OPEN

Consequences of Depletion of Susceptibles for Hazard Ratio Estimators Based on Propensity Scores

Bruce Fireman,[a] *Susan Gruber,*[b,c] *Zilu Zhang,*[b] *Robert Wellman,*[d] *Jennifer Clark Nelson,*[d] *Jessica Franklin,*[e] *Judith Maro,*[b] *Catherine Rogers Murray,*[b] *Sengwee Toh,*[b] *Joshua Gagne,*[e] *Sebastian Schneeweiss,*[e] *Laura Amsden,*[a] *and Richard Wyss*[e]

Abstract: We use simulated data to examine the consequences of depletion of susceptibles for hazard ratio (HR) estimators based on a propensity score (PS). First, we show that the depletion of susceptibles attenuates marginal HRs toward the null by amounts that increase with the incidence of the outcome, the variance of susceptibility, and the impact of susceptibility on the outcome. If susceptibility is binary then the Bross bias multiplier, originally intended to quantify bias in a risk ratio from a binary confounder, also quantifies the ratio of the instantaneous marginal HR to the conditional HR as susceptibles are depleted differentially. Second, we show how HR estimates that are conditioned on a PS tend to be between the true conditional and marginal HRs, closer to the conditional HR if treatment status is strongly associated with susceptibility and closer to the marginal HR if treatment status is weakly associated with susceptibility. We show that associations of susceptibility with the PS matter to the marginal HR in the treated (ATT) though not to the marginal HR in the entire cohort (ATE). Third, we show how the PS can be updated periodically to reduce depletion-of-susceptibles bias in conditional estimators. Although marginal estimators can hit their ATE or ATT targets consistently without updating the PS, we show how their targets themselves can be misleading as they are attenuated toward the null. Finally, we discuss implications for the interpretation of HRs and their relevance to underlying scientific and clinical questions. See video Abstract: http://links.lww.com/EDE/B727.

Keywords: Propensity score; Hazard ratio; Survival analysis; Depletion of susceptibles; Survivor bias; Noncollapsibility

(*Epidemiology* 2020;31: 806–814)

Susceptibles are defined as individuals whose baseline risk is relatively high. If the outcome can only happen to a person once—like death—then the prevalence of susceptibility in a cohort decreases over time because susceptibles tend to have earlier outcome events. Even in a randomized controlled trial (RCT), if the treatment reduces risk then the treated and untreated groups lose their susceptibles at different rates: the prevalence of susceptibility decreases faster in the untreated group than in the treated group. Consequently, the risk profile of the untreated survivors becomes more favorable than that of the treated survivors. This process, termed "differential depletion of susceptibles," poses challenges for the estimation and interpretation of hazard ratios (HRs) in survival analyses. Unless we ascertain and adjust for all aspects of susceptibility, the HR attenuates toward the null. This is a known source of selection bias, yet it is often overlooked.[1–5]

We use the acronym HRc to denote an HR that is conditional on susceptibility and HRm to denote an HR that is marginal (i.e., population-averaged over time). In an RCT—or a cohort study balanced by propensity score (PS) matching or weighting—the HRm equals the HRc initially, then the HRm diverges toward the null (unless it is already at the null) as outcomes occur. Our aims are to:

(1) show how the divergence of the HRm from the HRc is driven by the incidence of the outcome, the distribution of susceptibility, and the effect of susceptibility on the outcome.
(2) show how HRc estimators that condition on a PS tend to yield estimates between the HRc and HRm, closer to the HRc if the PS is strongly associated with the outcome (within each treatment group) and closer to the HRm if the PS is weakly associated with the outcome.

Submitted February 25, 2020; accepted July 27, 2020.

From the [a]Kaiser Permanente Division of Research, Oakland, CA; [b]Department of Population Medicine, Harvard Pilgrim Health Care Institute and Harvard Medical School, Boston, MA; [c]Putnam Data Sciences, LLC, Cambridge, MA; [d]Kaiser Permanente Washington Health Research Institute, Seattle, WA; and [e]Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA.

This project was funded by the Food and Drug Administration (FDA), HHSF223200910006I.

The authors report no conflicts of interest.

Disclosure: Contact Bruce Fireman at Bruce.Fireman@kp.org.

SDC Supplemental digital content is available through direct URL citations in the HTML and PDF versions of this article (www.epidem.com).

Correspondence: Bruce Fireman, Division of Research, KPNC, 2000 Broadway, Oakland, CA 94612. E-mail: bruce.fireman@kp.org.

ISSN: 1044-3983/20/3106-0806
DOI: 10.1097/EDE.0000000000001246

(3) show how a PS can be updated periodically so that PS-based methods can estimate the HRc without the selection bias that arises from depletion of susceptibles.

This work was undertaken to evaluate PS-based methods for the Sentinel Initiative, which monitors drug safety with data on over 200 million people.[6,7] Sentinel is interested in PS-based methods because they can preserve privacy by letting individual-level data remain at Sentinel's partner organizations.[8] A Sentinel workgroup evaluated PS-based methods that use matching, stratification, PS-based covariates, or inverse probability of treatment weighting (IPTW) in a wide range of simulated scenarios,[9] and found bias in PS-based estimators of the HRc, as was previously reported.[10–13] A subsequent Sentinel workgroup proposed a time-dependent PS to reduce this bias.[14] Here, we provide new insight into the bias, add refinements to the time-dependent PS, and point out features of research questions that can make the HRc—rather than the HRm or alternative effect measure—an appropriate target.

METHODS

Conceptual Framework

Table 1 shows attenuation of the HRm in a hypothetical RCT of a treatment that always reduces mortality by 50% (HRc = 0.5). One million subjects are randomized 1:1 to be treated or not. Half are susceptible (high risk), which multiplies risk ten-fold. Everyone is followed until death. Each row of the table summarizes an interval when 1% of subjects die. The first row shows the 10,000 earliest deaths; the last row shows the last 10,000. Columns 3–6 show the interval's deaths (and survivors) in four subgroups: high-risk treated, low-risk treated, high-risk untreated, and low-risk untreated. Given that 50% of subjects are treated, 50% are susceptible, treatment reduces mortality by 50%, and susceptibility multiplies mortality ten-fold, we have enough information to identify each subgroup's share of each row's 10,000 deaths. (All 100 rows are shown in eTable 1; http://links.lww.com/EDE/B712.)

Column 10 shows the instantaneous HRm (iHRm) in each interval: mortality among the treated divided by mortality among the untreated. Thus, the iHRm in interval 5 is:

$$[(3{,}112+326)/(237{,}758+248{,}755)]/[(5{,}915+648)/(225{,}972+247{,}515)] = 0.510.$$

The iHRm diverges from the HRc toward the null as susceptibles are depleted. It reaches a maximum of 0.95 when 57% of the cohort has died (row 57); then gradually returns to 0.50. For intuition about why the iHRm stops moving away from the HRc and starts returning to it, notice that after row 57 the RCT's untreated arm is so depleted of its susceptibles that mortality no longer exacerbates the imbalance in susceptibility but instead restores balance.

More insight into the trajectory of the iHRm comes from the bias multiplier "*b*" in column 9. The iHRm equals the HRc (always 0.5 in this illustration) multiplied by "*b*," which is calculated from the prevalence of susceptibility in the treated (column 7) and the untreated (column 8). These prevalences diverge as susceptibles are depleted differentially. Denote these prevalences p_1 and p_0 and let s denote the risk ratio for the effect of susceptibility on mortality; then this bias multiplier is:

$$b = [p_1(s-1)+1]/[p_0(s-1)+1]$$

Bross derived this bias multiplier over 50 years ago to quantify the bias in a risk ratio estimate attributable to an uncontrolled binary confounder.[15] Table 1 shows how Bross's "*b*" also quantifies the ratio of the instantaneous HRm to the HRc as susceptibles are depleted differentially. The instantaneous HRm can be calculated in two equivalent ways: using the numbers in columns 3–6, as shown for interval 5 above, or else by multiplying the HRc by Bross's "*b*."

This illustrates a similarity between the divergence of the iHRm from the HRc (caused by depletion of susceptibles) and the divergence of a biased risk ratio estimate from the true risk ratio (caused by a confounder). The differential depletion of susceptibles induces a selection bias that moves the HRm away from the HRc similarly to how confounders bias a risk ratio estimate. (See Smith and VanderWeele for analyses of selection bias that take the same analytic form as sensitivity analyses of confounding.[16])

The overall HRm is in the last column. For row *t*, it summarizes the instantaneous HRm's from the start through row *t*. Anchored to its history in this way, the overall HRm changes more gradually than does the instantaneous HRm. It reaches a maximum of 0.68 in row 69 and then gradually moves back to 0.63.

The overall HRm moves far from the HRc in this illustration because susceptibility has a large impact on risk—it multiplies risk ten-fold. The HRm and HRc would diverge less if susceptibility had less impact. This is shown in Figure 1 where we plot the divergence of the HRm and HRc while varying the prevalence of susceptibility and its effect on outcomes. The gap between the HRc and HRm widens with the cumulative incidence of the outcome until one of the treatment groups is so depleted of susceptibles that subsequent depletion no longer increases imbalance. Although the iHRm can cross the null as it diverges from the HRc (so that the iHRm might suggest the treatment is harmful at a time when the HRc indicates it is beneficial), the overall HRm does not cross the null as long as the HRc is constant.

Figure 2 plots the divergence of HRm and HRc in scenarios where susceptibility is continuous rather than dichotomous. The gap widens with increases in susceptibility's variance and therefore its impact on outcomes. Unlike

TABLE 1. Divergence of the Marginal Hazard Ratio (HRm) From the Conditional Hazard Ratio (HRc) as Susceptibles Are Depleted

Column 1	Column 2	Column 3	Column 4	Column 5	Column 6	Column 7	Column 8	Column 9	Column 10	Column 11
Time *t*	Total N Alive at End of *t*	Deaths (N) Treated High Risk	Deaths (N) Treated Low Risk	Deaths (N) Untreated High Risk	Deaths (N) Untreated Low Risk	Prev. of Hi-Risk in Treated[a]	Prev. of Hi-Risk in Untreated	Bross Bias Multiplier[b]	Instantaneous HRm at *t*	Overall HRm, t_0 Through *t*
0	1,000,000	(250,000)	(250,000)	(250,000)	(250,000)					
1	990,000	3,030 (246,970)	303 (249,697)	6,061 (243,939)	606 (249,394)	0.500	0.500	1.000	0.500	0.500
2	980,000	3,050 (243,920)	308 (249,389)	6,025 (237,914)	616 (248,778)	0.497	0.494	1.005	0.502	0.501
3	970,000	3,070 (240,849)	314 (249,075)	5,989 (231,924)	626 (248,152)	0.494	0.489	1.009	0.505	0.502
4	960,000	3,091 (237,758)	320 (248,755)	5,953 (225,972)	637 (247,515)	0.492	0.483	1.014	0.507	0.504
5	950,000	3,112 (234,647)	326 (248,429)	5,915 (220,057)	648 (246,867)	0.489	0.477	1.019	0.510	0.505
6–55	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
56	440,000	3,917 (44,137)	1,733 (210,888)	1,403 (7,205)	2,946 (177,769)	0.184	0.045	1.887	0.943	0.648
57	430,000	3,832 (40,306)	1,831 (209,057)	1,251 (5,954)	3,087 (174,683)	0.173	0.039	1.894	0.947	0.652
58	420,000	3,730 (36,576)	1,935 (207,123)	1,102 (4,852)	3,233 (171,450)	0.162	0.033	1.893	0.947	0.656
59	410,000	3,611 (32,964)	2,045 (205,078)	958 (3,894)	3,386 (168,064)	0.150	0.028	1.884	0.942	0.660
60–67	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
68	320,000	1,841 (8,556)	3,259 (180,821)	112 (204)	4,788 (130,418)	0.053	0.002	1.451	0.725	0.682
69	310,000	1,610 (6,945)	3,403 (177,418)	77 (127)	4,909 (125,509)	0.045	0.002	1.387	0.694	0.682
70	300,000	1,388 (5,558)	3,545 (173,873)	51 (77)	5,016 (120,493)	0.038	0.001	1.327	0.663	0.682
71–99	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
100	0	0 (0)	9,955 (0)	0 (0)	45 (0)	0.000	0.000	1.000	0.500	0.634

Lifetable of a hypothetical trial of a treatment that reduces mortality in a cohort with high and low susceptibility.
Time scaled as cumulative incidence: 100 time periods bounded by dates marking percentiles of death times.
HRc = 0.50 (treatment cuts risk 50%), susceptibility (high risk) multiplies mortality by 10, cohort is 50% high risk.
[a]Columns 7 and 8 show prevalences of susceptibility at interval's start; columns 3–6 show the *N* of survivors (parenthesized) at interval's end.
[b]Bross's formula for sensitivity of a risk ratio estimate to a confounder is shown here to yield the ratio of the iHRm (column 10) to HRc (0.5) at *t*.

Figure 1, these curves are not peaked; the gap never narrows—the more hazardous treatment group is always depleted faster of more susceptible survivors.

In these scenarios, divergence of the HRm and HRc is slight until outcomes occur in 1%–2% of the cohort. However, if an outcome only occurs in susceptibles and susceptibility is rare, then the HRm can diverge far from the HRc when incidence is high among the susceptibles regardless of how low it is in the overall cohort (eTable 2; http://links.lww.com/EDE/B712).

If we strengthen the treatment effect by moving the HRc farther from the null, the HRm stays at the same percentage of the distance from the HRc to the null (on the log scale). For example, consider a scenario where the HRm is halfway between the HRc and the null on the log scale: if we double the log HRc we approximately double the log HRm, widening the gap between the HRm and HRc, yet keeping the log HRm halfway between the log HRc and the null (eFigure 1; http://links.lww.com/EDE/B712).

Finally, if the treatment has no effect on outcomes, then depletion of susceptibles is not differential, imbalances do not arise, and the HRm and HRc do not diverge.

Simulations and Analyses

We conducted plasmode and Monte Carlo simulations.[17]

Plasmode simulations

We used de-identified Sentinel data on 39,472 new users of an anticoagulant, either rivaroxaban or warfarin.[18] We sampled with replacement to make simulated new-user cohorts with realistic covariate distributions and covariances. Treatments and outcomes were not sampled; they were allocated by mechanisms tailored to control the strength of the treatment effect, strength of confounding, outcome incidence, treatment prevalence, and amount of censoring. The scenario featured in Figure 3 included strong negative confounding such that an unadjusted analysis yields an HRc estimate of 1.0 when the truth is 2.0.

We used 15 covariates: age, sex, and 13 binary covariates with the most confounding potential, by Bross's formula.[15] We allocated treatment by a logistic function of the covariates. Time-to-event was assigned by a Weibull function of the covariates. We censored follow-up randomly (unrelated to covariates) at a rate inspired by the Sentinel data.

For each scenario we generated 1,000 datasets using SAS 9.3.[19] In each dataset, we estimated the PS using a logistic regression model consistent with the treatment-generating mechanism. We estimated the HRc and HRm by Cox regression using the PS in one or another of the ways under

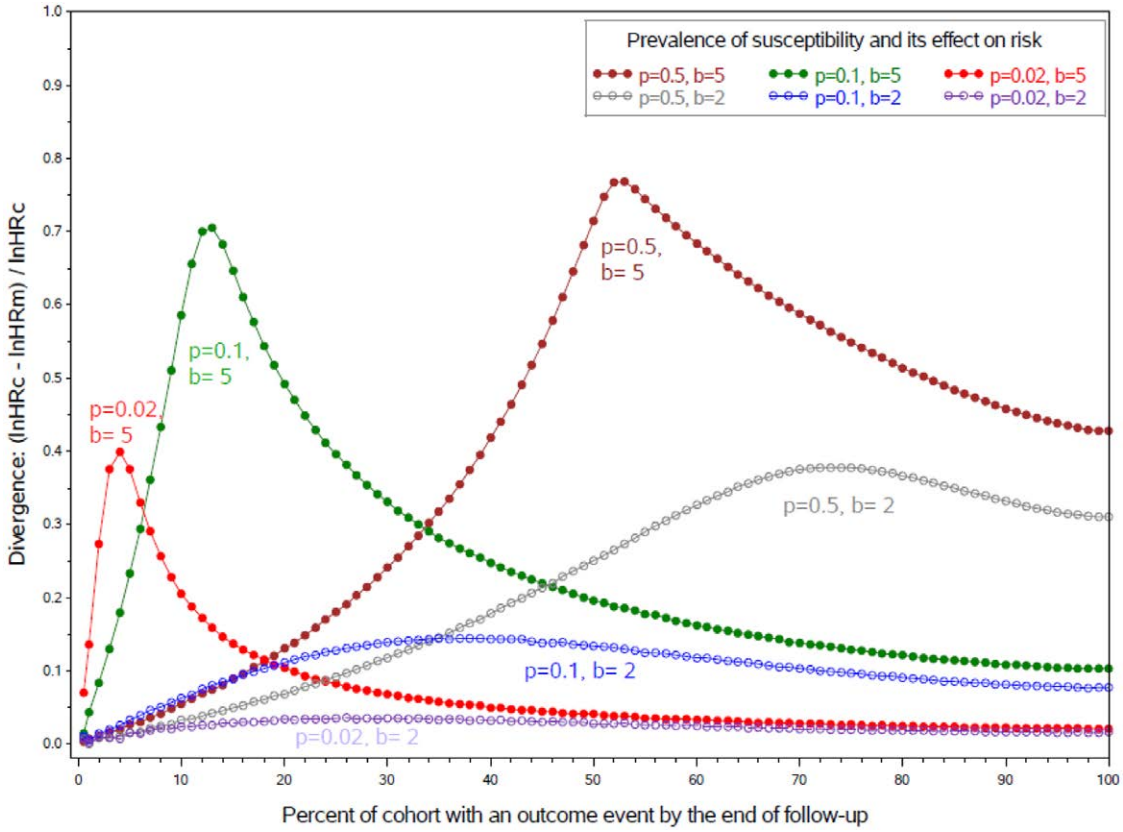

FIGURE 1. Divergence of the marginal HR (HRm) from a constant conditional HR (HRc) if susceptibility is binary by cumulative incidence, the prevalence (*p*) of susceptibility (*s*) and the effect (b) of susceptibility on risk, in a hypothetical randomized controlled trial where treatment doubles risk and time-to-event is proportional to exp(lnHRc × *tx* + *b* × *s*). These curves, shown here for HRc = 2, would be similar at any other non-null level of the HRc.

consideration. The "true" HRc was explicit in the data-generating mechanism; we found the corresponding "true" HRm by Cox regression in counterfactual cohorts followed for 2 years without censoring. The HRm's are standardized to either the entire cohort (ATE) or the treated group (ATT).

The estimators in the top nine rows of Figure 3 target the HRc. The first estimator provides a benchmark by adjusting for the individual covariates without a PS; the next eight estimators condition on the PS. They adjust for PS-based covariates as polynomial terms (row 2), dummy variables for PS deciles (row 3), or cubic B-splines with knots at quintiles of the PS among the treated (row 4). The next estimators use "greedy" nearest neighbor matching, either 1:1 (row 5) or 1:M with up to 10 comparators per treated subject (row 6). The estimators in rows 7–9 stratify on the PS using 10 strata (row 7), 20 strata (row 8), or fine stratification (5 subjects per stratum) (row 9). The estimators in rows 5–9 condition the outcome model so that each risk set is restricted to individuals in the same matched set or stratum.

The estimators in the bottom three rows of Figure 3 target an HRm, either ATE or ATT. Matching 1:1 is used for row 10; 1:M variable ratio matching is used for row 11. These matched HRm estimators differ from the matched HRc estimators in that they do not condition the outcome model on matched set; instead they fit unconditional Cox models. The IPTW estimators in row 12 use stabilized weights with robust variance estimation.[20]

Monte Carlo simulations

To clarify the divergence of the HRm and HRc (aim 1), we varied features of Table 1's hypothetical RCT. For Figure 1, the prevalence of susceptibility varied from 2% to 10% to 50%, and susceptibility's effect on mortality varied from an HRc of 2 to 5. For Figure 2, susceptibility was normally distributed, mean = 0, and variance either 0.5, 1.0, 1.5, 2, 3, or 4. Event times were exponential:

$$\text{time-to-event} = -\ln(u) / -\{\exp[suspectibility + (\ln HR_c \times Tx)]\},$$

where Tx is treatment status, u is random uniform.

To elucidate the bias in PS-based estimators of the HRc (aim 2), we varied the correlation of the PS with susceptibility, defined as a continuous risk score like Hansen's prognostic score.[21] Our risk score was the sum of 40 normally distributed covariates. Time-to-event was proportional to exp($riskscore$ + ($lnHR_c \times Tx$)). Treatment was based on the logit of: (the sum

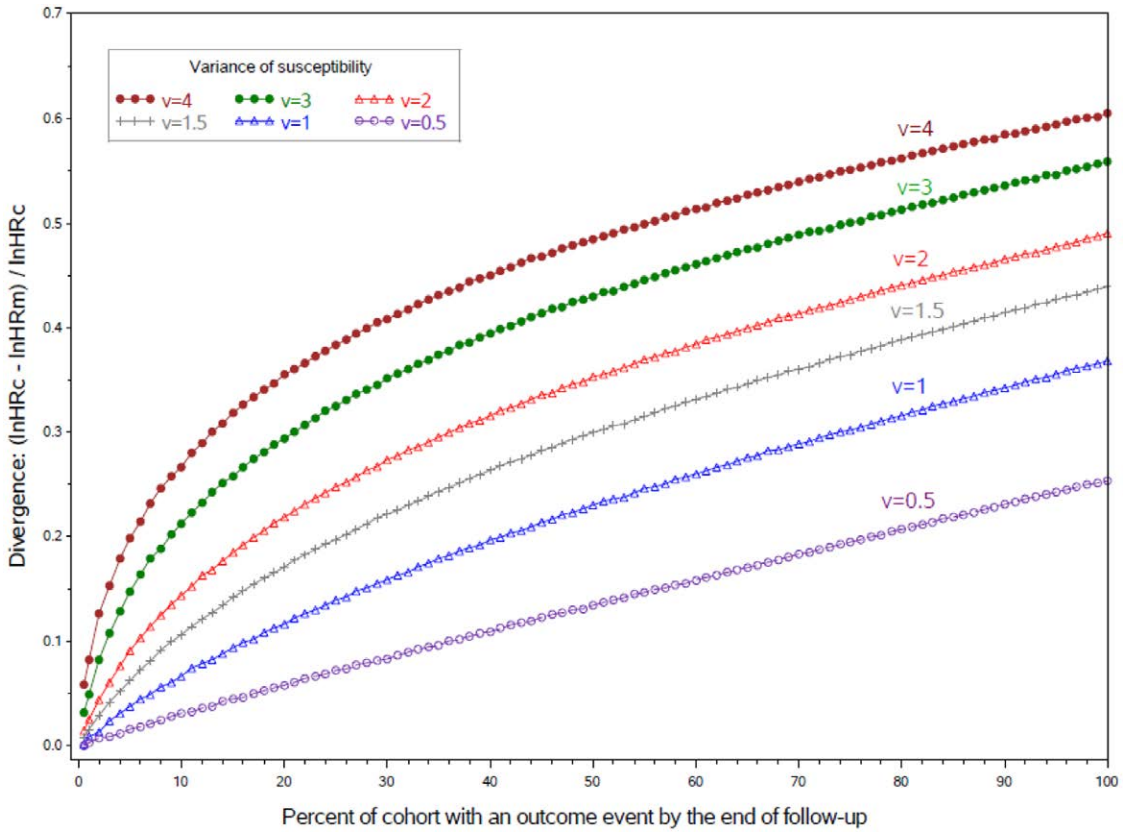

FIGURE 2. Divergence of the marginal HR (HRm) from a constant conditional HR (HRc) if susceptibility is continuous, by cumulative incidence and the variance of susceptibility (s), where s is normal with mean = 0, variance = v, in a hypothetical RCT where treatment doubles risk and time-to-event is proportional to exp(lnHRc × tx + sqrt(v) × s). These curves, shown here for HRc = 2, would be similar at any other non-null level of the HRc.

of k covariates) – (the sum of the other 40-k covariates). The correlation of the PS with the risk score was 0.75 if $k = 36$, 0.0 if $k = 20$, and -0.75 if $k = 4$.

To address our third aim, we developed a time-dependent PS following Wyss[14] in steps 1–4 below, adding refinements to the PS model and outcome model in steps 5–6 (details in eAppendix; http://links.lww.com/EDE/B712):

1. Make a baseline PS from baseline covariates that balances the treatment groups initially.
2. Assess whether the initial balance is sustained over time. If not, then…
3. Chop the timeline into intervals in which the PS can be updated. We made deciles each with 10% of the observed events.
4. Estimate the PS at the midpoint of each interval. Do NOT update the baseline covariates; update the function of them that predicts treatment status in individuals-still-at-risk.
5. The model for the updated PS may balance the treatment groups better if interactions are included even if none were needed for the baseline PS, because treatment status becomes related to the baseline covariates through the outcome mechanism as well as the initial treatment mechanism.
6. Specify and fit a Cox model for the outcome that conditions the HRc estimate on the updated PS and interactions—where interactions are cross products of earlier and later PS's.

We evaluated the time-varying PS in simulated cohorts, each with N = 100,000, HRc = 2.0, 40 normal covariates (mean = 0, SD = 0.2), and a correlation between PS and risk score of 0.0 or 0.75.

RESULTS

HR estimates from the plasmode simulations are compared in Figure 3 with the true HRc and both HRm's (ATE and ATT). The benchmark estimator (first row) was expected to be unbiased because it adjusts for covariates individually (without a PS) in a model consistent with the data-generating mechanism. As expected, it was within 1% of the true HRc.

The PS-based estimators of the HRc (rows 2–9 of Figure 3) all yielded estimates that were biased away from the true HRc = 2.0 toward the null. The estimates obtained by 1:1 matching averaged 5% below the HRc; the other conditional PS-based methods landed 10%–16% below the HRc. Although all PS-based conditional estimators were attenuated

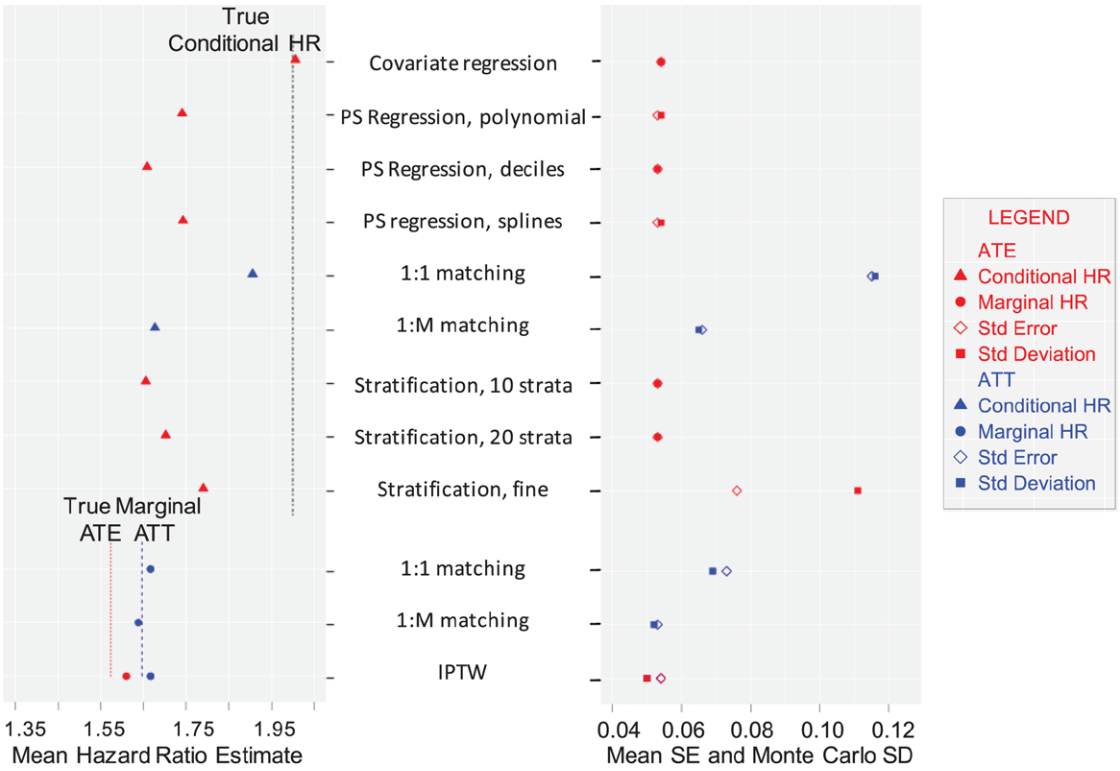

FIGURE 3. Mean hazard ratios (HRs) in the entire cohort (ATE) or the treated (ATT), and mean standard errors (SE) and Monte Carlo standard deviations (SD), by method.

toward the null, they were not attenuated as much as was the HRm (ATE).

The marginal ATE and ATT targets were 1.57 and 1.65, respectively, well below the HRc (2.0). There was little bias in the marginal estimators—each landed a little above its ATE or ATT target by amounts ranging from 1% to 3%. Censoring—even though it was "uninformative"—biased the HRm estimates slightly away from their target toward the HRc. As shown in Table 1, an overall HRm at t amounts to a summary measure of all instantaneous HRm's from the start of follow-up through t. Censoring reduces the contribution to this overall HRm of the later more-attenuated iHRm's relative to the contribution of earlier less-attenuated iHRm's (because more of the later outcomes are unobserved), unless each risk set is weighted by the probability that its anchoring event is censored. HRc estimators are similarly biased by censoring if the HRc is not constant.

Figure 3 also shows the precision of the estimators. Estimators that used 1:1 matching (rows 5 and 10) were less precise. They lost precision because some of the cohort was left unmatched and could not be informative. The conditional matched 1:1 estimator (row 5) was the least precise because whenever follow-up ended for one member of a matched pair, the other member could no longer be informative. For similar reasons, the estimator using fine stratification (row 9) was less precise than estimators using coarser stratification (rows 7–8).

Among PS-based estimators that were similarly precise (rows 2–4, 7–8, 11–12), the HRc estimators yielded slightly more powerful tests than the HRm estimators because the HRc estimates were a little farther from the null. The null was tested with the Wald statistic, which is a ratio of a log HR estimate to its standard error; when we compare Wald statistics with similar standard errors in their denominators, the Wald statistic whose numerator is farther from the null has more power.

In Monte Carlo simulations, as in plasmode simulations, the PS-based HRc estimates, shown in blue in Figure 4, were between the true HRc and the true HRm (ATE), shown in black. The PS-based estimates (traced by the blue curves) were closer to the HRc in scenarios where the PS was highly correlated with the risk score ($r = 0.84$ or 0.89) and closer to the HRm in scenarios where the PS was less correlated with the risk score ($r = 0.47$). In the scenario where the correlation of the PS with the risk score was 0, the blue curve tracing the PS-based HRc estimates follows the same trajectory as the black curve tracing the HRm (ATE).

The trajectory of the HRm (ATE) is the same in Figure 4 scenarios regardless of the PS-riskscore correlation, but the ATT varies with the correlation between the PS and risk score. We show the ATT for two of the scenarios (the red curves) to illustrate that the ATT is lower when the PS and risk score are positively correlated, and higher when they are negatively correlated. If most high-risk conditions are predictive of being

treated then high-risk individuals are upweighted in the ATT (though not the ATE), increasing imbalances from depletion-of-susceptibles and moving the ATT farther from the HRc.

The HR estimator that uses a time-varying PS is compared in Table 2 with estimators that use the individual covariates, IPTW, the baseline PS, or the risk score. When the HRc and HRm diverged, the covariate-adjusted HRc estimator stayed on its HRc target, and the IPTW-adjusted HRm estimator stayed on its HRm target, as expected. The baseline-PS-adjusted estimates approximated the HRm when the baseline PS was uncorrelated with the risk score and were between the HRc and HRm when the PS-risk score correlation was 0.75. When cumulative incidence was 95% and the PS-risk score correlation was 0 (bottom row), the baseline-PS estimator was most biased—43% below the true HRc on the log scale.

The time-varying PS yielded estimates that were near their target, the HRc = 2.0. When the bias in the baseline-PS estimator was 43% (bottom row) the bias was only 2% in the estimator using the time-varying PS. Whereas estimates using the time-varying PS were only slightly below their 2.0 target, the risk-score-adjusted estimates were even less biased—except when there were not enough events for a precise risk score (top rows of top panel). See the eAppendix; http://links.lww.com/EDE/B712 for more on adjustment for the time-varying PS, the risk score, or both.

DISCUSSION

We examined the consequences of depletion of susceptibles for HR estimators based on a PS. We found that PS-based estimates of the HRc are biased in an interesting way: they tend to be between the HRc and the HRm, closer to the HRc if the baseline PS is highly correlated with susceptibility and closer to the HRm if the PS is weakly correlated with susceptibility. When outcomes are infrequent or unaffected by the treatment, the HRm and HRc are nearly equal and this bias is negligible. However, when susceptibles are depleted differentially, a gap opens between the HRc and HRm, and PS-based HRc estimates fall into the gap. We described how this bias arises and how it can be reduced by updating the PS.

We found little bias in PS-based estimators that target a HRm. When susceptibles are depleted differentially, both kinds of HRm, ATE and ATT, diverge from the HRc and their PS-based estimators move with them (in the absence of censoring that tends to bias HR estimators toward their earlier levels). Our findings are consistent with Austin who found substantial bias in PS-based estimators of the HRc and negligible bias (in the absence of censoring) in PS-based estimators of the HRm.[10] Even though HRm estimators can hit their targets consistently, their interpretation should consider the contribution from differential depletion-of-susceptibles.

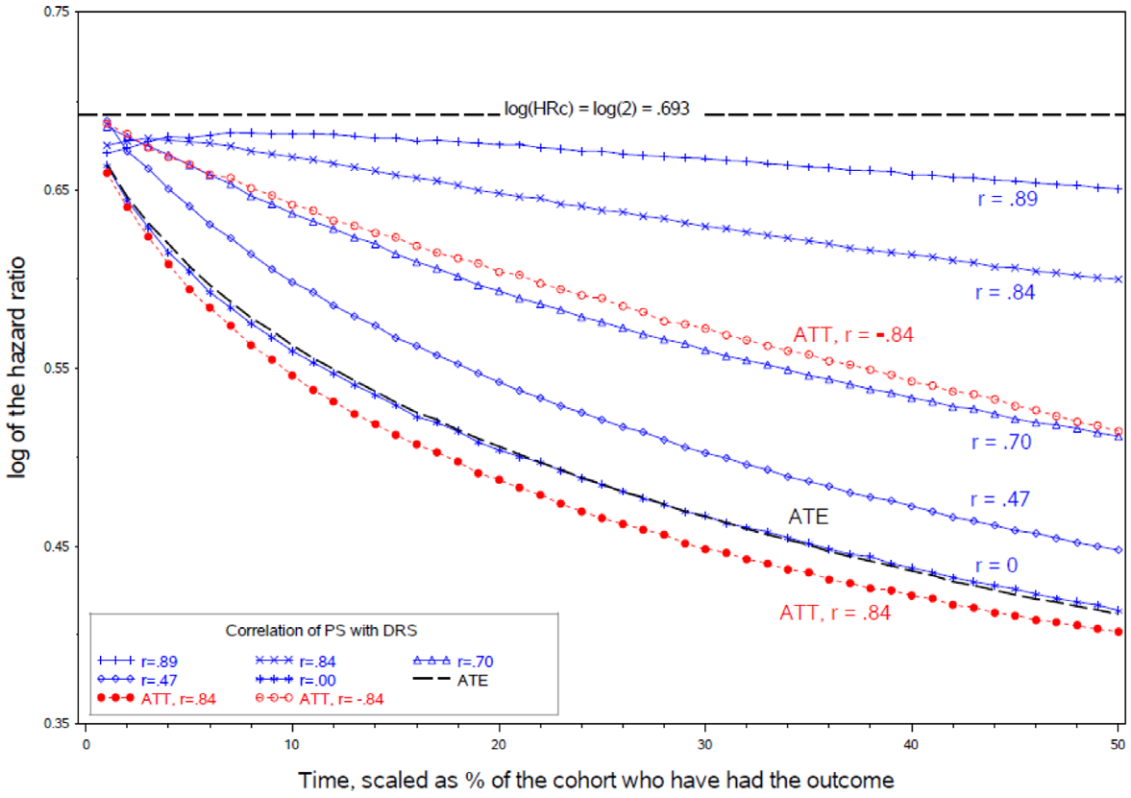

FIGURE 4. Conditional log hazard ratio (HRc) estimates over time by the PS-risk score correlation, in relation to the target log HRc and the corresponding log HRm.

Table 2. Conditional Hazard Ratio (HRc) Estimates Adjusted by an Updated Propensity Score (PS) Compared With HRc or HRm Adjusted by (a) Individual Covariates, (b) Risk Score, (c) IPTW, or (d) Baseline PS, by Cumulative Incidence

Cumulative Incidence at Endpoint, % of Cohort	(a) HRc by Covariates	(b) HRc by Risk Score	True HRm	(c) HRm by IPTW	(d) HRc by Baseline PS	(e) HRc by Updated PS
Correlation of PS with risk score is 0.75						
1	2.00	2.26	1.98	1.99	1.94	2.00
5	2.00	2.05	1.91	1.92	1.93	1.99
15	2.00	2.01	1.80	1.81	1.90	1.99
25	2.00	2.00	1.73	1.73	1.86	1.98
50	2.00	2.00	1.61	1.61	1.78	1.98
75	2.00	2.00	1.53	1.53	1.71	1.98
95	2.00	2.00	1.48	1.48	1.68	1.96
Correlation of PS with risk score is 0.00						
1	2.00	2.00	1.98	1.98	1.98	1.98
5	2.00	2.00	1.91	1.91	1.91	1.98
15	2.00	2.00	1.80	1.80	1.80	1.98
25	2.00	2.00	1.73	1.73	1.73	1.97
50	2.00	2.00	1.61	1.61	1.61	1.97
75	2.00	2.00	1.53	1.53	1.53	1.97
95	2.00	2.00	1.48	1.48	1.48	1.97

The true HRc = 2, each simulated cohort has $N = 10^5$, and the Monte Carlo 95% confidence interval for each mean HR estimate is <0.007 in width.

Given that the HRc and HRm can diverge, when is it appropriate to target the HRc and use a time-varying PS to estimate it?

An HR can be an appropriate target if it is plausible that the treatment effect would be multiplicative among the individuals-still-at-risk; and the HRc can be more relevant than the HRm to the underlying scientific and clinical issues if well-measured covariates affect risk. Imagine an RCT that randomly assigns people in September to be vaccinated or not against influenza and examines vaccine effectiveness (VE = (1 – HR) X 100%) each month of a December through March flu season. Imagine that the true HRc is stable at 0.5 yet influenza incidence is high, susceptibles are depleted differentially, and by March so many unvaccinated susceptibles have been infected (and are immune to reinfection this season) that Bross's bias multiplier is 1.89, as in row 57 of Table 1. If our research question is whether vaccine protection wanes (so that vaccination should be delayed until November for more timely protection), then the relevant target is the trajectory of the HRc—the trajectory of the HRm could be misleading insofar as it reflects depletion of susceptibles. As the HRm attenuates from 0.5 to 0.95, the corresponding VE estimate decreases from 50% to 5% due to selection bias. In this RCT, we should target the HRc, condition our analysis on the aspects of susceptibility that are measured (such as age and preexisting conditions), and assess the sensitivity of our findings to aspects of susceptibility that remain unmeasured.

An RCT for a SARS-CoV-2 vaccine can examine waning similarly. Waning can be assessed from the trajectory of the HRc adjusted for measured susceptibility. Unmeasured susceptibility may remain, yet we would reduce bias from depletion-of-susceptibles by targeting the HRc rather than the HRm. (See Ray on depletion-of-susceptibles bias in research on the waning of vaccine protection.[22])

Although our estimators use Cox regression, we need not assume that the HRm or HRc was constant—indeed their divergence implies that one or both changed. To examine a trend in an HRm or HRc, we can specify time-by-treatment interaction effects or divide the timeline into intervals and estimate the HR in each interval. If there is interest in the average effect during follow-up, the overall HR estimate can be interpreted as averaging the interval-specific HR's, as in Table 1.

Hernán's thoughtful article "The hazards of hazard ratios" considers two problems with HRs: first, they tend to be moving targets and second, they are prone to selection bias from depletion of susceptibles.[1] The depletion-of-susceptibles problem is our main concern in this article; the moving target problem is challenging when events are too sparse to ascertain the trajectory of the HRc, but it is not necessarily a reason to de-emphasize the HRc in favor of another effect measure, such as a risk difference or restricted mean survival time. When the treatment effect is expected to be multiplicative, we can target the HRc. If it may be strengthening or weakening, we can target its trajectory. Whereas Hernán duly emphasizes the value of marginal survival curves, the HRc and its trajectory answer some research questions more directly. It can be helpful to ascertain the treatment's effect on the individual survivors (as measured by the HRc) apart from

its effect on the mix of individuals who survive (which attenuates the HRm toward the null).

When could the time-varying PS be useful? In our simulations, adjustment for a time-varying PS reduced bias from depletion-of-susceptibles, but it is more burdensome than adjustment for individual covariates and less intuitive than a risk score. However, in a distributed data environment—such as the Sentinel Initiative—privacy concerns may preclude pooling individual covariates, and a risk score is not feasible unless outcome events are frequent at every site. If outcome incidence is high enough to move the HRm away from the HRc, and yet too low for precision in risk score estimation, the time-varying PS may be helpful.

This article has limitations. First, we do not address challenges to estimating an optimal PS. Second, our scenarios include no unmeasured confounding, time-varying confounding, misclassification, informative censoring, or missing data. These common sources of bias are outside our scope. Third, our findings are supported by illustrative simulated data rather than comprehensive evidence or proof.

In summary, we elucidated how the HRm and HRc diverge when susceptibles are depleted differentially. We showed how this biases PS-based estimators and how the bias can be reduced by updating the PS.

ACKNOWLEDGMENTS

This paper is based on the work of a Sentinel workgroup on hazard ratio estimators that use propensity scores. The authors gratefully acknowledge the contributions to this workgroup of Katherine Freitas, Christian Hampp, Rima Izem, Mark Levenson, and Yuequin Zhao. We also thank Ned Lewis for insightful comments.

REFERENCES

1. Hernán MA. The hazards of hazard ratios. *Epidemiology*. 2010;21:13–15.
2. Aalen OO, Cook RJ, Røysland K. Does Cox analysis of a randomized survival study yield a causal treatment effect? *Lifetime Data Anal*. 2015;21:579–593.
3. Martinussen T, Vansteelandt S, Andersen PK. Subtleties in the interpretation of hazard ratios. *Lifetime Data Anal*. 2020 Jul 11. [Epub ahead of print]
4. Steenland K, Karnes C, Darrow L, Barry V. Attenuation of exposure-response rate ratios at higher exposures: a simulation study focusing on frailty and measurement error. *Epidemiology*. 2015;26:395–401.
5. Stensrud MJ, Aalen JM, Aalen OO, Valberg M. Limitations of hazard ratios in clinical trials. *Eur Heart J*. 2019;40:1378–1383.
6. Platt R, Wilson M, Chan KA, Benner JS, Marchibroda J, McClellan M. The new Sentinel Network–improving the evidence of medical-product safety. *N Engl J Med*. 2009;361:645–647.
7. 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;364:498–499.
8. Toh S, Reichman ME, Houstoun M, et al. Multivariable confounding adjustment in distributed data networks without sharing of patient-level data. *Pharmacoepidemiol Drug Saf*. 2013;22:1171–1177.
9. Gruber S, Zhang Z, Wellman R, et al. Evaluation of Propensity Score Based Methods in Sentinel Study Settings Using Simulation Experiments. *Sentinel Coordinating Center*. May 29, 2019. https://www.sentinelinitiative.org/sentinel/methods/evaluation-propensity-score-based-methods-sentinel-study-settings-using-simulation. Accessed December 27, 2019.
10. Austin PC. The performance of different propensity score methods for estimating marginal hazard ratios. *Stat Med*. 2013;32:2837–2849.
11. Austin PC. The use of propensity score methods with survival or time-to-event outcomes: reporting measures of effect similar to those used in randomized experiments. *Stat Med*. 2014;33:1242–1258.
12. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behav Res*. 2011;46:399–424.
13. Hade EM, Lu B. Bias associated with using the estimated propensity score as a regression covariate. *Stat Med*. 2014;33:74–87.
14. Wyss R, Gagne JJ, Zhao Y, et al. Use of time-dependent propensity scores to adjust hazard ratio estimates in cohort studies with differential depletion of susceptibles. *Epidemiology*. 2020;31:82–89.
15. Bross ID. Spurious effects from an extraneous variable. *J Chronic Dis*. 1966;19:637–647.
16. Smith LH, VanderWeele TJ. Bounding bias due to selection. *Epidemiology*. 2019;30:509–516.
17. Franklin JM, Schneeweiss S, Polinski JM, Rassen JA. Plasmode simulation for the evaluation of pharmacoepidemiologic methods in complex healthcare databases. *Comput Stat Data Anal*. 2014;72:219–226.
18. Chrischilles EA, Gagne JJ, Fireman B, et al. Prospective surveillance pilot of rivaroxaban safety within the US Food and Drug Administration Sentinel System. *Pharmacoepidemiol Drug Saf*. 2018;27:263–271.
19. Statistical Analysis Software 9.3. Cary, NC: Statistical Analysis Software (SAS) Institute; 2002.
20. Lin DY, Wei LJ. The robust inference for the proportional hazards model. *J Am Stat Assoc*. 1989;84:1074–1078.
21. Hansen BB. The prognostic analogue of the propensity score. *Biometrika*. 2008;95:481–488.
22. Ray GT, Lewis N, Klein NP, Daley MF, Lipsitch M, Fireman B. Depletion-of-susceptibles bias in analyses of intra-season waning of influenza vaccine effectiveness. *Clin Infect Dis*. 2019;70:1484–1486.