Observational Studies 2 (2016) 51-64 Submitted 9/15; Published 8/16

Cohort restriction based on prior enrollment: Examining potential biases in estimating cancer and mortality risk

Susan M. Shortreed **shortreed.s@ghc.org**
Biostatistics Unit, Group Health Research Institute, Seattle, WA, U.S.A.
Department of Biostatistics, University of Washington, Seattle, WA, U.S.A.

Eric Johnson **johnson.ex@ghc.org**
Biostatistics Unit, Group Health Research Institute, Seattle, WA, U.S.A.

Carolyn M. Rutter **crutter@rand.org**
RAND Corporation, Santa Monica, CA, U.S.A.

Aruna Kamineni **kamineni.a@ghc.org**
Group Health Research Institute, Seattle, WA, U.S.A.

Karen J. Wernli **wernli.k@ghc.org**
Group Health Research Institute, Seattle, WA, U.S.A.

Jessica Chubak **chubak.j@ghc.org**
Group Health Research Institute, Seattle, WA, U.S.A.
Department of Epidemiology, University of Washington, Seattle, WA, U.S.A

Abstract

Electronic health records and administrative databases provide rich, longitudinal data for health-related research. These data cover large, diverse populations creating excellent research opportunities, but have limitations. In particular, information is available only for individuals who are enrolled in a particular health system; thus, studies often exclude individual's with short enrollment history. Such cohort restriction may cause selection bias in absolute risk estimates for the full enrollee population. We use hazard ratios (HRs) to estimate the association between length of prior enrollment and cancer and all-cause mortality risk. HRs different from one indicate restricted cohorts would produce biased risk estimates for the full enrollee population. Our study sample included 170,708 enrollees of a Western Washington healthcare delivery system. Unadjusted models found individuals with 10 or more years of prior enrollment had higher risk of cancer and death compared to those with less than 5 years prior enrollment (HRs ranged from $1.29 - 3.01$). Age- and sex-adjusted models accounted for much of this difference (HRs: $0.93 - 1.24$). Models adjusting for additional covariates had similar results (HRs: $0.91 - 1.14$). After evaluating potential selection bias, we conclude that, in this setting, age- and sex-standardizing risk estimates can remove most of the bias due to lengthy, prior-enrollment cohort restrictions. Before generalizing estimates based on a selected sample of patients meeting prior enrollment criteria, researchers should assess the potential for selection bias.

Keywords: Selection bias, electronic health records, enrollment history, cancer mortality risk estimation, administrative databases

1. Introduction

Data collected from electronic health records (EHRs) and administrative databases can provide rich longitudinal information on large and diverse patient populations to address important scientific questions (Ryan et al., 2010) Existing and newly forming partnerships among health care systems and research organizations are increasing access to these data sources for health care research (Chen et al., 2000; Reisinger et al., 2010; McCarty et al., 2011; Fleurence et al., 2014; Nelson et al., 2014; Ross et al., 2014). Notably, longitudinal data are available on increasingly larger cohorts of individuals, making it possible to estimate the risk of rare events following long-term exposures, such as cancer screening practices.

While large clinical and administrative databases provide many research opportunities, potential biases can limit their use. In particular, while age and sex are often available on all enrollees, more detailed patient information is generally only recorded when a person is seen at a particular health care facility or while insured by the health care organization. When an individual disenrolls from their health system, information on their subsequent care is no longer observable. Similarly, when an individual joins a health system, health care utilization received prior to enrollment is not observed.

The absence of data to form a complete picture of an individual's health care over a period of interest can be a particular challenge for epidemiologic studies designed to evaluate the relationship between exposures, like medication use or cancer screening behavior, and subsequent cancer risk. Long periods of enrollment might be required on study samples to ensure accurate exposure data capture from an EHR or administrative database. For example, average-risk individuals are recommended to begin screening for colorectal cancer at age 50 in the US (Preventive Services Task Force, 2008). Colonoscopy is a common method for colorectal cancer screening. Individuals who undergo a colonoscopy and have no high-risk finding are recommended to wait 10 years for their next screening test. Thus, to evaluate cancer risk after repeat colorectal cancer screening tests, long periods of continuous enrollment is required to obtain colonoscopy cancer screening behavior consistently among study participants.

It is common for studies using EHR or administrative data to restrict the study cohort to individuals with a minimum number of months of prior enrollment, with the required length of enrollment varying based on the scientific question of interest (McBean et al., 1993; Voordouw et al., 2004; Jackson et al., 2006; Hernandez-Diaz and Garcia Rodriguez, 2007; Bird et al., 2013; Black et al., 2013; Ryan et al., 2013; Turner et al., 2014; Gray et al., 2015; Nichols et al., 2015). This cohort restriction could result in bias when the risk of disease in the entire population is the estimand of interest and the risk of disease is associated with the cohort selection criteria (Hernn et al., 2004). Specifically, if individuals enrolled in health systems for long periods of time have a lower or higher risk of cancer, then estimates of cancer risk generated from these longer-term enrollees may not represent the true risk across the complete enrollee population. In this paper, we investigate the potential for selection bias based on length of prior enrollment in a study using EHR and administrative data. Our goal was to assess whether restricting cohort studies based on prior enrollment created a cohort with a different risk of cancer or death than the full enrollee population.

2. Methods

2.1 Study population

The study population includes individuals aged 40 years and older enrolled in Group Health on the cohort entry date of January 1st, 2008 who have selected a Group Health clinician for their primary care and reside in the Puget Sound Surveillance, Epidemiology, and End Results (SEER) (Cancer Institute and DCCPS and Surveillance Research Program and Surveillance Systems Branch, 2011) catchment area. Group Health is a not-for-profit mixed-model health care delivery system located in Washington State. We excluded all individuals with a known diagnosis of breast, cervical, colorectal, lung or prostate cancer on or before December 31st, 2007.

2.2 Outcomes and covariates

We considered the 5-year risk of incident breast, cervical, colorectal, lung, and prostate cancer as well as all-cause mortality. Cancer incidence data were obtained from Seattle-Puget Sound SEER (Cancer Institute and DCCPS and Surveillance Research Program and Surveillance Systems Branch, 2011). Deaths were collected from Group Health administrative data that draw on a number of sources including the Washington State Department of Health. Event data were obtained through December 31st, 2012, or up until individuals died or were censored due to disenrollment from Group Health.

We obtained a patients age and sex from administrative data available for all enrollees. Additionally, for individuals enrolled in Group Health for the year prior to cohort entry (i.e. from January 1 to December 31, 2007) we collected race, tobacco use, Charlson comorbidity score (Charlson et al., 1987) and body mass index from EHR data. We formed three groups of individuals based on their length of enrollment prior to January 1, 2008: less than 5 years, 5 to10 years, and 10 or more years.

2.3 Statistical analyses

We described differences in measured covariates between individuals enrolled for differing lengths of time prior to cohort entry using means and standard deviations for continuous-valued variables, with corresponding ANOVA p-values and percentages and raw numbers for categorical variables, with corresponding chi-squared p-values. We report the median and interquartile range for the length of observed follow-up for each of the enrollment groups as well as the proportion of individuals who were enrolled for the full 5-year follow-up period.

We estimated hazard ratios (HRs) comparing the 5-year risk of each outcome in the two groups with longest enrollment history (5 to 10 years and 10 or more years) compared to individuals with the shortest enrollment history (less than 5 years). HRs different from one imply unadjusted risk estimates obtained from a cohort restricted by enrollment history would be biased estimates of the true risk in the full enrollee population. We first estimated separate unadjusted Cox proportional hazard models for each cancer considered and all-cause mortality using length of prior enrollment as a predictor to estimate the differences in risk for individuals with varying amounts of prior enrollment.

We fit parsimonious age- and sex-adjusted Cox proportional hazards models. We note here that three of the cancers we considered (prostate, cervical, and breast) are sex-specific

diseases, thus corresponding parsimonious models included only age as a covariate. If age- and sex-adjusted HRs are different from one, then accounting for age and sex difference between the full enrollee population and the restricted cohort would not fully account for selection bias. Conversely, estimated HRs close to one indicate that risk estimates calculated in a restricted cohort standardized to the age and sex distribution of the full enrollee population would lead to unbiased risk estimates for the full population.

Lastly, we estimated a Cox model that adjusted for demographic information and general health history in addition to age and sex. Specifically, this accounted for patient age (40-49, 50-59, 60-69, 70-79, and 80+ years), sex, race and ethnicity (Asian, Non-Hispanic Black, Hispanic, Native Hawaiian/Pacific Islander, Native American/Alaskan Native, non-Hispanic White, Multiple Races, Unknown), tobacco use (never, ever current), Charlson score (0,1,2+), and body mass index (<25, 25-30, 30-35, 35-40, 40+ kg/m^2). This model was estimated in a cohort of individuals who were enrolled at least 1 year prior to January 1, 2008 to ensure covariate information was available. Thus, the shortest enrollment category for this model was less than 5 years but at least one year prior enrollment. Since this reference group definition is different than the reference group for the unadjusted and age- and sex-adjusted models, we refit both models restricting the reference group to patients with at least one year of prior enrollment.

We estimated and plotted Kaplan-Meier curves (Kaplan and Meier, 1958; Cole and Hernn, 2004) for the full enrollee population for all cancer outcomes considered and all-cause mortality. To visually compare results from a restricted cohort analysis age- and sex-standardized to the full enrollee population, we plotted inverse probability weighted Kaplan-Meier curves for the cohort restricted to those individuals enrolled in Group Health for 10 or more years. We constructed age- and sex-standardized inverse probability weights for each individual in the enrollment restricted cohort by calculating the probability of the individual belonging to an age- and sex-strata in the full enrollee population divided by the probability of belonging to the same strata in the restricted cohort. All analyses were performed in Stata Version 13.1 (StataCorp, 2013), and this study was approved by the Group Healths institutional review board.

3. Results

A total of 170,708 individuals aged 40 to105 years were included in our study sample. Of these, 31% had been enrolled in Group Health for less than 5 years (median: 1.9 years of prior enrollment), 17% had been enrolled in Group Health for between 5 and 10 years (median: 7.1 years), and 52% had been enrolled for 10 or more years at cohort entry (median: 15.1 years). We report the differences among measured covariates by length of prior enrollment in Table 1. Notably, those who had been enrolled for 10 or more years were on average older, 62.0 years with a standard deviation (sd) of 12.7 years compared to an average age of 55.8 years (sd=11.4 years) for individuals enrolled for 5 to 10 years and 52.8 years (sd=9.6 years) for those enrolled less than 5 years. The sex distribution was similar, about 54% female, across enrollment groups, as was body mass index, with a mean of about 29 kg/m^2 in all three groups.

The distribution of tobacco use was slightly different between the groups with 47% of individuals with the shortest prior enrollment having never smoked, compared to 56% and

58% in the groups with 5 to 10 years and 10 or more years of prior enrollment, respectively. As expected, since those with longer enrollment history were on average older, Charlson score increased as the length of prior enrollment increased. Of individuals who had 10 or more years of prior enrollment, 74% had a Charlson score of 0, while 80% of individuals with between $5-10$ years and 84% of those with less than 5 years of prior enrollment had a Charlson score of 0. All p-values associated with ANOVAs and chi-squared tests evaluating differences in measured covariates across the three enrollment groups were less than 0.001.

Length of follow-up was also slightly different between the three groups with 61% (n=32099) of individuals with the shortest enrollment history having the full 5 years of follow-up, while 77% (n=22768) of those with 5 to 10 years and 89% (n=78735) of those with more than 10 years of enrollment had the full 5 years of follow-up. The median length of follow-up time for all three enrollment groups was 5 years.

During the 5 year follow-up period, we observed 1514 breast, 387 cervical, 684 colorectal, 829 lung, and 982 prostate cancer incident events, and 9246 deaths due to any cause (Table 2). This corresponds to a 5-year mortality risk rate, estimated from the Kaplan-Meier curve in the full enrollee population, of 0.0625 with a 95% confidence interval (CI) of (0.0613, 0.0638). The 5-year incidence of each of the cancer types under consideration were as follows: 0.0105 (0.0100, 0.0110) for breast, 0.0027 (0.0024, 0.0030) for cervical, 0.0047 (0.0043, 0.0050) for colorectal, 0.0057 (0.0053, 0.0061) for lung, and 0.0068 (0.0064, 0.0072) for prostate.

Results from unadjusted models indicated a higher risk for all cancers considered and all-cause mortality in the group of individuals with the longest prior enrollment (>10 years) compared to those with less than 5 years of prior enrollment; HRs ranged from $1.29-3.01$, with all 95% CIs excluding one. HRs for individuals with between 5 and 10 years prior enrollment were closer to one, but still showed a higher risk of incident colorectal, lung, and prostate cancer as well as all-cause mortality compared with individuals with less than 5 years of prior enrollment (HRs ranged from $0.98-1.56$).

Adjusting for age and sex produced HRs close to one for both longer prior enrollment groups for all cancers considered and all-cause mortality (HRs ranged from 0.93 to 1.24). For all outcomes, models that adjusted for demographic information and general health history had similar HR estimates as the model that adjusted for only age and sex (HRs ranged from 0.91 to 1.14). Estimates from the unadjusted and age- and sex- adjusted models with the same reference group as the demographic and general health history model were very similar to the results presented in Table 2 (see Appendix). As shown in Figure 1, age- and sex-standardized Kaplan-Meier curves for the restricted cohort requiring 10+ years enrollment prior to cohort entry, produce Kaplan-Meier curves very similar to Kaplan-Meier curves constructed in the full enrollee population.

4. Conclusion

In this study, length of prior enrollment was strongly associated with risk of all cancers considered and mortality in unadjusted models. This increased risk appeared to be mostly due to the fact that individuals with the longest prior enrollment were on average older than the other enrollment groups. Accounting for age and sex in restricted cohort models for all outcomes resulted in HRs that were very near one, with all 95% CIs including one,

Table 1: Distribution of covariates by length of prior enrollment.

Covariates	Length of Group Health enrollment prior to January 1, 2008: less than 5 years (N=52582)	5-10 years† (N=29637)	10 or more years (N=88489)
Age: 40-49 years	43% (22443)	34% (10188)	16% (14369)
50-59 years	36% (18923)	34% (9960)	33% (28932)
60-69 years	16% (8345)	18% (5337)	24% (21667)
70-79 years	4% (1865)	10% (2870)	14% (12684)
80+ years	2% (1006)	4% (1282)	12% (10837)
Sex: female	54% (28433)	54% (16097)	55% (48965)
Body mass index* : < 25 kg/m^2	24% (9090)	27% (7946)	28% (24813)
25-30 kg/m^2	30% (11353)	33% (9867)	35% (31015)
30-35 kg/m^2	18% (6764)	19% (5575)	20% (17272)
35-30 kg/m^2	8% (3112)	8% (2437)	8% (7176)
40+ kg/m^2	6% (2158)	6% (1690)	5% (4570)
Unknown	15% (5873)	7% (2122)	4% (3643)
Tobacco use*: Never	47% (18082)	56% (16489)	58% (51608)
Previous	25% (9640)	28% (8260)	31% (27406)
Current	15% (5621)	13% (3828)	9% (8208)
Unknown	13% (5007)	4% (1060)	1% (1267)
Race*: Asian	9% (3403)	9% (2650)	5% (4827)
Non-Hispanic Black	4% (1584)	4% (1198)	4% (3218)
Hispanic	2% (712)	2% (571)	2% (1912)
Native Hawaiian/Pacific Islander	1% (257)	1% (199)	1% (442)
Native American/Alaskan Native	1% (266)	1% (252)	1% (643)
non-Hispanic White	56% (21382)	65% (19335)	77% (68464)
Multiple Races	2% (650)	2% (486)	1% (1230)
Unknown	26% (10096)	17% (4946)	9% (7753)
Charlson score*: 0	84% (32049)	80% (23615)	74% (65818)
1	11% (4339)	14%(4119)	16% (14194)
2+	5% (1962)	6% (1903)	10% (8477)
Number of years prior enrollment‡	1.92 (0.8, 3.2)	7.1 (6.1, 8.2)	15.1 (15.1, 15.1)
Number of years of follow-up‡	5.0 (2.75, 5.0)	5.0 (5.0, 5.0)	5.0 (5.0, 5.0)

*In order to obtain covariate information beyond age and sex, some minimal amount of enrollment criteria (1 year) was needed. This reduced the number of individuals in the shortest enrollment history group to 38350, which we used as the denominator for calculating percentages in the shortest enrollment period for body mass index, tobacco use, race/ethnicity and Charlson score.
†Enrollment group includes individuals enrolled at least 5 years but less than 10 years.
‡Median with interquartile range in parentheses.

Table 2: Comparison of cancer and all-cause mortality hazard ratios by prior enrollment duration.

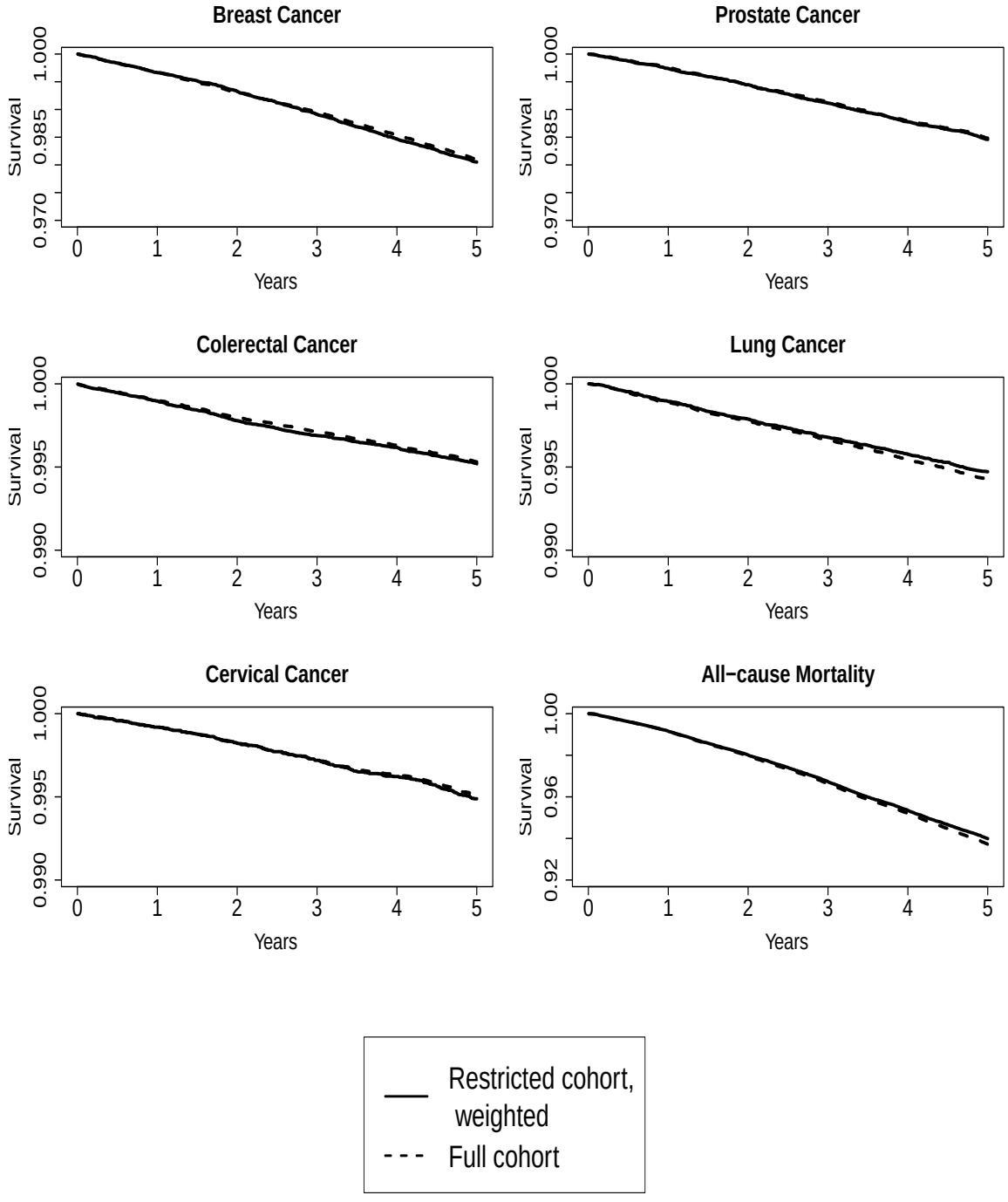
Length of prior enrollment	Breast cancer	Cervical cancer	Colorectal cancer	Lung cancer	Prostate cancer	All-cause mortality
			Number of events			
< 5 years	345	90	110	156	168	1144
≥ 5 & <10 years	258	56	102	133	143	1144
≥ 10 years	911	241	472	540	671	6958
			Unadjusted hazard ratios [95% CI]			
< 5 years	Ref	Ref	Ref	Ref	Ref	Ref
≥ 5 & <10 years	1.17	0.98	1.47	1.35	1.34	1.56
	[1.00‡, 1.38]	[0.70, 1.36]	[1.12, 1.93]	[1.07, 1.70]	[1.07, 1.68]	[1.44, 1.70]
≥ 10 years	1.29	1.31	2.17	1.74	2.05	3.01
	[1.15, 1.47]	[1.03, 1.67]	[1.76, 2.67]	[1.46, 2.08]	[1.73, 2.42]	[2.83, 3.20]
			Age- and sex-adjusted hazard ratios [95% CI]			
< 5 years	Ref	Ref	Ref	Ref	Ref	Ref
≥ 5 & <10 years	1.09	0.93	1.16	0.99	1.09	0.96
	[0.93, 1.28]	[0.67, 1.30]	[0.88, 1.52]	[0.78, 1.25]	[0.87, 1.36]	[0.89, 1.05]
≥ 10 years	1.09	1.13	1.24	0.86	1.17	0.98]
	[0.96, 1.25]	[0.88, 1.45]	[1.00, 1.55]	[0.72, 1.04]	[0.98, 1.39]	[0.92, 1.05]
			Demographic and general health history adjusted hazard ratios [95% CI]*			
< 5 years†	Ref	Ref	Ref	Ref	Ref	Ref
≥ 5 & <10 years	1.05	0.91	1.08	1.05	1.07	0.99
	[0.89,1.25]	[0.64, 1.30]	[0.81, 1.44]	[0.82, 1.35]	[0.84, 1.36]	[0.90, 1.08]
≥ 10 years	1.05	1.08	1.14	0.98	1.09	0.99
	[0.90,1.21]	[0.81, 1.43]	[0.90, 1.45]	[0.80, 1.21]	[0.90, 1.32]	[0.93, 1.07]

*Demographic and general health history model included as covariates: age, sex, race, Charlson score, tobacco use and body mass index.

†Reference group includes individuals who had at least 1 year prior enrollment and less than 5 years of enrollment, in contrast to unadjusted and age- and sex-adjusted models, which includes all individuals with less than 5 years prior enrollment, including those with less than 1 year enrollment.

‡Round forces bound to be equal to 1.00, but true confidence interval does not include one.

Figure 1: Kaplan-Meier curves in the full enrollee population (Full Cohort) and weighted Kaplan-Meier curves in the cohort restricted to individuals with at 10+ years prior enrollment (Restricted cohort, weighted). Weighted analyses standardize the restricted cohort to the age and sex distribution of the full enrollee population. We restricted the y-axis to allow for visual comparison of Kaplan-Meier curves.



and Kaplan-Meier curves very similar to those in the full enrollee population. Accounting for additional covariates resulted in minimal change in HRs; it appears that age and sex accounted for the majority of the risk differences by length of prior enrollment.

There is a growing literature on the impact of imposing enrollment criteria in records-based observational studies to select cohorts and capture covariate information on potential confounders. In particular, Roberts et al. studied the impact of varying the length of required prior enrollment in new user study designs for pharmacoepidemiologic studies (Roberts et al., 2015). It is common in such studies to require a washout period, a period of time in which no medicine dispensing has occurred, in order to define a cohort of "new users" of a medication. Roberts et al. (2015) found misclassification rates for new medication use of approximately 50% using a washout of 6 months and of 30% using a washout of 12 months, suggesting that a 12 month or less washout period may be insufficient for a new user study design (Roberts et al., 2015). Riis et al. (2015) found similar results in their analysis of lengths of look-back periods for new user studies, recommending two years or longer for some medication types (Riis et al., 2015). Enforcing longer washout periods for studies, requires restricting cohorts to populations of individuals who are enrolled and have coverage information available during the full time period.

Some work has reported on the uses of variable length look-back periods to define covariates. That is, authors assessed whether study investigators should use the same look-back time to define covariates for all individuals or use all available information even if the time frame that covariates are defined over varies across individuals included in the study. Both studies concluded that using all available information could reduce potential confounding, but both studies required a minimum amount of enrollment on all cohort members (Brunelli et al., 2013; Gilbertson et al., 2015). Gilbertson et al. (2015) acknowledged that defining a look-back period is a balancing act that takes into account potential misclassification bias on one hand and potential selection bias on the other. Our study focused on this latter type of bias investigators must weigh the impact of when designing a study. That is, the potential for selection bias to be induced by restricting a cohort to have long periods of prior enrollment in order to ascertain exposure and covariate information. As more studies implement enrollment restrictions, it is increasingly important to understand the impact of such restrictions on the validity of results.

This study was conducted using data gathered from Group Health members aged 40 years and older, thus the generalizability of these results to younger age groups and other outcomes should be explored. We were able to identify mortality and cancer events for individuals only if they were enrolled in Group Health during the follow-up period. Since, the length of follow-up across the different enrollment groups varied, our estimates could have been influenced by censoring; although the majority of individuals were enrolled in Group Health for all 5 years of follow-up. Additionally, we were only able to ascertain prior cancer occurrence for individuals while they were enrolled at Group Health; it is likely that we have more complete cancer information history for those individuals who were enrolled longer prior to January 1, 2008. Therefore, it is possible that we included more individuals with a prior cancer diagnosis in the group with less than 5 years prior enrollment than in the groups with longer enrollment history. Race and ethnicity, as well as tobacco use and body mass index, were missing on some individuals; had this information been observed on everyone it is possible that these covariates could have accounted for some additional

residual bias. Lastly, we were unable to ascertain family history of cancer, a risk factor for both cancer risk and mortality, to assess how adjustment for family history might further reduce residual bias.

We conclude that longitudinal, cohort studies investigating incident cancer and all-cause mortality risk that use a restricted cohort based on length of prior enrollment may produce valid risk estimates for the full enrollee population if analyses account for age and sex. Accounting for additional demographic and health history information could further reduce potential bias, but based on the results in this setting, age- and sex-adjusted models appear to adequately account for much of the apparent difference in cancer incidence and all-cause mortality among persons with different lengths of prior enrollment. This finding is extremely promising because age and sex are often completely captured in EHR and administrative databases. Researchers conducting studies using restricted cohorts should evaluate the potential for selection bias in their particular setting before making conclusions about the full enrollee population.

Acknowledgments

The collection of cancer incidence data used in this study was supported by the Cancer Surveillance System of the Fred Hutchinson Cancer Research Center, which is funded by Contract No. N01-CN-67009 and N01-PC-35142 from the Surveillance, Epidemiology and End Results (SEER) Program of the National Cancer Institute with additional support from the Fred Hutchinson Cancer Research Center and the State of Washington. This study was conducted as part of the NCI-funded consortium Population-Based Research Optimizing Screening through Personalized Regiments (PROSPR) (U54CA163261, Chubak). The overall aim of PROSPR is to conduct multi-site, coordinated, transdisciplinary research to evaluate and improve cancer screening processes. The seven PROSPR Research Centers reflect the diversity of US delivery system organizations.

Conflicts of interests: Dr. Shortreed has received funding from research grants awarded to GHRI by Bristol-Myers Squibb and Pfizer Inc and is a Co-Investigator on a grant awarded to GHRI from the Campbell Alliance, a consortium of pharmaceutical companies carrying out FDA-mandated studies regarding the safety of extended release opioids. She has also received funding to attend review panels and methods meetings through the Patient-Centered Outcome Research Institute.

References

Bird, S. T., Delaney, J. A., Brophy, J. M., Etminan, M., Skeldon, S. C., and Hartzema, A. G. (2013). Tamsulosin treatment for benign prostatic hyperplasia and risk of severe hypotension in men aged 40-85 years in the united states: risk window analyses using between and within patient methodology. *BMJ*, 347:f6320.

Black, M. H., Zhou, H., Takayanagi, M., Jacobsen, S. J., and Koebnick, C. (2013). Increased asthma risk and asthma-related health care complications associated with childhood obesity. *Am J Epidemiol*, 178(7):1120–8.

Brunelli, S. M., Gagne, J. J., Huybrechts, K. F., Wang, S. V., Patrick, A. R., Rothman, K. J., and Seeger, J. D. (2013). Estimation using all available covariate information versus a fixed look-back window for dichotomous covariates. *Pharmacoepidemiology and Drug Safety*, 22(5):542–550.

Cancer Institute and DCCPS and Surveillance Research Program and Surveillance Systems Branch, N. (2011). Surveillance, epidemiology, and end results (SEER) program research data (1973-2011).

Charlson, M. E., Pompei, P., Ales, K. L., and MacKenzie, C. R. (1987). A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *J Chronic Dis*, 40(5):373–83.

Chen, R. T., DeStefano, F., Davis, R. L., Jackson, L. A., Thompson, R. S., Mullooly, J. P., Black, S. B., Shinefield, H. R., Vadheim, C. M., Ward, J. I., and Marcy, S. M. (2000). The vaccine safety datalink: immunization research in health maintenance organizations in the usa. *Bull World Health Organ*, 78(2):186–94.

Cole, S. R. and Hernn, M. A. (2004). Adjusted survival curves with inverse probability weights. *Comput Methods Programs Biomed*, 75(1):45–9.

Fleurence, R. L., Curtis, L. H., Califf, R. M., Platt, R., Selby, J. V., and Brown, J. S. (2014). Launching pcornet, a national patient-centered clinical research network. *J Am Med Inform Assoc*, 21(4):578–82.

Gilbertson, D. T., Bradbury, B. D., Wetmore, J. B., Weinhandl, E. D., Monda, K. L., Liu, J., Brookhart, M. A., Gustafson, S. K., Roberts, T., Collins, A. J., and Rothman, K. J. (2015). Controlling confounding of treatment effects in administrative data in the presence of time-varying baseline confounders. *Pharmacoepidemiol Drug Saf.*

Gray, S. L., Anderson, M. L., Dublin, S., Hanlon, T. J., Hubbard, R., Walker, R., Yu, O., Crane, P. K., and Larson, E. B. (2015). Cumulative use of strong anticholinergics and incident dementia: A prospective cohort study. *JAMA Internal Medicine*, 175(3):401–407.

Hernandez-Diaz, S. and Garcia Rodriguez, L. A. (2007). Nonsteroidal anti-inflammatory drugs and risk of lung cancer. *Int J Cancer*, 120(7):1565–72.

Hernn, M. A., Hernandez-Diaz, S., and Robins, J. M. (2004). A structural approach to selection bias. *Epidemiology*, 15(5):615–625.

Jackson, L. A., Jackson, M. L., Nelson, J. C., Neuzil, K. M., and Weiss, N. S. (2006). Evidence of bias in estimates of influenza vaccine effectiveness in seniors. *Int J Epidemiol*, 35(2):337–44.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *J Am Stat Assoc*, 53(282):457–481.

McBean, A. M., Babish, J. D., and Warren, J. L. (1993). Determination of lung cancer incidence in the elderly using medicare claims data. *Am J Epidemiol*, 137(2):226–34.

McCarty, C. A., Chisholm, R. L., Chute, C. G., Kullo, I. J., Jarvik, G. P., Larson, E. B., Li, R., Masys, D. R., Ritchie, M. D., Roden, D. M., Struewing, J. P., and Wolf, W. A. (2011). The emerge network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. *BMC Med Genomics*, 4:13.

Nelson, J. C., Shortreed, S. M., Yu, O., Peterson, D., Baxter, R., Fireman, B., Lewis, N., McClure, D., Weintraub, E., Xu, S., Jackson, L. A., and on behalf of the Vaccine Safety Datalink project (2014). Integrating database knowledge and epidemiological design to improve the implementation of data mining methods that evaluate vaccine safety in large healthcare databases. *Statistical Analysis and Data Mining*, 7(5):337–351.

Nichols, G. A., Schroeder, E. B., Karter, A. J., Gregg, E. W., Desai, J., Lawrence, J. M., O'Connor, P. J., Xu, S., Newton, K. M., Raebel, M. A., Pathak, R. D., Waitzfelder, B., Segal, J., Lafata, J. E., Butler, M. G., Kirchner, H. L., Thomas, A., Steiner, J. F., and Group, S.-D. S. (2015). Trends in diabetes incidence among 7 million insured adults, 2006-2011: the supreme-dm project. *Am J Epidemiol*, 181(1):32–9.

Preventive Services Task Force, U. (2008). Screening for colorectal cancer: U.s. preventive services task force recommendation statement. *Ann Intern Med*, 149(9):627–37.

Reisinger, S. J., Ryan, P. B., O'Hara, D. J., Powell, G. E., Painter, J. L., Pattishall, E. N., and Morris, J. A. (2010). Development and evaluation of a common data model enabling active drug safety surveillance using disparate healthcare databases. *J Am Med Inform Assoc*, 17(6):652–62.

Riis, A. H., Johansen, M. B., Jacobsen, J. B., Brookhart, M. A., Sturmer, T., and Stovring, H. (2015). Short look-back periods in pharmacoepidemiologic studies of new users of antibiotics and asthma medications introduce severe misclassification. *Pharmacoepidemiol Drug Saf*, 24(5):478–85.

Roberts, A. W., Dusetzina, S. B., and Farley, J. F. (2015). Revisiting the washout period in the incident user study design: why 6-12 months may not be sufficient. *J Comp Eff Res*, 4(1):27–35.

Ross, T. R., Ng, D., Brown, J. S., Pardee, R., and Hornbrook, M. C. (2014). The hmo research network virtual data warehouse: A public data model to support collaboration. *eGEMs*, 2(1).

Ryan, P., Suchard, M. A., Schuemie, M. J., and Madigan, D. (2013). Learning from epidemiology: Interpreting observational database studies for the effects of medical products. *Statistics in Biopharmaceutical Research*, pages null–null.

Ryan, P., Welebob, E., Hartzema, A. G., Stang, P., and Overhage, J. M. (2010). Surveying us observational data sources and characteristics for drug safety needs. *Pharmaceutical Medicine*, 24(4):231–238.

StataCorp (2013). Stata statistical software: Release 13.

Turner, J. A., Saunders, K., Shortreed, S. M., LeResche, L., Riddell, K., Rapp, S. E., and Von Korff, M. (2014). Chronic opioid therapy urine drug testing in primary care: prevalence and predictors of aberrant results. *J Gen Intern Med*, 29(12):1663–71.

Voordouw, A. C., Sturkenboom, M. C., Dieleman, J. P., Stijnen, T., Smith, D. J., van der Lei, J., and Stricker, B. H. (2004). Annual revaccination against influenza and mortality risk in community-dwelling elderly persons. *JAMA*, 292(17):2089–95.

Appendix

In the main manuscript we report results for unadjusted and age- and sex-adjusted hazard ratios in the full Group Health populations regardless of how long individuals were enrolled prior to January 1, 2008. In order to ascertain demographic and general health history information individuals in the shortest enrollment category with less than 1 year of Group Health enrollment were excluded from the Cox model used to estimate hazard ratios for the demographic and general health history adjusted model reported in the main manuscript. This appendix reports the results of sensitivity analyses (Table 3) excluding individuals with less than 1 year of Group Health enrollment from the shortest enrollment category for all models (including unadjusted and age- and sex- adjusted models). The results reporting the hazard ratios from the demographic and general health history adjusted model are nearly identical to those reported in Table 2 of the main manuscript.

Table 3: Comparison of cancer and all-cause mortality hazard ratios by prior enrollment duration.

Length of prior enrollment	Breast cancer	Cervical cancer	Colorectal cancer	Lung cancer	Prostate cancer	All-cause mortality
			Number of events			
$\geq$ 1 & <5 years†	261	72	92	118	130	922
$\geq$ 5 & <10 years	258	56	102	133	143	1144
$\geq$ 10 years	911	241	472	540	671	6958
			Unadjusted hazard ratios [95% CI]			
$\geq$ 1 & <5 years†	Ref	Ref	Ref	Ref	Ref	Ref
$\geq$ 5 & <10 years	1.16	0.92	1.32	1.35	1.32	1.47
	[0.98, 1.38]	[0.65, 1.30]	[1.00, 1.76]	[1.05, 1.73]	[1.04, 1.68]	[1.35, 1.60]
$\geq$ 10 years	1.29	1.23	1.96	1.74	2.02	2.83
	[1.12, 1.48]	[0.95, 1.60]	[1.56, 2.44]	[1.42, 2.12]	[1.67, 2.43]	[2.63, 3.03]
			Age- and sex-adjusted hazard ratios [95% CI]			
$\geq$ 1 & <5 years†	Ref	Ref	Ref	Ref	Ref	Ref
$\geq$ 5 & <10 years	1.10	0.88	1.08	1.04	1.12	0.97
	[0.92, 1.31]	[0.62, 1.25]	[0.81, 1.43]	[0.81, 1.33]	[0.88, 1.42]	[0.89, 1.06]
$\geq$ 10 years	1.11	1.06	1.15	0.90	1.20	0.99
	[0.96, 1.28]	[0.81, 1.39]	[0.92, 1.46]	[0.74, 1.11]	[0.99, 1.45]	[0.92, 1.06]
			Demographic and general health history adjusted hazard ratios [95% CI]*			
$\geq$ 1 & <5 years†	Ref	Ref	Ref	Ref	Ref	Ref
$\geq$ 5 & <10 years	1.05	0.91	1.08	1.05	1.07	0.99
	[0.89, 1.26]	[0.64, 1.30]	[0.81, 1.44]	[0.82, 1.35]	[0.84, 1.36]	[0.90, 1.08]
$\geq$ 10 years	1.05	1.08	1.14	0.98	1.09	0.99
	[0.90, 1.21]	[0.82, 1.43]	[0.90, 1.45]	[0.80, 1.21]	[0.90, 1.32]	[0.92, 1.07]

*Demographic and general health history model included as covariates: age, sex, race, Charlson score, tobacco use and body mass index.

†Reference group includes individuals who had at least 1 year prior enrollment and less than 5 years of enrollment.