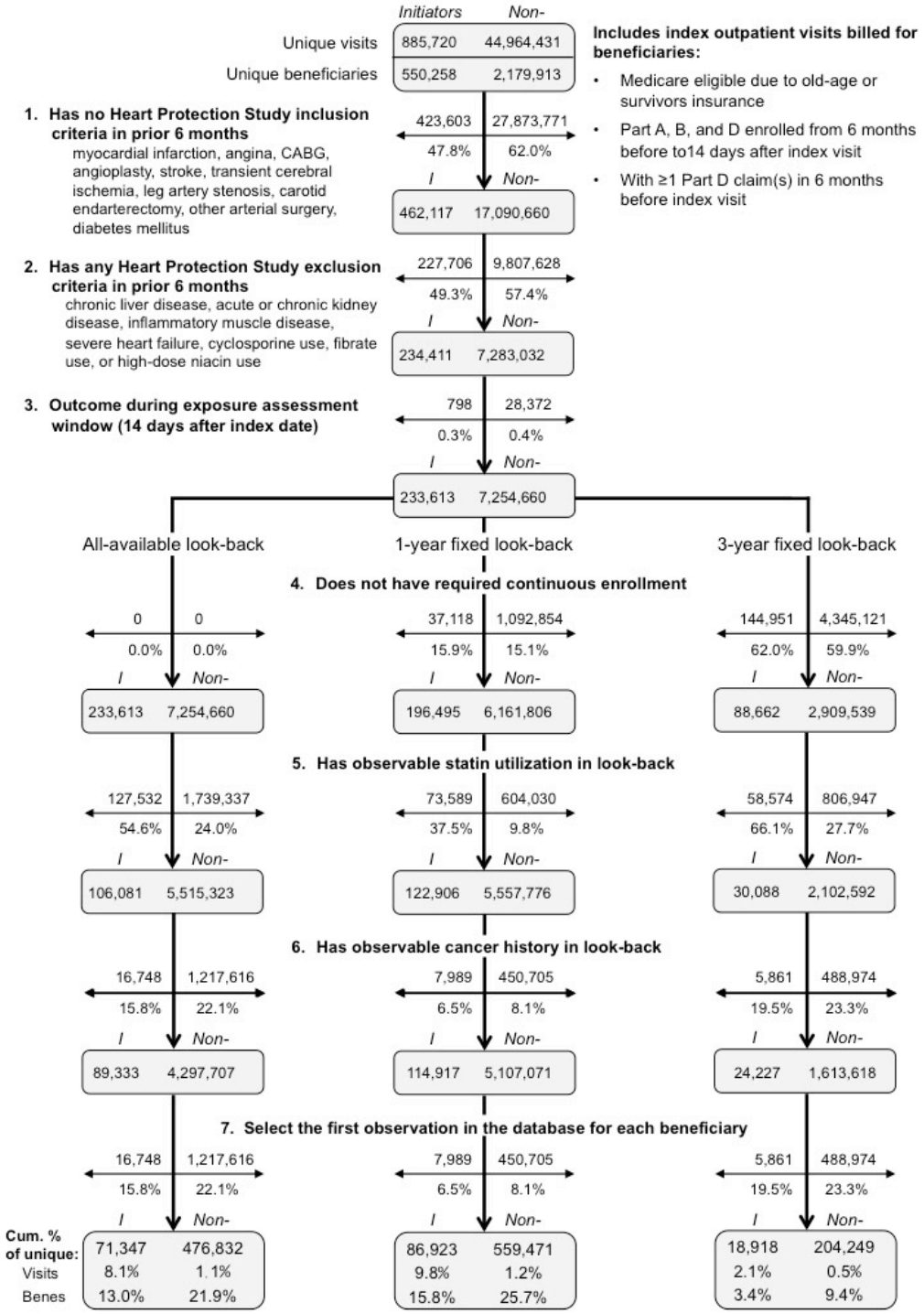


Figure A4.3.2. Box-and-whisker plots displaying the amount of observable Part A, B, and D history at baseline using the all-available approach, contrasting A) initiators vs. non-initiators, B) patients with and without cancer outcomes within six months, C) initiators of high-potency vs. low-potency statins, D) patients who died within 2-years vs. those who did not

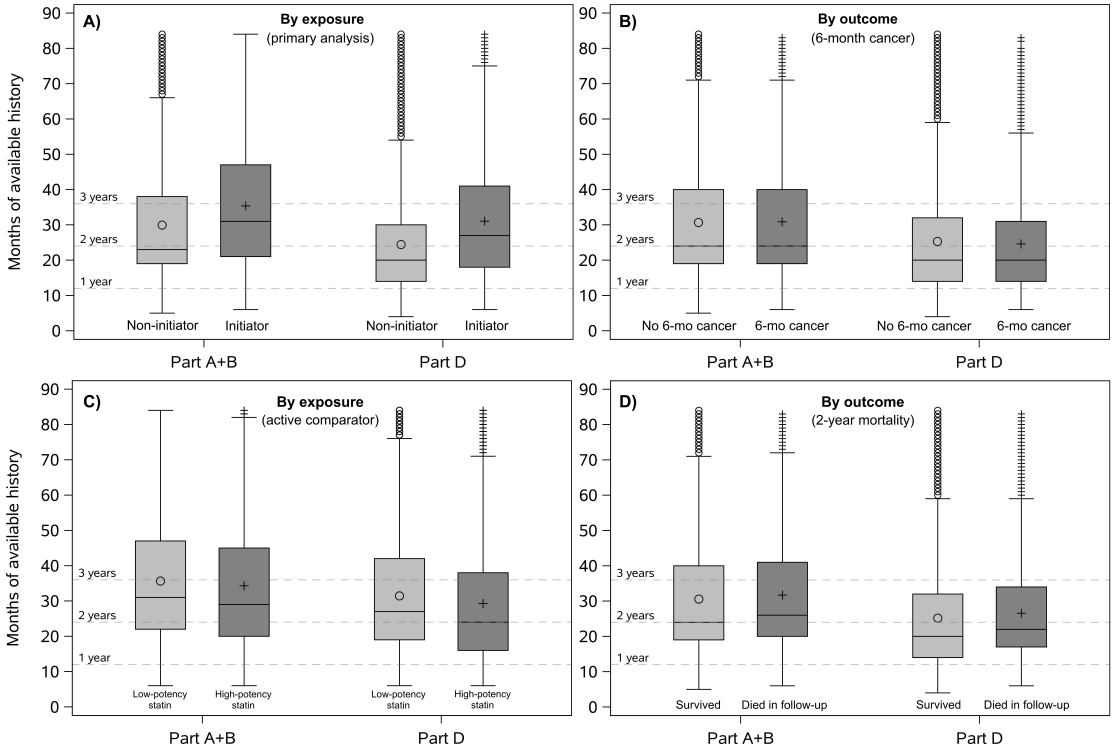


Figure A4.3.3. Proportion of the 1-year fixed and 3-year fixed cohort with observable history in the database of A) statin use and B) cancer for the 1-year and 3-year look-back approaches in the active comparator sub-analyses

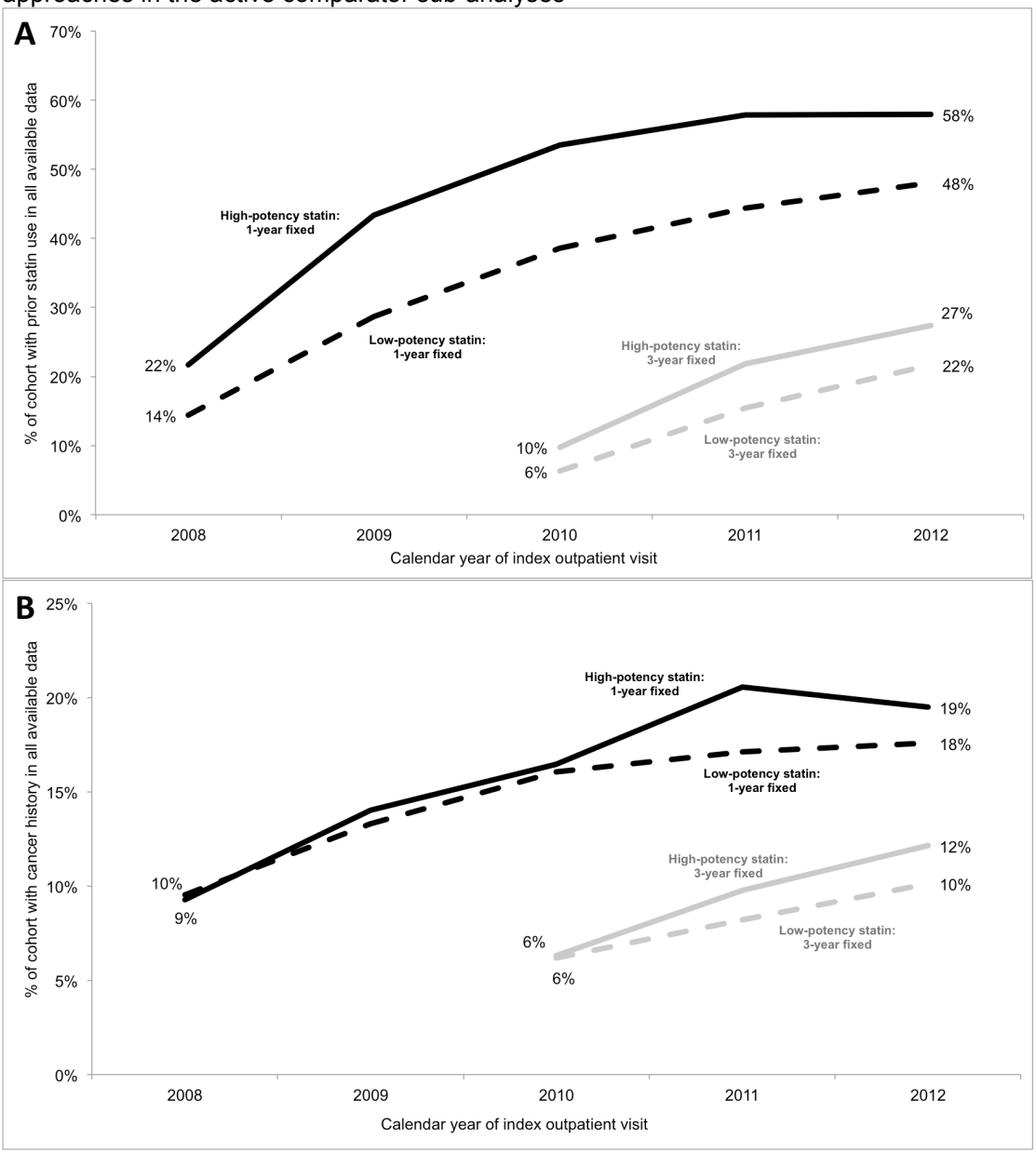


Figure A4.3.4. Propensity score distributions for 6-month cancer outcome, using different look-back approaches

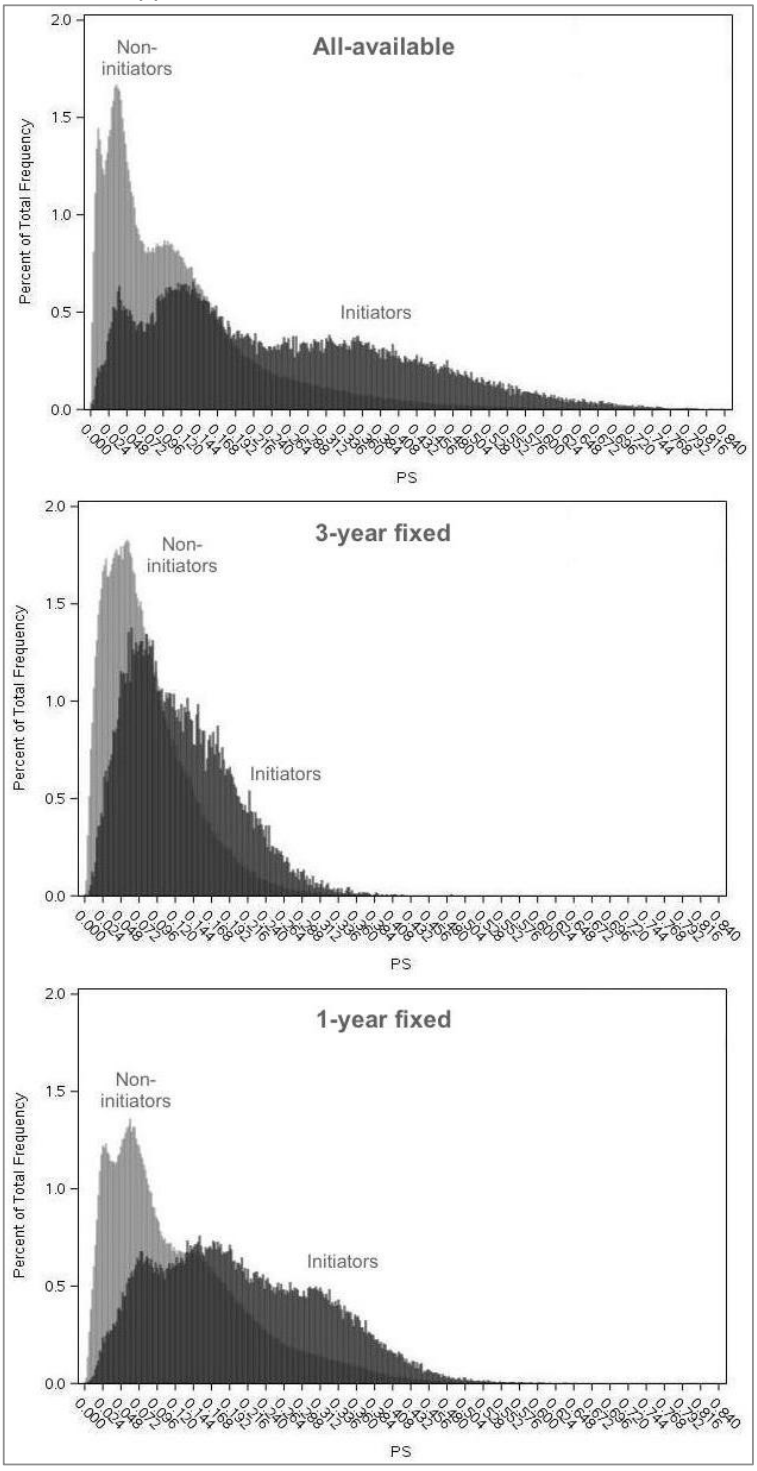


Figure A4.3.5. Propensity score distributions for the 2-year mortality outcome, using different look-back approaches

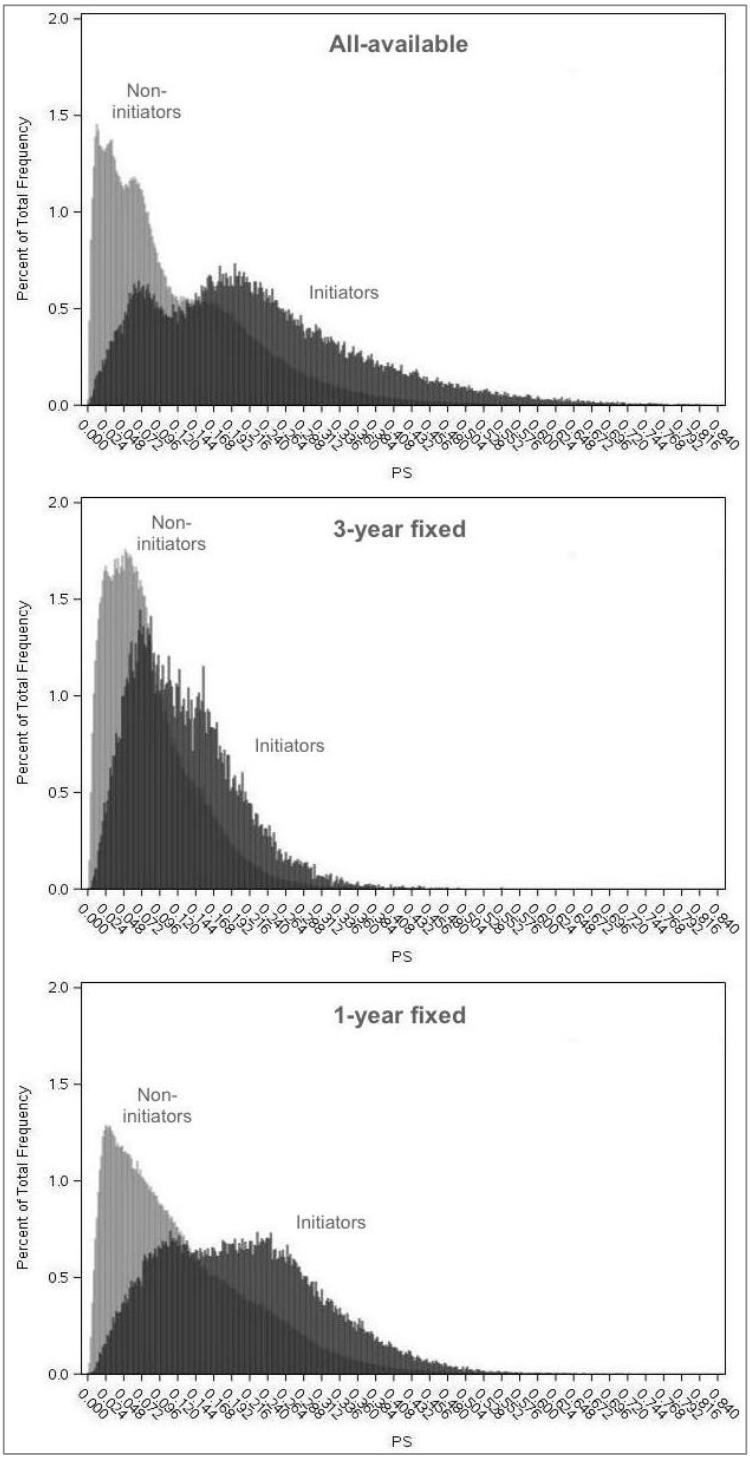


Figure A4.3.6. Crude cumulative incidence curves and risk difference at six months for incident cancer

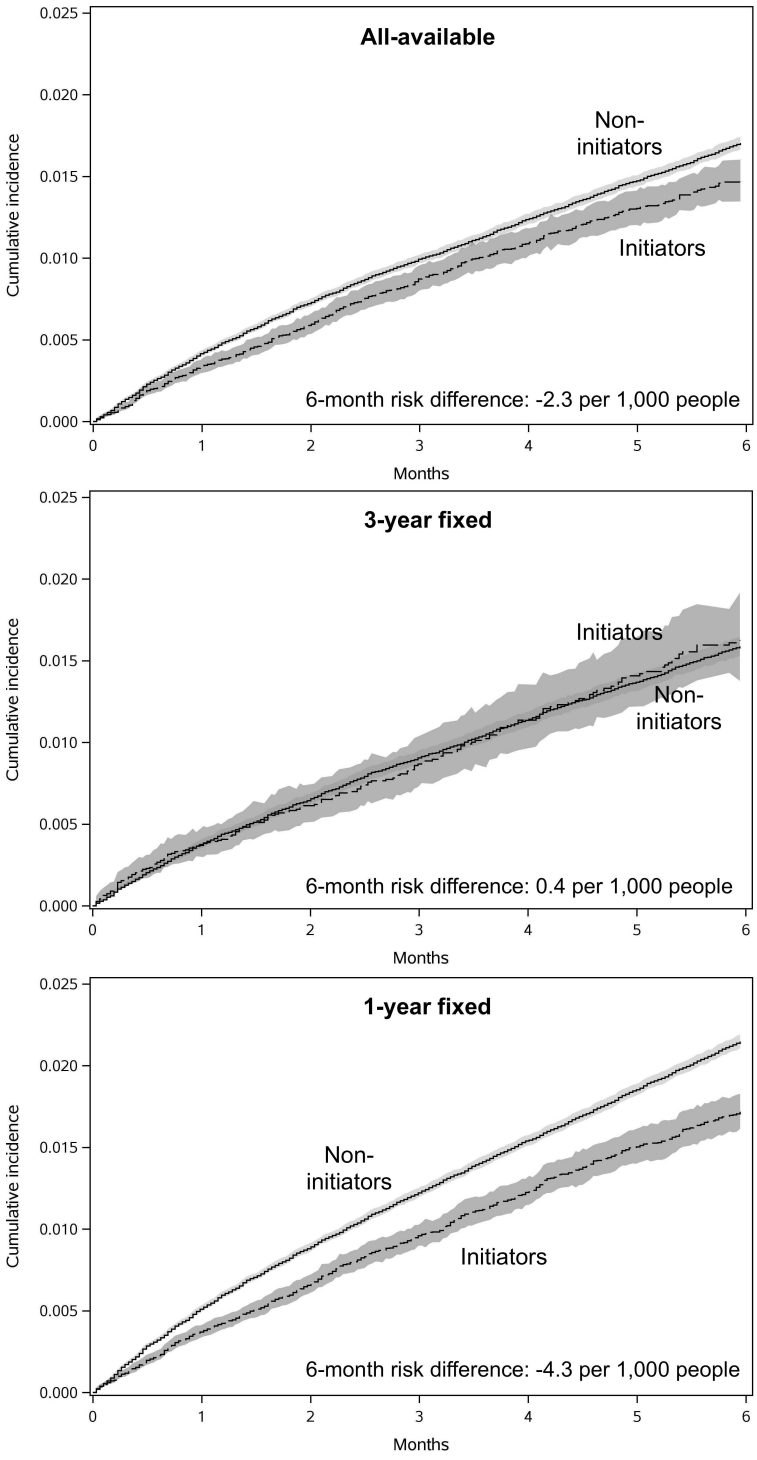
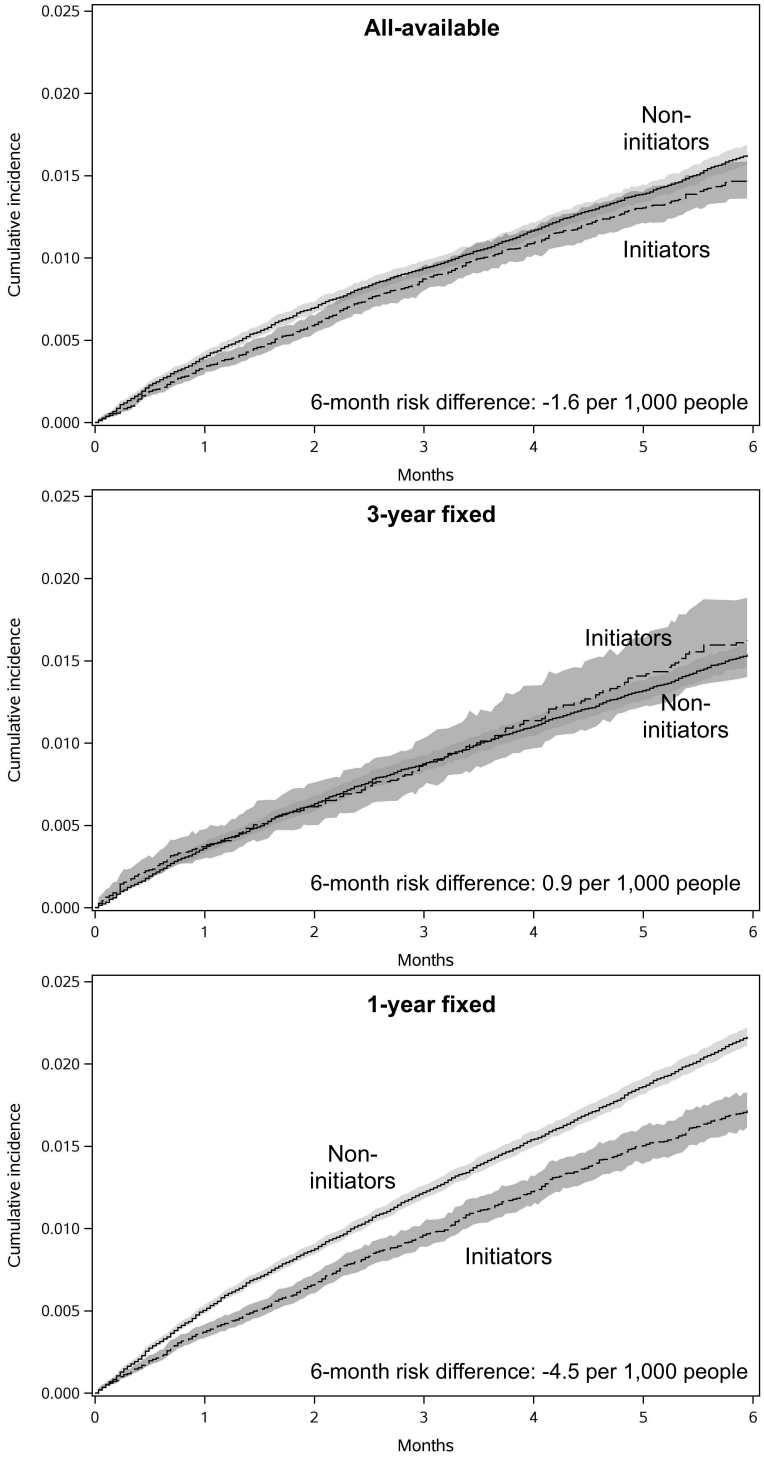


Figure A4.3.7. Standardized mortality ratio weighted cumulative incidence curves and risk difference at six months for incident cancer



APPENDIX 5.1: SUPPLEMENTAL DESCRIPTION OF AIM 2 METHODS

A5.1.1 Origin of parameters (log-RRs) for relationship between MCBS covariates and simulated outcomes

In the interest of producing a realistic simulation, we used an empirical approach to 1) select the MCBS covariates used to simulate exposures and outcomes, and 2) to define the pre-specified β-coefficients (log-RRs) for the relationship between covariates and the outcome. We assessed time to all-cause mortality within one year after the index date (using mortality information in Part A/B Medicare claims data), and censored follow-up if a subject dis-enrolled from Medicare Part A or B. We then fit a multivariate Cox proportional hazards model to estimate hazard ratios (HRs) for the relationships between each MCBS variable and mortality within one year. The covariates which had pronounced relationships (HRs) with the mortality outcome were selected as "true" confounders in the plasmode simulation. Next, we fit a second Cox model containing the narrowed set of MCBS variables. Finally, we set the pre-specified β-coefficients (log-RRs) in the outcome simulation to be approximately equal to the log-HR for each covariate as estimated by this outcome model.

We chose to base parameters on the relationship between covariates and mortality (as opposed to inpatient hospitalizations) because 1) the covariates which predict mortality are more likely to be important across a range of analyses that may seek to apply our findings, 2) mortality is well-classified in the data, and 3) identifying mortality incidence does not require consideration of competing risks.

A5.1.2 Generating pre-specified exposure and outcome prevalence by calibrating intercept terms in simulation equations

In this study, we used simulation to explore the performance of different look-back approaches across a range of scenarios. In order to understand how different simulated features (e.g. a protective vs. harmful exposure effect) impact the performance of these

approaches, we compare scenarios which differ with respect to only that feature. However, varying one parameter in a simulation often has unintended consequences which can obscure findings. In particular, we wanted to ensure that varying different study parameters did not affect the proportion of subjects exposed or the mean outcome rate. To achieve this, we allowed the intercept term in exposure and outcome probability equations to vary between scenarios such that each generated pre-specified prevalences.

A5.1.2.1 Calibration of the intercept term in the exposure probability simulation equation

To calculate the intercept term for each scenario's exposure probability equation, we derived the "calibration equation" below using 1) the exposure probability equation, 2) the pre-specified simulation parameters for each scenario, and 3) the distribution of each covariate in the source cohort. We applied this equation separately among users who truly did have prior statin use and users who were statin naïve.

$$\beta_0 = -\text{Ln}(p_{exp}^{-1} - 1) - (\vec{\beta}_{continuous} * \vec{\mu}_{continuous}) - (\vec{\beta}_{binary} * \vec{p}_{binary})$$

where:

β_0 = calibrated intercept term

p_{exp} = pre-specified prevalence of exposure

$\vec{\beta}_{continuous}$ = a vector of pre-specified log-ORs for the effect of each continuous MCBS confounder on the likelihood of exposure

$\vec{\mu}_{continuous}$ = a vector of the mean value for each continuous MCBS confounder in the source dataset

$\vec{\beta}_{binary}$ = a vector of pre-specified log-ORs for the effect of each continuous MCBS confounder on the likelihood of exposure

$\vec{p}_{binary}$ = a vector of the proportion with each binary MCBS confounder in the source dataset.

A5.1.2.2 Calibration of the intercept term in the outcome rate simulation equation

To calculate the intercept term for each scenario's outcome rate equation, we derived the "calibration equation" below using 1) the outcome ate equation, 2) the pre-specified simulation parameters for each scenario, and 3) the distribution of each covariate in the source cohort. we derived the equation and used it to calculate the intercept term in the exposure probability simulation equation. We applied this equation separately among users who truly did have prior statin use and users who were statin naïve.

$$\beta_0 = -\text{Ln}(\mu_{\text{outcome}}) - \beta_{exp} * p_{exp} - (\vec{\beta}_{continuous} * \vec{\mu}_{continuous}) - (\vec{\beta}_{binary} * \vec{p}_{binary})$$

where:

β_0 = calibrated intercept term

μ_{outcome} = the pre-specified mean outcome rate

β_{exp} = the log-RR (or RD) for the effect of interest

p_{exp} = pre-specified prevalence of exposure

$\vec{\beta}_{continuous}$ = a vector of pre-specified log-RRs for the effect of each continuous MCBS confounder on the likelihood of exposure

$\vec{\mu}_{continuous}$ = a vector of the mean value for each continuous MCBS confounder in the source dataset

$\vec{\beta}_{binary}$ = a vector of pre-specified log-RRs for the effect of each continuous MCBS confounder on the likelihood of exposure

$\vec{p}_{binary}$ = a vector of the proportion with each binary MCBS confounder in the source dataset.

A5.1.3 Additional considerations when interpreting estimates of residual confounding bias due to inadequate ascertainment of covariates

Our approach to isolating the residual confounding bias caused by inadequate ascertainment of covariates is imperfect. We produce this estimate by using the perfect

MCBS covariate to select the cohort, in theory eliminating the component of bias caused by inclusion of prior users. Since the prevalence of some confounding variables (e.g. diabetes) varies between people with and without prior statin use (Table 5.2), excluding prior users may affect the prevalence of confounders. This could increase or decrease the amount of confounding in the estimate. However, back-of-the-envelope calculations using the bias estimates presented in Figure 5.3 and Figure A5.3.1 reveal that summing estimates of the two component biases approximately equals the net bias estimate. Thus, we do not believe this has a meaningful impact on findings.

A5.1.4 Description of sub-analysis using the all-available look-back approach after imposing differential information accuracy by outcome and/or exposure status

In order to explore the performance all-available look-backs in scenarios with highly differential information inaccuracy, we conducted sub-analyses in which we intentionally left-truncated data histories differentially by exposure status, outcome status, and both simultaneously. In the tables below, we describe for each scenario the amount of data history that we assessed within each sub-group.

Differential information accuracy with respect to exposure status

	3yr(E=1)|1yr(E=0): More look-back available among initiators		*3yr(E=0)|1yr(E=1)*: Less look-back available among initiators	
	Outcome > 0	Outcome = 0	Outcome > 0	Outcome = 0
Initiators	3 years	3 years	1 year	1 year
Non-initiators	1 year	1 year	3 years	3 years

Differential information accuracy with respect to outcome status

	3yr(O=1)/1yr(O=0): More look-back available among those with outcomes		*3yr(O=0)|1yr(O=1)*: Less look-back available among those with no outcomes	
	Outcome > 0	Outcome = 0	Outcome > 0	Outcome = 0
Initiators	3 years	1 year	1 year	3 years
Non-initiators	3 years	1 year	1 year	3 years

Differential information accuracy with respect to exposure and outcome status

Scenario 1 & 2: More look-back available among those with outcomes

	Diff info #1: More look-back available among initiators		*Diff info #2:* Less look-back available among initiators	
	Outcome > 0	Outcome = 0	Outcome > 0	Outcome = 0
Initiators	3 years	1.5 years	0.5 years	1 year
Non-initiators	1 year	0.5 years	1.5 years	3 years

Scenario 3 & 4: More look-back available among those with no outcomes

	Diff info #3: More look-back available among initiators		*Diff info #4:* Less look-back available among initiators	
	Outcome > 0	Outcome = 0	Outcome > 0	Outcome = 0
Initiators	1.5 years	3 years	1 year	0.5 years
Non-initiators	0.5 years	1 year	3 years	1.5 years

A5.1.5 Description of additional sub-analyses

We completed a range of sub-analyses to assess the robustness of our findings regarding the relative performance of the look-back approaches. We explored scenarios where exposure reduced the outcome rate and scenarios where it increased outcome rate. In the primary analysis, we allowed the impact of MCBS covariates on simulated exposures and outcomes to vary depending on whether they occurred proximally (within the last year) or distally (before the last year) relative to the index interview. We conducted two sub-analyses in which the effect of these variables on simulated exposures and outcomes was fixed: one in which only proximal covariates had an impact and a second in which all covariates (proximal or distal) had the same impact. In the primary analyses, we did not trim the propensity score distribution or truncate weights. However, we explored results with 2% asymmetric trimming of the propensity score distribution (i.e. restricted the cohort to those with propensity scores above the 2nd percentile among the exposed and below the 98th percentile among the unexposed.

In all scenarios studied, 2% asymmetric trimming had a negligible impact on estimates. However, one exception was that in the sub-analysis with time-fixed confounding

(by variables occurring only the prior year), the look-back approaches produced estimates with identical bias (0.29) after trimming. Further description of characteristics and results from the sub-analyses where we simulated time-fixed confounding are available in Web Table 4.

APPENDIX 5.2: AIM 2 SUPPLEMENTAL TABLES

Table A5.2.1. Coding definitions for MCBS survey variables (true confounders in the simulation)

MCBS survey variable	MCBS code definition (original MCBS dataset variables in **bold**)
Statin use	*In any interview not occurring after the index interview* **FDB_GNN** contains "VASTATIN"
Sex: female	**H_SEX** = 2
Age (years)	Age = **interview_date** – **h_dob**
BMI (kg/m^2)	HEIGHT_m = (**HEIGHTFT** * 0.3048) + (**HEIGHTIN** * 0.0254) WEIGHT_kg = **WEIGHT** * 0.453592 BMI = (WEIGHT_kg / (HEIGHT_m * HEIGHT_m));
Routine cancer screening	Females: (**MAMMOGRM** = 1) or (**PAPSMEAR** = 1) Males: (**DIGTEXAM** = 1) or (**BLOODTST** = 1)
Serious (non-skin) cancer [any]	(**OCCLUNG** = 1) or (**OCCCOLON** = 1) or (**OCCUTER** = 1) or (**OCCOVARY** = 1) or (**OCCSTOM** = 1) or (**OCCCERVX** = 1) or (**OCCKIDNY** = 1) or (**OCCBRAIN** = 1) or (**OCCTHROA** = 1) or (**OCCHEAD** = 1)
Within prior year	In any and (**D_CANCER** = 1)
Before prior year	In any and (**D_CANCER** = 0)
Major cardiovascular event [any]	(**OCMYOCAR** = 1) or (**OCCHD** = 1) or (**OCCFAIL** = 1) or (**OCSTROKE** = 1)
Within prior year	(**D_MYOCAR** = 1) or (**D_CHD** = 1) or (**D_CFAIL** = 1) or (**D_STROKE** = 1)
Before prior year	In any and not in prior year
ACE/ARB use	*In either the index interview or the two prior Summer/Winter interviews* ACE-Is: **FDB_GNN** contains "PRIL" or “PRILO” ARBs: **FDB_GNN** contains "SARTAN"
Diabetes (ever)	(**OCBETES** = 1) or (**OCBETES** = 3) or (**OCBETES** = 4)
Frailty (ADL)	Any subject experiencing difficulty with any activity of daily living (below)
Bathing	(**HPPDBATH** = 1) and ((**DONTBATH** = 1) or (**HELPBATH** = 1))
Dressing	(**HPPDDRES** = 1) and ((**DONTDRES** = 1) or (**HELPDRES** = 1))
Eating	(**HPPDEAT** = 1) and ((**DONTEAT** = 1) or (**HELPEAT** = 1))
Getting out of bed/chair	(**HPPDCHAR** = 1) and ((**DONTCHAR** = 1) or (**HELPCHAR** = 1))
Walking	(**HPPDWALK** = 1) and ((**DONTWALK** = 1) or (**HELPWALK** = 1))
Using the toilet	(**HPPDTOIL** = 1) and ((**DONTTOIL** = 1) or (**HELPTOIL** = 1))

Table A5.2.2. Simulated relationships between covariates assessed in the MCBS survey data and exposures/outcomes

	Time-stratified confounding		Time-fixed confounding: proximal		Time-fixed confounding: any	
MCBS survey variables	Exposure OR	Outcome RR	Exposure OR	Outcome RR	Exposure OR	Outcome RR
Sex: female	0.50	0.56	0.50	0.54	0.50	0.56
Age (years)	0.996	1.06	0.996	1.06	0.996	1.06
Age (squared term)	1.001	-	1.001	-	1.001	-
BMI (kg/m^2)	1.015	0.72	1.015	0.72	1.015	0.71
BMI (squared term)	1.002	1.005	1.002	1.005	1.002	1.005
Routine cancer screen	0.50	0.61	0.50	0.61	0.50	0.62
ACE inhibitor or ARB	0.50	0.50	0.50	0.50	0.50	0.50
Diabetes (ever)	2.00	2.00	2.00	2.00	2.00	2.00
Frailty	2.00	3.02	2.00	3.12	2.00	3.11
Time-varying confounders						
Serious (non-skin) cancer						
≤ 1 year	4.00	3.75	4.00	3.40	-	-
> 1 year	2.00	1.96	-	-	-	-
ever	-	-	-	-	3.00	2.26
Major CV event						
≤ 1 year	4.00	2.22	4.00	1.80	-	-
> 1 year	2.00	1.77	-	-	-	-
ever	-	-	-	-	3.00	1.89

Table A5.2.3. Frequencies and proportions of study population with exposure and outcome frequencies for each look-back approach, after excluding those with prior statin use (median results among 1,500 iterations)

	Time-stratified confounding		Time-fixed confounding: proximal		Time-fixed confounding: any	
	Total N	Outcomes %[a]	Total N	Outcomes %[a]	Total N	Outcomes %[a]
Truth						
Initiator	1,467.0	37.8%	1,457.0	36.0%	1,475.0	38.5%
Non-initiator	2,688.0	46.7%	2,697.0	47.7%	2,678.0	46.3%
Total	4,155.0	43.6%	4,154.0	43.6%	4,153.0	43.6%
All-available						
Initiator	1,538.0	38.9%	1,538.0	37.5%	1,539.0	39.2%
Non-initiator	2,587.0	46.2%	2,586.0	47.3%	2,586.0	45.8%
Total	4,125.0	43.5%	4,124.0	43.6%	4,125.0	43.4%
1-year						
Initiator	1,870.0	40.6%	1,862.0	39.0%	1,872.0	41.1%
Non-initiator	2,791.0	46.2%	2,688.0	49.3%	2,787.0	45.7%
Total	4,661.0	44.0%	4,550.0	45.1%	4,659.0	43.9%
3-year						
Initiator	1,251.0	40.3%	1,246.5	38.8%	1,255.0	40.8%
Non-initiator	2,063.0	48.9%	2,071.0	50.0%	2,060.0	48.5%
Total	3,314.0	45.6%	3,317.5	45.8%	3,315.0	45.6%

[a] This % shows the proportion of subjects who had at least one occurrence of the outcome (O>0).

Table A5.2.4. Simulation characteristics and overview of results for each look-back approach, for time-stratified and time-fixed confounding simulations (median results among 1,500 iterations)

	% of cohort in equipoise[a]		Exclusions (%)[a]		Prior statin use (% of final cohort)[a]		Rate ratio (true effect = 0.40)			Rate difference (true effect = -0.08)		
	Initiator	Non-initiator	Cont. enroll.	Prior statin use	Initiator	Non-initiator	Crude	IPTW	C.I. width	Crude	IPTW	C.I. width
Time-stratified confounding (+/- 1-yr)												
MCBS	58%	61%	0%	58%	0%	0%	0.70	0.40 (0.36, 0.45)	0.24	0.54	-0.08 (-0.21, 0.03)	0.24
Time-stratified adjustment (+/- 6-mo)												
All avail.	76%	80%	1%	58%	14%	5%	0.74	0.59 (0.53, 0.67)	0.24	0.50	0.26 (0.14, 0.39)	0.24
1-yr	77%	81%	7%	50%	26%	10%	0.80	0.65 (0.58, 0.72)	0.22	0.53	0.30 (0.19, 0.42)	0.23
3-yr	78%	82%	22%	58%	13%	4%	0.73	0.60 (0.53, 0.68)	0.26	0.55	0.32 (0.17, 0.47)	0.30
Time-fixed adjustment												
All avail.	80%	84%	1%	58%	14%	5%	0.74	0.61 (0.54, 0.69)	0.24	0.50	0.30 (0.18, 0.43)	0.25
1-yr	79%	83%	7%	50%	26%	10%	0.80	0.66 (0.59, 0.73)	0.21	0.53	0.32 (0.20, 0.43)	0.23
3-yr	81%	85%	22%	58%	13%	4%	0.73	0.63 (0.55, 0.71)	0.25	0.55	0.37 (0.22, 0.53)	0.31
Sub-analyses[b]												
3yr(E=0)|1yr(E=1)	60%	61%	1%	53%	27%	5%	0.78	0.65 (0.57, 0.73)	0.25	0.49	0.29 (0.16, 0.44)	0.27
3yr(E=1)|1yr(E=0)	58%	65%	1%	54%	15%	11%	0.76	0.61 (0.54, 0.69)	0.24	0.52	0.29 (0.16, 0.42)	0.25
3yr(O=0)|1yr(O>0)	78%	82%	1%	54%	21%	7%	0.84	0.70 (0.63, 0.77)	0.20	0.59	0.35 (0.23, 0.46)	0.23
3yr(O>0)/1yr(O=0)	81%	84%	1%	53%	22%	9%	0.70	0.61 (0.54, 0.68)	0.24	0.42	0.25 (0.14, 0.37)	0.23
Differential info #1[c]	54%	55%	1%	53%	16%	11%	0.73	0.56 (0.49, 0.63)	0.25	0.47	0.20 (0.06, 0.34)	0.27
Differential info #2[d]	47%	46%	1%	52%	27%	6%	0.81	0.73 (0.64, 0.84)	0.28	0.52	0.41 (0.26, 0.58)	0.32
Differential info #3[e]	52%	57%	1%	53%	16%	11%	0.80	0.71 (0.62, 0.80)	0.25	0.56	0.40 (0.28, 0.54)	0.26
Differential info #4[f]	55%	58%	1%	52%	28%	6%	0.85	0.65 (0.56, 0.74)	0.28	0.57	0.27 (0.13, 0.43)	0.30
Time-fixed confounding (<1-yr only)												
MCBS	62%	66%	0%	58%	0%	0%	0.63	0.40 (0.35, 0.45)	0.23	0.40	-0.07 (-0.20, 0.03)	0.23
Time-stratified adjustment (+/- 6-mo)												
All avail.	79%	82%	1%	58%	14%	5%	0.69	0.58 (0.51, 0.65)	0.24	0.40	0.22 (0.11, 0.35)	0.25
1-yr	80%	83%	7%	50%	27%	10%	0.73	0.62 (0.56, 0.69)	0.21	0.39	0.24 (0.13, 0.35)	0.23
2-yr	79%	82%	13%	55%	16%	6%	0.69	0.59 (0.52, 0.67)	0.24	0.41	0.25 (0.12, 0.38)	0.26
3-yr	79%	83%	22%	58%	13%	4%	0.67	0.58 (0.51, 0.66)	0.26	0.41	0.25 (0.11, 0.41)	0.29
Time-fixed adjustment												
All avail.	82%	85%	1%	58%	14%	5%	0.69	0.59 (0.53, 0.67)	0.24	0.40	0.25 (0.12, 0.38)	0.26
1-yr	82%	85%	7%	50%	27%	10%	0.73	0.62 (0.56, 0.70)	0.22	0.39	0.24 (0.13, 0.35)	0.23
2-yr	81%	85%	13%	55%	16%	6%	0.69	0.61 (0.54, 0.69)	0.26	0.41	0.28 (0.15, 0.43)	0.29
3-yr	83%	86%	22%	58%	13%	4%	0.67	0.59 (0.52, 0.68)	0.26	0.41	0.28 (0.14, 0.45)	0.31
Sub-analyses[b]												

	% of cohort in equipoise[a]		Exclusions (%)[a]		Prior statin use (% of final cohort)[a]		Rate ratio (true effect = 0.40)			Rate difference (true effect = -0.08)		
	Initiator	Non-initiator	Cont. enroll.	Prior statin use	Initiator	Non-initiator	Crude	IPTW	C.I. width	Crude	IPTW	C.I. width
3yr(E=0)|1yr(E=1)	61%	62%	1%	53%	28%	5%	0.71	0.61 (0.54, 0.69)	0.26	0.36	0.21 (0.07, 0.36)	0.29
3yr(E=1)|1yr(E=0)	61%	67%	1%	54%	15%	11%	0.70	0.58 (0.52, 0.66)	0.24	0.40	0.22 (0.11, 0.33)	0.22
3yr(O=0)|1yr(O>0)	80%	84%	1%	54%	21%	7%	0.77	0.66 (0.60, 0.73)	0.20	0.45	0.26 (0.16, 0.38)	0.22
3yr(O>0)/1yr(O=0)	82%	85%	1%	53%	23%	9%	0.64	0.59 (0.52, 0.67)	0.25	0.30	0.20 (0.09, 0.32)	0.23
Differential info #1[c]	56%	56%	1%	53%	17%	10%	0.67	0.53 (0.47, 0.60)	0.25	0.36	0.14 (0.02, 0.26)	0.24
Differential info #2[d]	46%	46%	1%	52%	27%	6%	0.73	0.70 (0.60, 0.81)	0.30	0.38	0.34 (0.18, 0.52)	0.34
Differential info #3[e]	54%	59%	1%	53%	16%	11%	0.73	0.68 (0.60, 0.77)	0.25	0.43	0.33 (0.22, 0.46)	0.24
Differential info #4[f]	56%	59%	1%	52%	28%	6%	0.75	0.59 (0.51, 0.68)	0.29	0.41	0.16 (0.03, 0.32)	0.29
Time-fixed confounding (any)												
MCBS	53%	56%	0%	58%	0%	0%	0.72	0.40 (0.35, 0.45)	0.25	0.57	-0.07 (-0.21, 0.04)	0.25
Time-stratified adjustment (+/- 6-mo)												
All avail.	75%	79%	1%	58%	14%	5%	0.74	0.59 (0.53, 0.67)	0.23	0.50	0.27 (0.15, 0.39)	0.24
1-yr	76%	80%	7%	50%	26%	11%	0.82	0.66 (0.59, 0.73)	0.22	0.55	0.32 (0.21, 0.43)	0.22
2-yr	76%	80%	13%	55%	16%	6%	0.76	0.62 (0.55, 0.70)	0.24	0.56	0.33 (0.20, 0.46)	0.25
3-yr	77%	81%	22%	58%	12%	5%	0.74	0.61 (0.53, 0.68)	0.25	0.56	0.33 (0.19, 0.47)	0.28
Time-fixed adjustment												
All avail.	79%	84%	1%	58%	14%	5%	0.74	0.62 (0.55, 0.69)	0.24	0.50	0.31 (0.19, 0.44)	0.25
1-yr	79%	82%	7%	50%	26%	11%	0.82	0.67 (0.60, 0.74)	0.21	0.55	0.34 (0.23, 0.45)	0.23
2-yr	80%	84%	13%	55%	16%	6%	0.76	0.65 (0.57, 0.73)	0.24	0.56	0.38 (0.24, 0.53)	0.29
3-yr	81%	85%	22%	58%	12%	5%	0.74	0.64 (0.56, 0.72)	0.25	0.56	0.39 (0.25, 0.54)	0.30
Sub-analyses[b]												
3yr(E=0)|1yr(E=1)	60%	60%	1%	53%	27%	5%	0.80	0.65 (0.58, 0.72)	0.23	0.51	0.29 (0.16, 0.41)	0.25
3yr(E=1)|1yr(E=0)	58%	64%	1%	54%	14%	11%	0.77	0.64 (0.56, 0.72)	0.26	0.54	0.34 (0.20, 0.47)	0.27
3yr(O=0)|1yr(O>0)	77%	81%	1%	54%	21%	7%	0.86	0.71 (0.64, 0.78)	0.20	0.62	0.37 (0.26, 0.48)	0.22
3yr(O>0)/1yr(O=0)	81%	83%	1%	53%	22%	9%	0.71	0.61 (0.55, 0.69)	0.23	0.43	0.27 (0.16, 0.39)	0.23
Differential info #1[c]	53%	55%	1%	53%	16%	11%	0.74	0.58 (0.51, 0.66)	0.27	0.49	0.25 (0.10, 0.39)	0.29
Differential info #2[d]	47%	46%	1%	52%	26%	6%	0.82	0.73 (0.64, 0.83)	0.26	0.54	0.40 (0.26, 0.54)	0.27
Differential info #3[e]	51%	57%	1%	53%	16%	11%	0.81	0.73 (0.64, 0.83)	0.26	0.58	0.44 (0.31, 0.58)	0.27
Differential info #4[f]	54%	57%	1%	52%	28%	6%	0.85	0.63 (0.56, 0.71)	0.24	0.58	0.26 (0.13, 0.38)	0.25

[a] These results reflect the results from the simulation of a homogenous rate ratio. However, the values of these statistics were nearly identical in simulations of a homogenous rate difference.

[b] In the sub-analyses listed below, we restricted the cohort to respondents with at least 3-years continuous enrollment then intentionally censored the look-back selectively by exposure and outcome status.

[c] Differential Info #1: 3-yr(E=1|O>0), 2-yr(E=1|O=0), 1-yr(E=0|O>0), 6-mo(E=0|O=0)

[d] Differential Info #2: 3-yr(E=1|O=0), 2-yr(E=0|O=0), 1-yr(E=1|O>0), 6-mo(E=1|O=0)

[e] Differential Info #3: 3-yr(E=0|O>0), 2-yr(E=1|O>0), 1-yr(E=0|O=0), 6-mo(E=0|O>0)
[f] Differential Info #4: 3-yr(E=0|O=0), 2-yr(E=0|O>0), 1-yr(E=1|O=0), 6-mo(E=1|O>0)

APPENDIX 5.3: AIM 2 SUPPLEMENTAL FIGURES

Figure A5.3.1. Forest plot displaying performance of various look-back approaches when estimating IPTW-adjusted estimates of the rate difference in simulations of an absolute exposure effect.
Bias estimates are accompanied by 95% confidence intervals (CIs) and corresponding estimates of the root-mean-squared-error (rMSE). For each look-back approach, the plot includes estimates of the net (total) bias, residual confounding bias due to misclassification of covariates, and bias due to inclusion of prior statin users. (*Note: the corresponding figure for the rate ratio analysis is presented in Figure 5.3.)*

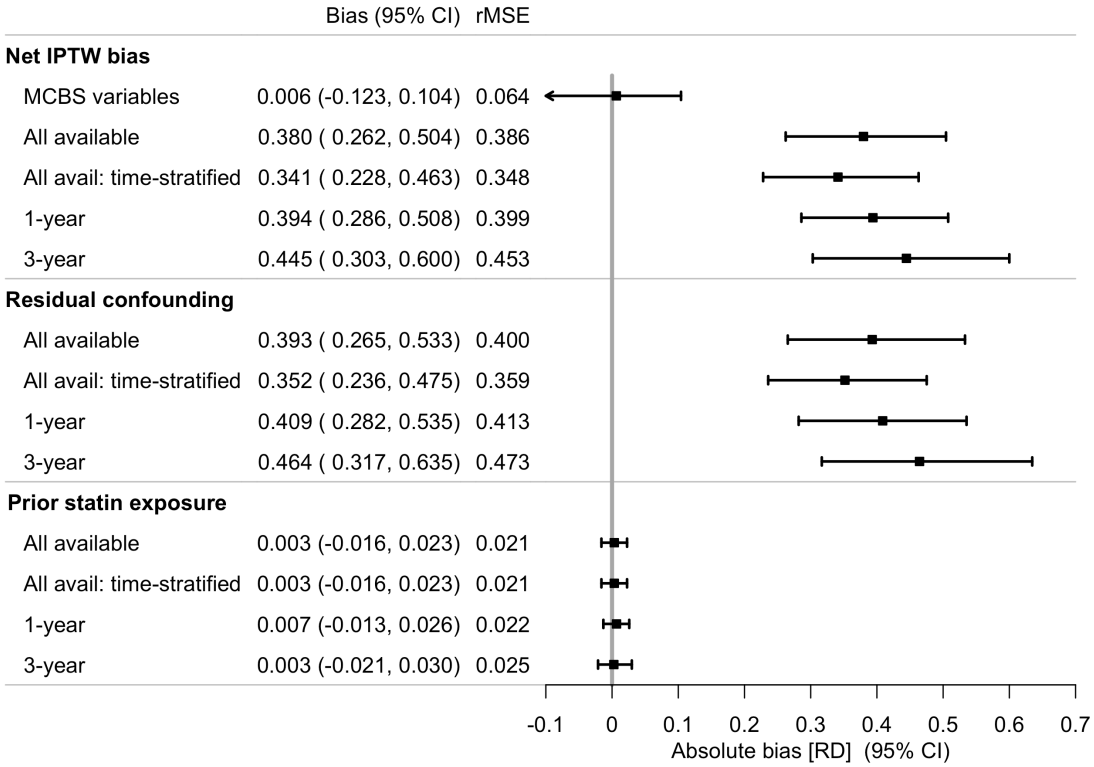
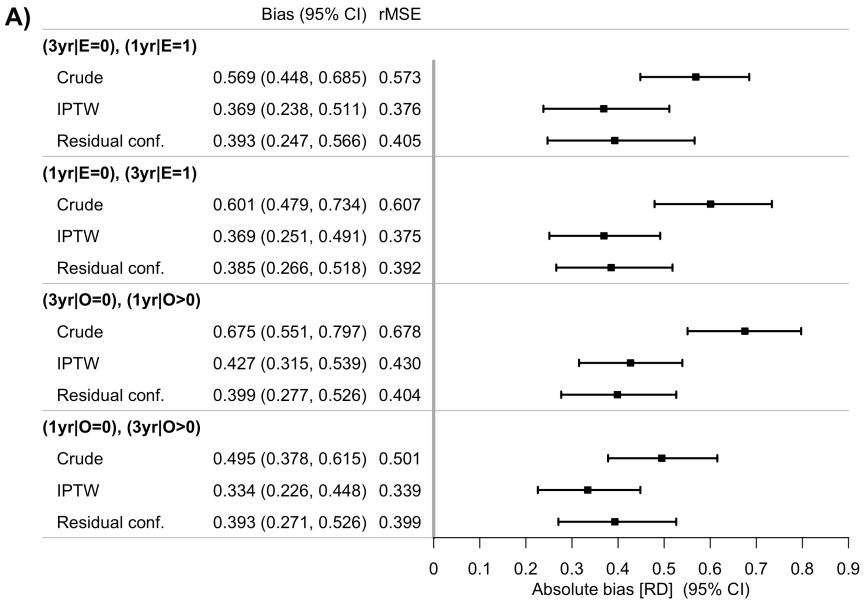


Figure A5.3.2. Forest plot displaying performance of the all-available look-back approach in sub-analyses with highly differential information accuracy A) by exposure or outcome status, and B) by exposure and outcome status (simultaneously) when estimating IPTW-adjusted estimates of the rate difference in simulations of a multiplicative exposure effect.
E=exposure, O=outcome. Bias estimates are accompanied by 95% confidence intervals (CIs) and corresponding estimates of the root-mean-squared-error (rMSE). For each sub-analysis, the plot includes estimates of the net (total) bias and residual confounding bias due to misclassification of covariates. Partial control for confounding holds in all analyses since the residual confounding bias due to misclassification of covariates is closer to zero than the crude (unadjusted) bias. *(Note: the corresponding figure for the rate ratio analysis is presented in Figure 5.4.)*



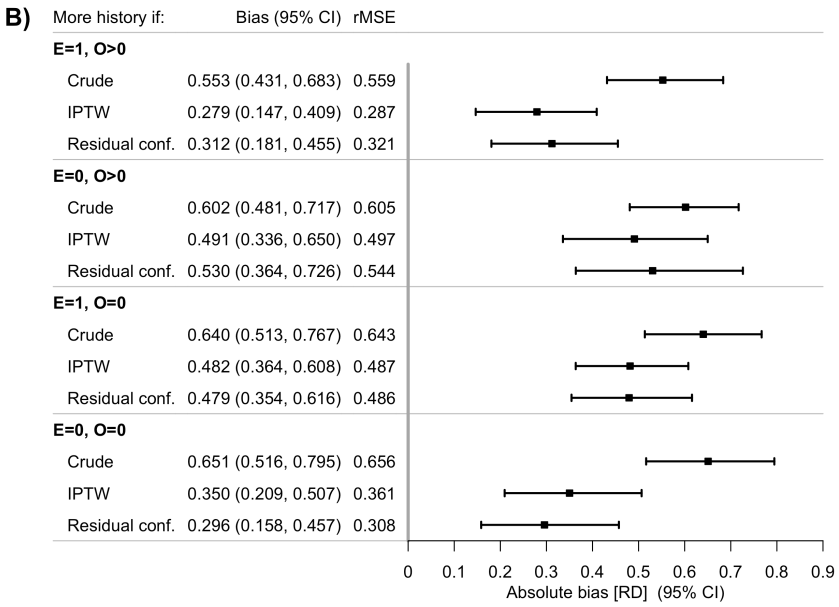
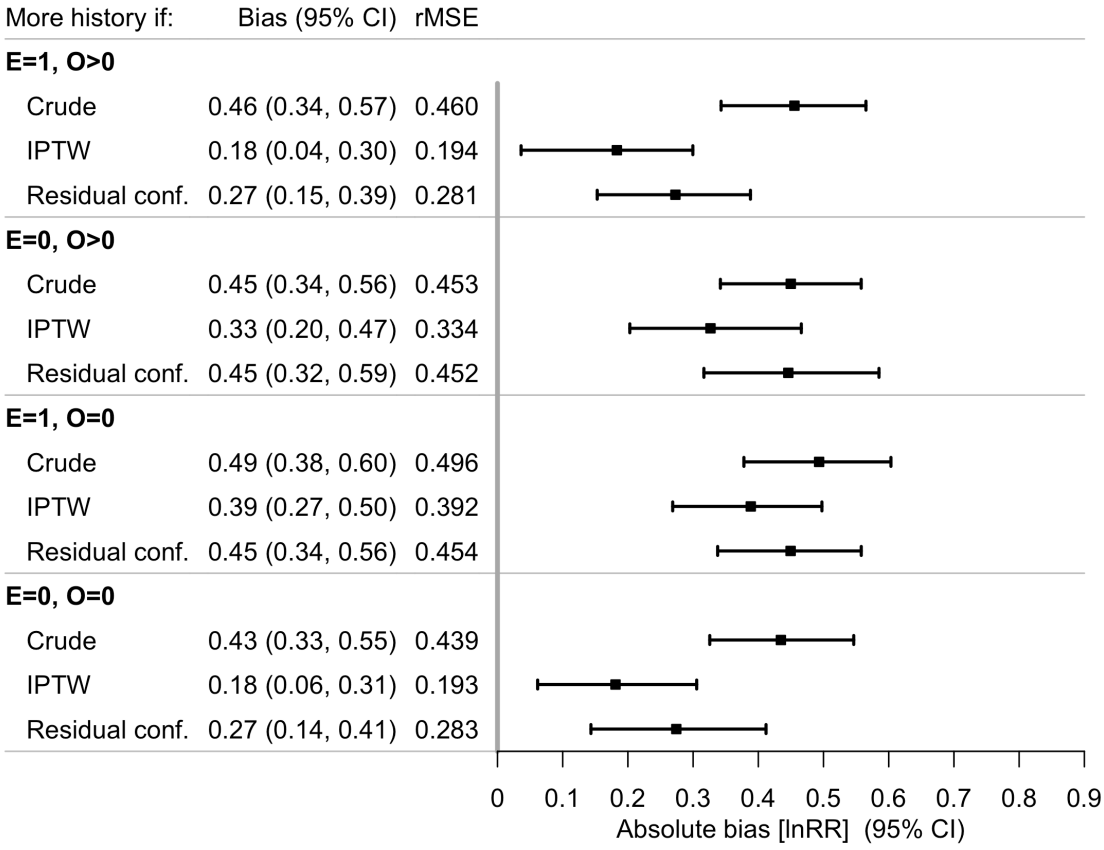


Figure A5.3.3. Forest plot displaying performance of the all-available look-back approach in four sub-analyses with highly differential information accuracy by both exposure and outcome (simultaneously) when estimating IPTW-adjusted estimates of the rate ratio in simulations of a harmful multiplicative exposure effect.
E=exposure, O=outcome. Bias estimates are accompanied by 95% confidence intervals (CIs) and corresponding estimates of the root-mean-squared-error (rMSE). For each sub-analysis, the plot includes estimates of the net (total) bias and residual confounding bias due to misclassification of covariates. In the scenario where more database history was available among non-initiators (E=0) and people with outcomes (O>0), adjusting for confounding had no effect on the estimate. *(Note: the corresponding figure simulating a protective effect of interest analysis is presented in Figure 5.4b.)*



REFERENCES

1. Safran C, Bloomrosen M, Hammond WE, et al. Toward a national framework for the secondary use of health data: an American Medical Informatics Association White Paper. *J Am Med Inform Assn.* 2007;14(1):1-9.

2. Velentgas P, Dreyer NA, Nourjah P, Smith SR, Torchia MM, eds. Developing a protocol for observational comparative effectiveness research: a user's guide. AHRQ Publication No. 12(13)-EHC099. Rockville, MD: Agency for Healthcare Research and Quality; January 2013. https://effectivehealthcare.ahrq.gov/topics/observational-cer-protocol/research/. Accessed September 26, 2017.

3. Fisher ES, Whaley FS, Krushat WM, et al. The accuracy of Medicare's hospital claims data: progress has been made, but problems remain. *Am J Public Health.* 1992;82(2):243-248.

4. Brunelli SM, Gagne JJ, Huybrechts KF, et al. Estimation using all available covariate information versus a fixed look-back window for dichotomous covariates. *Pharmacoepidemiol Drug Saf.* 2013;22(5):542-550.

5. Glynn RJ, Monane M, Gurwitz JH, Choodnovskiy I, Avorn J. Agreement between drug treatment data and a discharge diagnosis of diabetes mellitus in the elderly. *Am J Epidemiol.* 1999;149(6):541-549.

6. Iezzoni LI, Foley SM, Daley J, Hughes J, Fisher ES, Heeren T. Comorbidities, complications, and coding bias. Does the number of diagnosis codes matter in predicting in-hospital mortality? *JAMA.* 1992;267(16):2197-2203.

7. Rothman KJ, Greenland S, Lash TL. Case-Control Studies. In: Seigafuse S, Bierig L, eds. *Modern Epidemiology.* 3rd ed. Philadelphia (PA): Lippincott Williams & Wilkins; 2008:121-127.

8. Greenland S. The effect of misclassification in the presence of covariates. *Am J Epidemiol.* 1980;112(4):564-569.

9. Walker AM, Lanes SF. Misclassification of covariates. *Stat Med.* 1991;10(8):1181-1196.

10. Brenner H. Bias due to non-differential misclassification of polytomous confounders. *J Clin Epidemiol.* 1993;46(1):57-63.

11. Schneeweiss S, Avorn J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. *J Clin Epidemiol.* 2005;58(4):323-337.

12. Wacholder S. When measurement errors correlate with truth: surprising effects of nondifferential misclassification. *Epidemiology (Cambridge, Mass.).* 1995;6(2):157-161.

13. Wacholder S, McLaughlin JK, Silverman DT, Mandel JS. Selection of controls in case-control studies. I. Principles. *Am J Epidemiol.* 1992;135(9):1019-1028.

14. Nakasian SS, Rassen JA, Franklin JM. Effects of expanding the look-back period to all available data in the assessment of covariates. *Pharmacoepidemiol Drug Saf.* 2017;26(8):890-899.

15. Afilalo J, Duque G, Steele R, Jukema JW, de Craen AJ, Eisenberg MJ. Statins for secondary prevention in elderly patients: a hierarchical bayesian meta-analysis. *J Am Coll Cardiol.* 2008;51(1):37-45.

16. Roberts CG, Guallar E, Rodriguez A. Efficacy and safety of statin monotherapy in older adults: a meta-analysis. *J Gerontol A-Biol.* 2007;62(8):879-887.

17. Franklin JM, Schneeweiss S, Polinski JM, Rassen JA. Plasmode simulation for the evaluation of pharmacoepidemiologic methods in complex healthcare databases. *Comput Stat Data An.* 2014;72:219-226.

18. Gadbury GL, Xiang Q, Yang L, Barnes S, Page GP, Allison DB. Evaluating statistical methods using plasmode data sets in the age of massive public databases: an illustration using false discovery rates. *PLoS genetics.* 2008;4(6):e1000098.

19. Brookhart MA, Wyss R, Layton JB, Stürmer T. Propensity score methods for confounding control in nonexperimental research. *Circ Cardiovasc Qual Outcomes.* 2013;6(5):604-611.

20. Marshall JR, Hastrup JL. Mismeasurement and the resonance of strong confounders: uncorrelated errors. *Am J Epidemiol.* 1996;143(10):1069-1078.

21. Strom BL. Overview of Automated Databases in Pharmacoepidemiology. In: Strom BL, Kimmel ES, eds. *Textbook of Pharmacoepidemiology.* 5th ed. West Sussex, England: John Wiley & Sons, Ltd; 2006:167-171.

22. Rodriguez EM, Staffa JA, Graham DJ. The role of databases in drug postmarketing surveillance. *Pharmacoepidemiol Drug Saf.* 2001;10(5):407-410.

23. Arana A, Rivero E, Egberts TCG. What do we show and who does so? An analysis of the abstracts presented at the 19th ICPE. *Pharmacoepidemiol Drug Saf.* 2004;13(suppl 1):S330-S331.

24. Ray WA. Improving automated database studies. *Epidemiology (Cambridge, Mass.).* 2011;22(3):302-304.

25. Madigan D, Ryan PB, Schuemie M, et al. Evaluating the impact of database heterogeneity on observational study results. *Am J Epidemiol.* 2013;178(4):645-651.

26. Glynn RJ, Knight EL, Levin R, Avorn J. Paradoxical relations of drug treatment with mortality in older persons. *Epidemiology (Cambridge, Mass.).* 2001;12(6):682-689.

27. Rothman KJ, Greenland S, Lash TL. Introduction to Regression Modeling. In: Seigafuse S, Bierig L, eds. *Modern Epidemiology.* 3rd ed. Philadelphia (PA): Lippincott Williams & Wilkins; 2008:381-455.

28. Ray WA. Evaluating medication effects outside of clinical trials: new-user designs. *Am J Epidemiol.* 2003;158(9):915-920.

29. Schisterman EF, Cole SR, Platt RW. Overadjustment bias and unnecessary adjustment in epidemiologic studies. *Epidemiology (Cambridge, Mass.).* 2009;20(4):488-495.

30. Csizmadi I, Collet JP. Bias and Confounding in Pharmacoepidemiology. In: Strom BL, Kimmel ES, eds. *Textbook of Pharmacoepidemiology.* 5th ed. West Sussex, England: John Wiley & Sons, Ltd; 2006:261-275.

31. Hallas J, Gaist D, Bjerrum L. The waiting time distribution as a graphical approach to epidemiologic measures of drug utilization. *Epidemiology (Cambridge, Mass.).* 1997;8(6):666-670.

32. Hallas J, Stovring H. Templates for analysis of individual-level prescription data. *Basic Clin Pharmacol Toxicol.* 2006;98(3):260-265.

33. Riis AH, Johansen MB, Jacobsen JB, Brookhart MA, Sturmer T, Stovring H. Short look-back periods in pharmacoepidemiologic studies of new users of antibiotics and asthma medications introduce severe misclassification. *Pharmacoepidemiol Drug Saf.* 2015.

34. Gardarsdottir H, Heerdink ER, Egberts AC. Potential bias in pharmacoepidemiological studies due to the length of the drug free period: a study on antidepressant drug use in adults in the Netherlands. *Pharmacoepidemiol Drug Saf.* 2006;15(5):338-343.

35. Gardarsdottir H, Souverein PC, Egberts TC, Heerdink ER. Construction of drug treatment episodes from drug-dispensing histories is influenced by the gap length. *J Clin Epidemiol.* 2010;63(4):422-427.

36. Stovring H. Selection bias due to immigration in pharmacoepidemiologic studies. *Pharmacoepidemiol Drug Saf.* 2007;16(6):681-686.

37. Lipworth L, Friis S, Blot WJ, et al. A population-based cohort study of mortality among users of ibuprofen in Denmark. *Am J Ther.* 2004;11(3):156-163.

38. Vestergaard P, Rejnmark L, Mosekilde L. Fracture risk associated with use of antiepileptic drugs. *Epilepsia.* 2004;45(11):1330-1337.

39. Crane J, Pearce N, Flatt A, et al. Prescribed fenoterol and death from asthma in New Zealand, 1981-83: case-control study. *Lancet.* 1989;1(8644):917-922.

40. Buist AS, Burney PGJ, Feinstein AR, et al. Fenoterol and fatal asthma. *Lancet.* 1989;1(8646):1071-1071.

41. Elwood JM, Skegg DCG. Review of studies relating prescribed fenoterol to deaths from asthma in New Zealand. *Wellington: Department of Health.* 1989.

42. Morrison AS, MacMahon B, Suginura T. Control of cigarette smoking in evaluating the association of coffee drinking and bladder cancer. *Coffee and Health. New York: Cold Spring Harbor Laboratory.* 1984:127-134.

43. Cederlof R, Doll R, Fowler B, Friberg L, Nelson N, Vouk V. Air pollution and cancer: risk assessment methodology and epidemiological evidence. *Environ Health Perspect.* 1978;22:1-12.

44. Simon D, Yen S, Cole P. Coffee drinking and cancer of the lower urinary tract. *J Natl Cancer Inst.* 1975;54(3):587-591.

45. Hartge P, Hoover R, West DW, Lyon JL. Coffee drinking and risk of bladder cancer. *J Natl Cancer Inst.* 1983;70(6):1021-1026.

46. Marrett LD, Walter SD, Meigs JW. Coffee drinking and bladder cancer in Connecticut. *Am J Epidemiol.* 1983;117(2):113-127.

47. Cox B, Elwood JM. The effect on the stratum-specific odds ratios of nondifferential misclassification of a dichotomous covariate. *Am J Epidemiol.* 1991;133(2):202-207.

48. Feinstein AR, Sosin DM, Wells CK. The Will Rogers phenomenon. Stage migration and new diagnostic techniques as a source of misleading statistics for survival in cancer. *New Engl J Med.* 1985;312(25):1604-1608.

49. Savitz DA, Baron AE. Estimating and correcting for confounder misclassification. *Am J Epidemiol.* 1989;129(5):1062-1071.

50. Kaufman JS, Cooper RS, McGee DL. Socioeconomic status and health in blacks and whites: the problem of residual confounding and the resiliency of race. *Epidemiology (Cambridge, Mass.).* 1997;8(6):621-628.

51. Kristensen P. Bias from nondifferential but dependent misclassification of exposure and outcome. *Epidemiology (Cambridge, Mass.).* 1992;3(3):210-215.

52. Rothman KJ, Greenland S, Lash TL. Validity in Epidemiologic Studies. In: Seigafuse S, Bierig L, eds. *Modern Epidemiology.* 3rd ed. Philadelphia (PA): Lippincott Williams & Wilkins; 2008:128-147.

53. Hernan MA, Cole SR. Invited Commentary: Causal diagrams and measurement bias. *Am J Epidemiol.* 2009;170(8):959-962; discussion 963-954.

54. Greenland S, Pearl J. Adjustments and their consequences—collapsibility analysis using graphical models. *Int Stat Rev.* 2011;79(3):401-426.

55. Pearl J. On measurement bias in causal inference. In: Grunwald P, Spirtes P, eds. *Proceedings of the 26th Conference on Uncertainty and Artificial Intelligence.* Corvallis, WA: AUAI Press; 2012:425-432.

56. Kuroki M, Pearl J. Measurement bias and effect restoration in causal inference. *Biometrika.* 2014;101(2):423-437.

57. Pearl J. Causality and Structural Models in Social Science and Economics. In: *Causality.* 2nd ed. New York, NY: Cambridge University Press; 2009:133-172.

58. Kupper LL. Effects of the use of unreliable surrogate variables on the validity of epidemiologic research studies. *Am J Epidemiol.* 1984;120(4):643-648.

59. Fewell Z, Davey Smith G, Sterne JA. The impact of residual and unmeasured confounding in epidemiologic studies: a simulation study. *Am J Epidemiol.* 2007;166(6):646-655.

60. Funk MJ, Landi SN. Misclassification in Administrative Claims Data: Quantifying the Impact on Treatment Effect Estimates. *Curr Epidemiol Rep.* 2014;1(4):175-185.

61. Alderman BW, Baron AE, Savitz DA. Cautions in the use of antecedents as surrogates for confounders. *Am J Epidemiol.* 1993;137(11):1259-1272.

62. Tzonou A, Kaldor J, Smith PG, Day NE, Trichopoulos D. Misclassification in case-control studies with two dichotomous risk factors. *Rev Epidemiol Sante Publique.* 1986;34(1):10-17.

63. Fung KY, Howe GR. Methodological issues in case-control studies. III--The effect of joint misclassification of risk factors and confounding factors upon estimation and power. *Int J Epidemiol.* 1984;13(3):366-370.

64. Kaldor J, Clayton D. Latent class analysis in chronic disease epidemiology. *Stat Med.* 1985;4(3):327-335.

65. Espeland MA, Hui SL. A general approach to analyzing epidemiologic data that contain misclassification errors. *Biometrics.* 1987;43(4):1001-1012.

66. Sturmer T, Schneeweiss S, Avorn J, Glynn RJ. Adjusting effect estimates for unmeasured confounding with validation data using propensity score calibration. *American Journal of Epidemiology.* 2005;162(3):279-289.

67. Greenland S, Robins JM. Confounding and misclassification. *Am J Epidemiol.* 1985;122(3):495-506.

68. Ogburn EL, VanderWeele TJ. On the nondifferential misclassification of a binary confounder. *Epidemiology (Cambridge, Mass.).* 2012;23(3):433-439.

69. Ahlbom A, Steineck G. Aspects of misclassification of confounding factors. *Am J Ind Med.* 1992;21(1):107-112.

70. Flegal KM, Keyl PM, Nieto FJ. Differential misclassification arising from nondifferential errors in exposure measurement. *Am J Epidemiol.* 1991;134(10):1233-1244.

71. Brenner H, Blettner M. Misclassification bias arising from random error in exposure measurement: implications for dual measurement strategies. *Am J Epidemiol.* 1993;138(6):453-461.

72. Gustafson P, Le Nhu D. Comparing the effects of continuous and discrete covariate mismeasurement, with emphasis on the dichotomization of mismeasured predictors. *Biometrics.* 2002;58(4):878-887.

73. Schneeweiss S, Glynn RJ, Tsai EH, Avorn J, Solomon DH. Adjusting for unmeasured confounders in pharmacoepidemiologic claims data using external information: the example of COX2 inhibitors and myocardial infarction. *Epidemiology (Cambridge, Mass.).* 2005;16(1):17-24.

74. Lunt M, Glynn RJ, Rothman KJ, Avorn J, Sturmer T. Propensity score calibration in the absence of surrogacy. *Am J Epidemiol.* 2012;175(12):1294-1302.

75. Schneeweiss S, Rassen JA, Glynn RJ, Avorn J, Mogun H, Brookhart MA. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology (Cambridge, Mass.).* 2009;20(4):512-522.

76. Marshall J. Invited commentary: Fewell and colleagues--fuel for debate. *Am J Epidemiol.* 2007;166(6):656-658; discussion 659-661.

77. Fewell Z, Smith GD, Sterne JAC. Fewell et al. Respond to “Fuel for Debate”. *Am J Epidemiol.* 2007;166(6):659-661.

78. Armstrong BG, Whittemore AS, Howe GR. Analysis of case-control data with covariate measurement error: application to diet and colon cancer. *Stat Med.* 1989;8(9):1151-1163; discussion 1165-1156.

79. Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Stürmer T. Variable selection for propensity score models. *Am J Epidemiol.* 2006;163(12):1149-1156.

80. Sturmer T, Wyss R, Glynn RJ, Brookhart MA. Propensity scores for confounder adjustment when assessing the effects of medical interventions using nonexperimental study designs. *J Intern Med.* 2014;275(6):570-580.

81. Hutchison GB, Shapiro S. Lead time gained by diagnostic screening for breast cancer. *J Natl Cancer Inst.* 1968;41(3):665-681.

82. Feinleib M, Zelen M. Some pitfalls in the evaluation of screening programs. *Arch Environ Health.* 1969;19(3):412-415.

83. Prentice RL. Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika.* 1982;69(2):331-342.

84. Pepe MS, Self SG, Prentice RL. Further results on covariate measurement errors in cohort studies with time to response data. *Stat Med.* 1989;8(9):1167-1178; discussion 1179.

85. Gorfine M, Hsu L, Prentice RL. Nonparametric correction for covariate measurement error in a stratified Cox model. *Biostatistics (Oxford, England).* 2004;5(1):75-87.

86. Morgan RO, Virnig BA, DeVito CA, Persily NA. The Medicare-HMO revolving door--the healthy go in and the sick go out. *New Engl J Med.* 1997;337(3):169-175.

87. Heart Protection Study Collaborative G. The effects of cholesterol lowering with simvastatin on cause-specific mortality and on cancer incidence in 20,536 high-risk people: a randomised placebo-controlled trial [ISRCTN48489393]. *BMC Medicine.* 2005;3:6-7015-7013-7016.

88. Rothman KJ. Induction and latent periods. *Am J Epidemiol.* 1981;114(2):253-259.

89. Fine JP, Gray RJ. A proportional hazards model for the subdistribution of a competing risk. *J Am Stat Assoc.* 1999;94(446):496-509.

90. Lau B, Cole SR, Gange SJ. Competing risk regression models for epidemiologic data. *Am J Epidemiol.* 2009;170(2):244-256.

91. Wolin KY, Carson K, Colditz GA. Obesity and cancer. *Oncologist.* 2010;15(6):556-565.

92. Smith RA, von Eschenbach AC, Wender R, et al. American Cancer Society guidelines for the early detection of cancer: update of early detection guidelines for prostate, colorectal, and endometrial cancers. Also: update 2001--testing for early lung cancer detection. *CA Cancer J Clin.* 2001;51(1):38-75; quiz 77-80.

93. Yang P, Zhou Y, Chen B, et al. Overweight, obesity and gastric cancer risk: results from a meta-analysis of cohort studies. *Eur J Cancer (Oxford, England : 1990).* 2009;45(16):2867-2873.

94. Turati F, Tramacere I, La Vecchia C, Negri E. A meta-analysis of body mass index and esophageal and gastric cardia adenocarcinoma. *Ann Oncol.* 2013;24(3):609-617.

95. Pischon T, Lahmann PH, Boeing H, et al. Body size and risk of renal cell carcinoma in the European Prospective Investigation into Cancer and Nutrition (EPIC). *Int J Cancer.* 2006;118(3):728-738.

96. Key TJ, Appleby PN, Reeves GK, et al. Body mass index, serum sex hormones, and breast cancer risk in postmenopausal women. *J Natl Cancer Inst.* 2003;95(16):1218-1226.

97. American Cancer Society. *Cancer Facts & Figures 2014.* 2014; http://www.cancer.org/acs/groups/content/@research/documents/webcontent/acspc-042151.pdf. Accessed September 26, 2017.

98. Gawron A, Hou L, Ning H, Berry JD, Lloyd-Jones DM. Lifetime risk for cancer death by sex and smoking status: the lifetime risk pooling project. *Cancer Cause Control.* 2012;23(10):1729-1737.

99. Lin DY, Wei L-J. The robust inference for the Cox proportional hazards model. *J Am Stat Assoc.* 1989;84(408):1074-1078.

100. Walker AM, Patrick AR, Lauer MS, et al. A tool for assessing the feasibility of comparative effectiveness research. *J Comp Effect Res.* 2013;2013(1):11-20.

101. Institute P-COR. PCORnet: The National Patient-Centered Clinical Research Network. 2017; https://www.pcori.org/research-results/pcornet-national-patient-centered-clinical-research-network. Accessed September 26, 2017.

102. Behrman RE, Benner JS, Brown JS, McClellan M, Woodcock J, Platt R. Developing the Sentinel System--a national resource for evidence development. *N Engl J Med.* 2011;364(6):498-499.

103. Psaty BM, Breckenridge AM. Mini-Sentinel and regulatory science--big data rendered fit and functional. *N Engl J Med.* 2014;370(23):2165-2167.

104. Conover MM, Jonsson Funk M. Uniform vs. all-available look-backs to identify exclusion criteria in observational cohort studies [abstract]. 2015.

105. Yang D, Dalton JE. *A unified approach to measuring the effect size between two groups using SAS®.* 2012; http://support.sas.com/resources/papers/proceedings12/335-2012.pdf. Acessed September 26, 2017.

106. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika.* 1983;70(1):41-55.

107. Patorno E, Grotta A, Bellocco R, Schneeweiss S. Propensity score methodology for confounding control in health care utilization databases. *Epidemiol Biostat Public Health.* 2013;10(3).

108. Setoguchi S, Glynn RJ, Avorn J, Mogun H, Schneeweiss S. Statins and the risk of lung, breast, and colorectal cancer in the elderly. *Circulation.* 2007;115(1):27-33.

109. Jukema JW, Bruschke AV, van Boven AJ, et al. Effects of lipid lowering by pravastatin on progression and regression of coronary artery disease in symptomatic men with normal to moderately elevated serum cholesterol levels. The Regression Growth Evaluation Statin Study (REGRESS). *Circulation.* 1995;91(10):2528-2540.

110. Pitt B, Mancini GB, Ellis SG, Rosman HS, Park JS, McGovern ME. Pravastatin limitation of atherosclerosis in the coronary arteries (PLAC I): reduction in atherosclerosis progression and clinical events. PLAC I investigation. *J Am Coll Cardiol.* 1995;26(5):1133-1139.

111. Stürmer T, Schneeweiss S, Brookhart MA, Rothman KJ, Avorn J, Glynn RJ. Analytic strategies to adjust confounding using exposure propensity scores and disease risk scores: nonsteroidal antiinflammatory drugs and short-term mortality in the elderly. *Am J Epidemiol.* 2005;161(9):891-898.

112. Lash TL, Fox MP, Fink AK. *Applying Quantitative Bias Analysis to Epidemiologic Data.* Springer New York; 2009.

113. Centers for M, Medicaid S. Medicare Current Beneficiary Survey (MCBS). In. Vol 20152015. https://www.cms.gov/Research-Statistics-Data-and-Systems/Research/MCBS/index.html. Accessed September 26, 2017.

114. Petroski J, Ferraro D, Chu A. Ever enrolled Medicare population estimates from the MCBS Access to Care files. *Medicare Medicaid Res Rev.* 2014;4(2).

115. Mann E. Differences between the Cost and Use and Access to Care modules. In. Vol 20152012. https://www.resdac.org/resconnect/articles/103. Accessed September 26, 2017.

116. Faurot KR, Jonsson Funk M, Pate V, et al. Using claims data to predict dependency in activities of daily living as a proxy for frailty. *Pharmacoepidemiol Drug Saf.* 2015;24(1):59-66.

117. Katz S, Downs TD, Cash HR, Grotz RC. Progress in development of the index of ADL. *Gerontologist.* 1970;10(1):20-30.

118. Noelker LS, Browdie R. Sidney Katz, MD: a new paradigm for chronic illness and long-term care. *Gerontologist.* 2014;54(1):13-20.

119. Depalma G, Xu H, Covinsky KE, et al. Hospital readmission among older adults who return home with unmet need for ADL disability. *Gerontologist.* 2013;53(3):454-461.

120. Keeler E, Guralnik JM, Tian H, Wallace RB, Reuben DB. The impact of functional status on life expectancy in older persons. *J Gerontol A Biol Sci Med Sci.* 2010;65(7):727-733.

121. Young JC, Stürmer T, Lund JL, Funk MJ. Predictors of prevalent statin use among older adults identified as statin initiators based on Medicare claims data. *Pharmacoepidemiol Drug Saf.* 2016;25(7):836-843.

122. Wacholder S, Silverman DT, McLaughlin JK, Mandel JS. Selection of controls in case-control studies. II. Types of controls. *Am J Epidemiol.* 1992;135(9):1029-1041.

123. Conover MM, Howell JO, Wu JM, Kinlaw AC, Dasgupta N, Jonsson Funk M. Incidence of opioid-managed pelvic pain after hysteroscopic sterilization versus laparoscopic sterilization, U.S. 2005-2012 (In-Press). *Pharmacoepidemiol Drug Saf.* 2015.

124. Rothman KJ, Greenland S, Lash TL. Measures of Effect and Measures of Association. In: Seigafuse S, Bierig L, eds. *Modern Epidemiology.* 3rd ed. Philadelphia (PA): Lippincott Williams & Wilkins; 2008:61.

125. Cadmus-Bertram L, Marcus BH, Patterson RE, Parker BA, Morey BL. Use of the Fitbit to Measure Adherence to a Physical Activity Intervention Among Overweight or Obese, Postmenopausal Women: Self-Monitoring Trajectory During 16 Weeks. *JMIR Mhealth Uhealth.* 2015;3(4):e96.

126. Wood WA, Basch E. From Intuition to Execution: Realizing the Potential of Wearables in Oncology. *J Oncol Pract.* 2017;13(2):90-92.

127. Avery TR, Kulldorff M, Vilk Y, et al. Near real-time adverse drug reaction surveillance within population-based health networks: methodology considerations for data accrual. *Pharmacoepidemiol Drug Saf.* 2013;22(5):488-495.

128. Hoy MB. An Introduction to the Blockchain and Its Implications for Libraries and Medicine. *Med Ref Serv Q.* 2017;36(3):273-279.

129. Yue X, Wang H, Jin D, Li M, Jiang W. Healthcare Data Gateways: Found Healthcare Intelligence on Blockchain with Novel Privacy Risk Control. *J Med Syst.* 2016;40(10):218.

130. Bradley CJ, Penberthy L, Devers KJ, Holden DJ. Health services research and data linkages: issues, methods, and directions for the future. *Health Serv Res.* 2010;45(5 Pt 2):1468-1488.

131. Bell RM, Keesey J, Richards T. The urge to merge: linking vital statistics records and Medicaid claims. *Med Care.* 1994;32(10):1004-1018.