

# Standardizing Discrete-Time Hazard Ratios With a Disease Risk Score

David B. Richardson\*, Alexander P. Keil, Jessie K. Edwards, Alan C. Kinlaw, and Stephen R. Cole

\* Correspondence to Dr. David B. Richardson, Department of Epidemiology, School of Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599 (email: david.richardson@unc.edu).

Initially submitted June 11, 2019; accepted for publication April 14, 2020.

The disease risk score (DRS) is a summary score that is a function of a potentially large set of covariates. The DRS can be used to control for confounding by the covariates that went into estimation of the DRS and obtain a standardized estimate of an exposure's effect on disease. However, to date, literature on the DRS has not addressed analyses that focus on estimation of survival or hazard functions, which are common in epidemiologic analyses of cohort data. Here, we propose a method for standardization of hazard ratios using the DRS in longitudinal analyses of the association between a binary exposure and an outcome. This approach to handling a potentially large set of covariates through a model-based approach to standardization may provide a useful tool for cohort analyses of hazard ratios and may be particularly well-suited to settings where an exposure propensity score is difficult to model. Simulations are used in this paper to illustrate the approach, and an empirical example is provided.

cohort studies; disease risk score; regression analysis; standardization

Abbreviations: DRS, disease risk score; EPS, exposure propensity score; IPE, inverse probability of exposure; SMR, standardized mortality ratio.

Suppose that an investigator wants to estimate the association between an exposure variable and a binary outcome variable while using standardization to control for confounding. One approach to achieving standardization in a regression model framework is to use a balancing score, such as the exposure propensity score (EPS) or the disease risk score (DRS), to reduce a potentially large vector of covariates to a single scalar summary variable (1–5). For risk ratios or prevalence ratios, recent work has demonstrated that this can be done using inverse probability of exposure (IPE) weights or DRS-based methods (3, 5–7). For hazard ratios, previous work has demonstrated how IPE weights can be used to obtain a standardized estimate in a proportional hazards model framework (8–10). However, an approach to the use of the DRS for analyses of hazard ratios has not been proposed.

We propose an approach for estimating the effect of a binary point exposure on the hazard of an outcome in a regression model framework that leverages the DRS for covariate control. The DRS may be preferable to the EPS for multivariable standardization if the exposure is extremely

rare (11, 12), relatively novel (i.e., newly emerging), or difficult to model (13).

## METHODS

Consider a study in which deaths have been ascertained without loss to follow-up for a closed cohort. Let  $X_i$  denote a binary point exposure of primary interest and  $Z_i$  denote baseline covariates (such as age at study entry, race, and sex), noting that  $i$  indexes cohort members and we use subscript  $i$  to denote the values of variables for cohort member  $i$ .

Suppose the cohort information is recorded in discrete time, meaning that continuous time has been divided into a sequence of contiguous time periods of equal duration (e.g., person-years). Define study entry as time 0 and administrative censoring at the end of the study,  $\tau$ . Denote by  $T_i$  person  $i$ 's failure time (possibly occurring after  $\tau$ , in which case the person was observed through the end of study), and let  $d_i$  denote an indicator of failure during the study period ( $d = 1$ ) or censoring ( $d = 0$ ). Denote the time of last observation for

person  $i$ ,  $T_i^*$ , equal to  $T_i$  or  $\tau$ , whichever occurs first. Here we focus on analysis in a cohort mortality study, but we note that the proposed methods readily extend to analysis of incident first events and also could apply to repeated independent events.

A person-oriented data structure for a cohort mortality study may include 1 row of data per person in the cohort (Figure 1A), which records  $i$ ,  $X_i$ ,  $\mathbf{Z}_i$ ,  $\tau$ ,  $T_i$ , and  $d_i$ . Alternatively, a person-period data structure includes multiple rows of data per person in the cohort (Figure 1B). Let  $j$  index discrete time from study entry ( $j = 0$ ) and let subscript  $j$  denote the values of variables in time period  $j$ . Note that, throughout,  $X_i$  and  $\mathbf{Z}_i$  denote baseline covariates that do not change over time  $j$ . Denote by  $Y_{ij}$  a binary time-varying indicator of the outcome status associated with person  $i$  in period  $j$  that takes a value of 0 except at time  $T_i$ , when  $Y_{ij}$  is assigned the value of the binary indicator of failure status for person  $i$ ,  $d_i$ . In addition, for each record in the data structure, we define a time-varying variable,  $q_{ij}$ , that equals 1 for periods  $j \leq T_i^*$  and 0 for periods  $T_i^* < j \leq \tau$ .

For the purposes of defining the estimand of interest, let  $h(j|\mathbf{Z}, X)$  denote the discrete-time hazard function, and let  $h^x(j|\mathbf{Z})$  denote the potential discrete-time hazard function for the intervention setting  $X = x$ . Suppose that  $i = 1$  to  $p$  represent the exposed ( $X = 1$ ) cohort members (i.e., the cohort has been sorted by  $X$ ). We define a ratio of standardized potential discrete-time hazards contrasting  $X = 1$  with  $X = 0$ , where the target population is the exposed ( $X = 1$ ), as

$$\frac{\sum_{i=1}^p \sum_{j=0}^{\tau} h^1(j|\mathbf{Z}) S^1(j-1|\mathbf{Z}_1) / \sum_{i=1}^p \sum_{j=0}^{\tau} S^1(j-1|\mathbf{Z}_1)}{\sum_{i=1}^p \sum_{j=0}^{\tau} h^0(j|\mathbf{Z}) S^0(j-1|\mathbf{Z}_1) / \sum_{i=1}^p \sum_{j=0}^{\tau} S^0(j-1|\mathbf{Z}_1)}, \quad (1)$$

where  $S^x(j|\mathbf{Z}_i)$  denotes the probability of surviving through time  $j$ ; that is,

$$S^x(j|\mathbf{Z}_i) = 1 - \sum_{u=0}^j h^x(u|\mathbf{Z}) S^x(u-1|\mathbf{Z}_i).$$

We define  $S^x(-1|\mathbf{Z}_i) = 1$ , and we assume that the discrete-time hazard in any time interval  $j$  is suitably small for this approximation. Note that the quantities in the numerator and denominator of this standardized ratio measure have the form of expected numbers of events divided by expected times in the study, and consequently these quantities could be termed rates.

### Disease risk score

Using just the records for at-risk person-periods observed among the unexposed ( $q_{ij} = 1$  and  $X_i = 0$ ), we can estimate the discrete-time hazard of the outcome in the absence of exposure, which we refer to as the DRS.

Since  $Y_{ij}$  is bounded by 0 and 1, the DRS can be modeled as having a logistic dependence on a set of predictors by fitting a pooled logistic model of the form

$$\log \left( \frac{\Pr(Y_{ij} = 1|X = 0)}{1 - \Pr(Y_{ij} = 1|X = 0)} \right) = \alpha_j + \mathbf{Z}_i \beta,$$

**Table 1.** Definitions of the Outcome Variable ( $m_{ijk}$ ) and the Antilog of an Offset ( $n_{ijk}$ ) in a Proposed Regression Model Fitted to an Expanded Data Structure to Obtain a Standardized Estimate of the Discrete-Time Hazard Ratio

$k^a$	$k = 1$	$k = 0$
$m_{ijk}$	$Y_{ij}^b$	$g(j \mathbf{Z}_i)S(j \mathbf{Z}_i)^{c,d}$
$n_{ijk}$	$q_{ij}^e$	$S(j \mathbf{Z}_i)$

<sup>a</sup> Index for 2 rows of information for each person-period.  $k = 1$  corresponds to the information for estimation of the numerator of the ratio of standardized discrete-time hazards, and  $k = 0$  corresponds to the information for estimation of the denominator of the ratio.

<sup>b</sup>  $Y_{ij}$  is a binary indicator of case status for person  $i$  in period  $j$ .

<sup>c</sup>  $g(j, \mathbf{Z})$  is the disease risk score.

<sup>d</sup>  $S(j, \mathbf{Z})$  is the survival function.

<sup>e</sup>  $q_{ij}$  is a binary indicator of whether person  $i$  was “at risk” in period  $j$ .

where the vector of parameters,  $\alpha_j$ , describes the baseline hazard function and  $\beta$  is a vector of parameters associated with covariates  $\mathbf{Z}$ . Estimation of a vector of  $j$  parameters associated with discrete time intervals of the baseline logit hazard function,  $\alpha_j$ , may be inefficient; and often a smooth function of time,  $j$ , might be specified using, for example, splines or polynomial bases. Also note that product terms for interaction between the time scale and covariates can of course be included in the model, if appropriate. Using the estimated coefficients from the fitted model, the DRS (Figure 1C) may be calculated for all members of the cohort,  $i$ , and all periods,  $j$ , as  $g(j|\mathbf{Z}) = \text{expit}(\hat{\alpha}_j + \mathbf{Z}_i \hat{\beta})$ .

### Estimating standardized ratios

Estimation of the ratio of standardized discrete-time hazards in equation 1 may be done using the DRS. We proceed by assuming  $h^1(j|\mathbf{Z}) = \hat{h}(j|\mathbf{Z}, X = 1)$ , under consistency, and  $h^0(j|\mathbf{Z}) = g(j|\mathbf{Z})$ , under the additional assumptions of correct model specification and conditional exchangeability.

Suppose that we expand the person-period data set to include 2 rows for each person-period: The first corresponds to the information needed for estimation of the numerator of the standardized ratio, and the second corresponds to information needed for estimation of the denominator of the standardized ratio (Figure 1D). Let  $k \in 0, 1$  index these 2 rows of information for each person-period. The expanded data set includes rows indexed by person  $i$ , period  $j$ , and stratum  $k$ .

In this expanded data set,  $k = 1$  corresponds to the information needed for estimation of the numerator of the ratio (Table 1). When  $k = 1$ ,  $m_{ijk} = 1$  and  $n_{ijk} = 1$  correspond to a binary indicator of case status for person  $i$  in period  $j$ ,  $Y_{ij}$ , and an indicator of whether person  $i$  was “at risk” in period  $j$ ,  $q_{ij}$ , respectively.

A)

$i$	$X_i$	$Z_i$	$\tau$	$T_i$	$d_i$
1	0	3	4	4	0
2	1	2	4	3	1

B)

$i$	$j$	$X_i$	$Z_i$	$Y_{ij}$	$q_{ij}$
1	0	0	3	0	1
1	1	0	3	0	1
1	2	0	3	0	1
1	3	0	3	0	1
1	4	0	3	0	1
2	0	1	2	0	1
2	1	1	2	0	1
2	2	1	2	0	1
2	3	1	2	1	1
2	4	1	2	0	0

C)

$i$	$j$	$X_i$	$Z_i$	$Y_{ij}$	$q_{ij}$	$g(j, Z_i)$
1	0	0	3	0	1	0.0047
1	1	0	3	0	1	0.0047
1	2	0	3	0	1	0.0048
1	3	0	3	0	1	0.0049
1	4	0	3	0	1	0.0049
2	0	1	2	0	1	0.0056
2	1	1	2	0	1	0.0057
2	2	1	2	0	1	0.0057
2	3	1	2	1	1	0.0058
2	4	1	2	0	0	0.0059

D)

$i$	$j$	$k$	$X_i$	$Z_i$	$Y_{ij}$	$q_{ij}$	$g(j, Z_i)$	$S(j Z_i)$	$g(j Z_i)S(j Z_i)$	$m_{ijk}$	$n_{ijk}$
1	0	1	0	3	0	1	0.0047	1.0000	0.0047	0.0000	<b>1.0000</b>
1	0	0	0	3	0	1	0.0047	<b>1.0000</b>	0.0047	0.0047	<b>1.0000</b>
1	1	1	0	3	0	1	0.0047	0.9539	0.0045	0.0000	<b>1.0000</b>
1	1	0	0	3	0	1	0.0047	<b>0.9539</b>	0.0045	0.0045	<b>0.9539</b>
1	2	1	0	3	0	1	0.0048	0.9094	0.0044	0.0000	<b>1.0000</b>
1	2	0	0	3	0	1	0.0048	<b>0.9094</b>	0.0044	0.0044	<b>0.9094</b>
1	3	1	0	3	0	1	0.0049	0.8664	0.0042	0.0000	<b>1.0000</b>
1	3	0	0	3	0	1	0.0049	<b>0.8664</b>	0.0042	0.0042	<b>0.8664</b>
1	4	1	0	3	0	1	0.0049	0.8249	0.0041	0.0000	<b>1.0000</b>
1	4	0	0	3	0	1	0.0049	<b>0.8249</b>	0.0041	0.0041	<b>0.8249</b>
2	0	1	1	2	0	1	0.0056	1.0000	0.0056	0.0000	<b>1.0000</b>
2	0	0	1	2	0	1	0.0056	<b>1.0000</b>	0.0056	0.0056	<b>1.0000</b>
2	1	1	1	2	0	1	0.0057	0.9453	0.0053	0.0000	<b>1.0000</b>
2	1	0	1	2	0	1	0.0057	<b>0.9453</b>	0.0053	0.0053	<b>0.9453</b>
2	2	1	1	2	0	1	0.0057	0.8929	0.0049	0.0000	<b>1.0000</b>
2	2	0	1	2	0	1	0.0057	<b>0.8929</b>	0.0049	0.0049	<b>0.8929</b>
2	3	1	1	2	1	1	0.0058	0.8427	0.0049	1.0000	<b>1.0000</b>
2	3	0	1	2	1	1	0.0058	<b>0.8427</b>	0.0049	0.0049	<b>0.8427</b>
2	4	1	1	2	0	0	0.0059	0.7948	0.0047	0.0000	<b>0.0000</b>
2	4	0	1	2	0	0	0.0059	<b>0.7948</b>	0.0047	0.0047	<b>0.7948</b>

**Figure 1.** Possible data structures for 2 people in a cohort study. A) A person-level data structure; B), a person-period data structure; C) a person-period data structure with a disease risk score appended; D) an expanded data structure with 2 records per person-period. In these data structures,  $i$  indexes cohort members,  $X_i$  is a binary point exposure of primary interest,  $Z_i$  denotes baseline covariates,  $\tau$  is the administrative end of the study,  $T_i$  is person  $i$ 's failure time,  $d_i$  denotes a binary indicator of failure during the study period ( $d = 1$ ) or censoring ( $d = 0$ ),  $j$  indexes discrete time,  $Y_{ij}$  is a binary indicator of case status for person  $i$  in period  $j$ ,  $q_{ij}$  is a binary indicator of whether person  $i$  was "at risk" in period  $j$ ,  $g(j|Z)$  is the disease risk score,  $k$  is an index for 2 rows of information for each person-period (where  $k = 1$  corresponds to the information for estimation of the numerator of the ratio of standardized discrete-time hazards and  $k = 0$  corresponds to the information for estimation of the denominator of the ratio),  $S(j, Z)$  is the survival function,  $m_{ijk}$  is the outcome variable in the proposed regression model for obtaining a standardized estimate of the discrete-time hazard ratio (gray fill illustrates which values inform calculation of  $m_{ijk}$ ), and  $n_{ijk}$  is used to define an offset in the proposed regression model for obtaining a standardized estimate of the discrete-time hazard ratio (boldface text illustrates which values inform calculation of  $n_{ijk}$ ).

In the expanded data set,  $k = 0$  corresponds to the information needed for estimation of the denominator of the ratio (Table 1). When  $k = 0$ ,  $m_{ijk} = 0$  and  $n_{ijk} = 0$  correspond to

$g(j|Z_i)S(j|Z_i)$  and  $S(j|Z_i) = 1 - \sum_{t=0}^j g(t|Z_i)S(t-1|Z_i)$ , respectively, where  $S(j, Z)$  is the survival function and we define  $S(-1|Z_i) = 1$ .

Using just the records with  $X = 1$ , a regression model fitted to the expanded data structure may take the form

$$\log(E[m_{ijk}]) = \delta_0 + \delta_1 k + \log(n_{ijk}),$$

where the antilog of estimated parameter  $\delta_1$  estimates the standardized ratio comparing the observed discrete-time hazard at  $X = 1$  with the expected discrete-time hazard among persons with  $X = 1$  had their exposure been set to  $X = 0$ . When  $n_{ijk} = 0$ , a small constant is added to allow

calculation of  $\log(n_{ijk})$ . Estimation of robust confidence intervals is recommended given the 2-stage regression (first estimation of the DRS and second fitting the marginal structural model) (14). In the Web Appendix (available at <https://academic.oup.com/aje>), we provide illustrative SAS code (SAS Institute, Inc., Cary, North Carolina) for obtaining discrete-time hazard ratios with both robust confidence intervals and bootstrapped confidence intervals.

### Simulation example

We use simulated data under a cohort study design to demonstrate the implementation of the proposed approach and as confirmatory of the mathematical results (rather than as a comprehensive analysis of finite sample properties). Data were simulated for 1,000 cohort studies with 10,000 members in each cohort. In each simulation, at baseline we generated 10 covariates, denoted  $Z_1$ – $Z_{10}$ . Among these,  $Z_1$ – $Z_4$  were confounders associated with both exposure and outcome,  $Z_5$ – $Z_7$  were exposure predictors, and  $Z_8$ – $Z_{10}$  were outcome predictors.  $Z_1$ ,  $Z_3$ ,  $Z_5$ ,  $Z_6$ ,  $Z_8$ , and  $Z_9$  were random binary variables, and the others were continuous variables assigned as the absolute value of standard normal random variables with mean 0 and variance 1. The relationships between variables follow the structure described by Lee et al. (15), with correlations induced between several of the variables (Figure 1). We generated a random binary exposure,  $X$ , with an exposure prevalence of approximately 10%; we encoded dependence of  $X$  on covariates by specifying that  $X$  took a value of 1 with probability  $1/\{1 + \exp[-(-0.1 - 1 \times Z_1 - 0.5 \times Z_2 - 0.5 \times Z_3 - 0.5 \times Z_4 - 0.5 \times Z_5 - 0.5 \times Z_6 - 0.5 \times Z_7)]\}$ .

Each person entered the study at  $t = 0$  and was followed until  $t = 20$ . Each person-year, we generated a random binary outcome  $Y$  in which we encoded dependence of the outcome on  $X$ , covariates, and time in the study. Two simulation scenarios were examined. In the first simulation scenario, there was homogeneity of stratum-specific ratios, such that  $Y$  took a value of 1 with probability  $\exp(-3 + 0.1 \times \log(t) + 1 \times X - 0.5 \times Z_1 - 0.1 \times Z_2 - 0.5 \times Z_3 - 0.1 \times Z_4 - 0.5 \times Z_8 - 0.5 \times Z_9 - 0.1 \times Z_{10})$ . In the second simulation scenario, there was heterogeneity of stratum-specific ratios, such that  $Y$  took a value of 1 with probability  $\exp(-3 + 0.1 \times \log(t) + 1 \times X - 0.5 \times Z_1 - 0.1 \times Z_2 - 0.5 \times Z_3 - 0.1 \times Z_4 - 0.5 \times Z_8 - 0.5 \times Z_9 - 0.1 \times Z_{10} + 1 \times X \times Z_1)$ .

First, we estimated the crude discrete-time hazard ratio by fitting a regression model for  $Y$  as a function of  $X$ , and we estimated a covariate-adjusted hazard ratio by fitting a regression model for  $Y$  as a function of  $X$ ,  $\log(t)$ , and  $Z_1$ – $Z_{10}$ , including each covariate as a main effect in the regression model (and not including any product terms for interaction between covariates or between covariates and the exposure of interest,  $X$ ).

Next, we estimated the DRS by fitting a regression model to predict  $Y$  as a function of  $\log(t)$  and  $Z_1$ – $Z_{10}$  among the unexposed ( $X = 0$ ), including each covariate as a main effect in the regression model (and not including any product terms for interaction between covariates). We used the method described in this paper (and the SAS code shown in the Web

Appendix) to obtain a standardized estimate of the discrete-time hazard ratio, and we used the robust variance estimator.

We also estimated a standardized ratio using an IPE-weighted marginal structural log binomial model, with robust variance (16). We fitted a logistic model to each simulated cohort at baseline to predict  $X$  as a function of  $Z_1$ ,  $Z_2$ ,  $\dots$ ,  $Z_{10}$ , including each covariate as a main effect in the regression model (and not including any product terms for interaction between covariates). The predicted probability of exposure from the fitted model is the EPS. We constructed standardized mortality ratio (SMR) weights to estimate the standardized ratio where the target population is the exposed group (2, 17); exposed cohort members are given a weight of 1, while unexposed cohort members are given weights that are defined as the ratio of the estimated propensity score to 1 minus the estimated propensity score. We also constructed stabilized IPE weights to estimate the ratio for scenarios where the target population is the total study population; exposed cohort members are given a weight defined as the ratio of the marginal probability of exposure to the estimated propensity score, while unexposed cohort members are given a weight defined as the ratio of 1 minus the marginal probability of exposure to 1 minus the estimated propensity score. In each analysis, we summarized results from 1,000 simulated cohorts by computing the mean log discrete-time hazard ratio (log hazard ratio), the estimated standard deviation of the 1,000 log hazard ratios, and the average estimated robust standard error of the log hazard ratio. Simulations were conducted over the 1,000 cohorts, and the mean values of the exposure group-specific log hazard ratios and associated 95% confidence intervals were calculated.

### Empirical example

We used the method described in this paper to estimate a standardized ratio, and also estimated a standardized ratio using an IPE-weighted marginal structural log binomial model, in empirical data from Kalbfleisch and Prentice (18) that were derived from the Veteran's Administration Lung Cancer Trial, a study in which there was imbalance in covariates across treatment arms. The data concerned a cohort of 137 male patients with advanced, inoperable lung cancer who were given either standard therapy or experimental chemotherapy. The exposure of primary interest was a binary variable coded "1" for treatment by the test chemotherapy or "0" for standard therapy, and the outcome of interest was vital status (1 = deceased, 0 = censored), with survival time recorded in days; in the current analysis, we assumed administrative censoring at 999 days. Covariates included histological type of the tumor (coded 1 for squamous cell carcinoma, 2 for small cell carcinoma, 3 for adenocarcinoma, or 4 for large cell carcinoma), age at study entry (in years), Karnofsky performance score for overall patient status at study entry (range, 0–100), time between diagnosis and start of the study (in months), and a binary indicator for whether the patient had received prior therapy. We derived standardized estimates of the discrete-time hazard ratio for scenarios where the target population was the exposed group and used the robust variance estimator.

**Table 2.** Estimated Difference in the Mean Log Discrete-Time Hazard Ratio, Empirical Standard Error, and Average Estimated Standard Error for 1,000 Cohorts With 1,000 Observations

Simulation Setup	Simulation Result		
	Log HR	ESE	ASE <sup>a</sup>
Simulation scenario 1 (homogeneity)			
Crude	1.13	0.05	0.05
Adjusted in multivariable outcome model	1.00	0.05	0.05
EPS-weighted (ATE)	0.93	0.06	0.06
EPS-weighted (ATT)	0.93	0.05	0.05
DRS-based (equation 1; ATT)	0.93	0.05	0.04
Simulation scenario 2 (heterogeneity)			
Crude	1.38	0.04	0.04
Adjusted in multivariable outcome model	1.28	0.05	0.05
EPS-weighted (ATE)	1.38	0.05	0.06
EPS-weighted (ATT)	1.18	0.05	0.05
DRS-based (equation 1; ATT)	1.18	0.05	0.04

Abbreviations: ASE, average standard error; ATE, average treatment effect in the total population; ATT, average treatment effect in the treated; DRS, disease risk score; EPS, exposure propensity score; ESE, empirical standard error; HR, hazard ratio.

<sup>a</sup> For crude and multivariable-adjusted estimates, ASE is the average standard error estimated across 1,000 simulated cohorts. For DRS- and EPS-based estimates, ASE is the average robust standard error.

## RESULTS

### Simulations

In the first simulation scenario, there was homogeneity of stratum-specific ratios. The crude estimate differed from the value specified under the simulation setup (Table 2). The covariate-adjusted estimate was equivalent to the specified value.

The hazard ratio is not a collapsible measure of association, meaning that the standardized marginal estimate of the hazard ratio need not equal the covariate-adjusted hazard ratio estimate, even in the absence of confounding (19, 20). Consistent with that fact, the DRS-based estimate does not equal the covariate-adjusted estimate. However, the DRS-based method yielded an estimate of association that was equivalent to the estimates obtained from marginal structural log-binomial models using IPE weights and SMR weights (Table 2). Note that when there is homogeneity of the effect measure across levels of covariates, the choice of target population does not affect estimates of the standardized ratios, and the IPE- and SMR-weighted estimates will be equivalent.

In the second simulation scenario, there was heterogeneity of stratum-specific ratios (Table 2). Again, the DRS-based method yielded an estimate of association that was equivalent to the estimates obtained from marginal structural log-binomial models using SMR weights. Given heterogeneity of the discrete-time hazard ratio across levels of covariates, the choice of target population does matter, and the IPE-weighted estimate of the ratio (which targets the total population) differs from the estimate obtained using SMR-type weights (which, like the proposed DRS-based method, targets the exposure effect among the exposed).

### Empirical example

The empirical cohort included 137 patients, of whom 68 (49.6%) received chemotherapy and 128 (93.4%) were deceased by the end of follow-up. Chemotherapy-treated patients were more likely to have squamous cell carcinoma (57% vs. 43%) or adenocarcinoma (67% vs. 33%) than standard therapy patients. Chemotherapy-treated patients were slightly older than standard-therapy patients (mean age = 59 years vs. 57 years). Chemotherapy-treated patients were similar to standard therapy patients with respect to Karnofsky performance score, time between diagnosis and the start of the study, and proportion of patients who had received another type of therapy before the current one.

Fitting the proposed DRS-based model yielded a standardized discrete-time hazard ratio estimate of 0.85 (95% confidence interval: 0.60, 1.18). Fitting a marginal structural model with SMR weights yielded a standardized hazard ratio estimate of 0.86 (95% confidence interval: 0.55, 1.33) when the target population was the exposed cohort. Similarly to the simulations, the proposed DRS-based model yielded an estimated discrete-time hazard ratio that was very similar to the SMR-weighted estimate obtained using marginal structural models with weights derived using the EPS.

## DISCUSSION

This paper has focused on implementation of regression models using a DRS to obtain estimates of standardized discrete-time hazard ratios. The proposed DRS-based model may be appealing in study settings where the EPS is difficult to model well. In simulations and in our empirical example, it was feasible to estimate both the DRS and the EPS; however, in practice there are settings in which one approach or the other may be preferable or most feasible. Use of the DRS may be attractive, for example, if the exposure of interest is newly emerging or extremely rare, because under such conditions it may be difficult to reliably model the EPS. When exposure does not occur in some strata of covariates, the EPS may be difficult or impossible to estimate. For example, in a study of the safety or effectiveness of a newly introduced therapy, the DRS may be appealing, relative to the EPS, if there are historical data with which to model the DRS before introduction of the new therapy. Similar settings arise in occupational and environmental studies, where a novel environmental contaminant or a change in industrial process occurs, if there are historical data with which to model the DRS prior to that change. Our proposed approach focuses



on estimation of the exposure effect among the exposed, and therefore it pertains only to strata of covariates where there were exposed cohort members. The proposed approach leverages a DRS model fitted using the data for the unexposed cohort members to estimate the potential outcomes expected among the exposed had they been unexposed.

An issue that has received attention in the literature on the DRS is what group to use to fit the model for estimation of the DRS (13). We note that in the proposed approach we split the data, estimating the DRS only in the records for the unexposed and estimating the standardized discrete-time hazard ratio only using records for the exposed. The problem of “overfitting” to the unexposed by estimating the DRS in the same sample as that used for the calculated predictions may be minimized in this way; however, further work on this topic is warranted. We note that the DRS,  $g(j, \mathbf{Z})$ , could be estimated in a well-chosen out-of-sample group, and estimation of the discrete-time hazard ratio would proceed in the same fashion as when the DRS is estimated in an unexposed in-sample group.

One challenge to using a DRS-based method is the difficulty of assessing whether the model for estimation of the DRS is appropriately specified. Wyss et al. (21) have suggested a “dry-run” analysis in which the referent group is partitioned into pseudo-exposed and pseudo-unexposed groups so that differences in the observed covariates resemble differences between the actual exposed and unexposed populations but the adjusted (pseudo-)exposure–outcome association is expected to be null. Under their proposed approach, a DRS model is evaluated by its ability to retrieve an unconfounded null estimate after adjustment in this “dry-run” analysis. Such an approach could also be applied in the setting considered in the current paper, where the target of estimation is the standardized discrete-time hazard ratio.

In this paper we focused on analyses of a single outcome of interest, setting aside the issue of competing risks. We note that this approach can be extended to analyses of cause-specific mortality in which we model 2 (or more) competing risks for mortality. As in standard estimates of the cause-specific hazard ratio, common risk factors should be controlled for in the analytical approach (e.g., via regression adjustment or weighting), and such approaches may be necessary with DRS-based estimates of the discrete-time hazard ratio as well.

Classical approaches to standardization of rates commence with categorization of covariates and cross-classification of person-time and events into these categories. The alternative approach considered here commences with a balancing score, which we refer to as the DRS. Using a regression model, this approach can allow us to reduce a potentially large vector of covariates to a single scalar summary variable; the approach accommodates continuous covariates as well as ordinal or nominal categorical variables, and rather than complete saturation of the model, as implemented classically by the cross-classification of information by levels of all covariates, simpler models for the DRS may be fitted if appropriate.

Hansen (6) laid out a basis for understanding the DRS as a method for inducing prognostic balance in observed data for studies of response to treatment. Here we have

extended the use of the DRS to induce prognostic balance in time-to-event analyses, by which, in the absence of right-censoring, we are referring to a setting where the potential time to the event of interest under the control treatment is independent of covariates,  $Z_i$ , conditional on the DRS. In our setting of estimation of a ratio of standardized discrete-time hazards, where the target population is the exposed, this is achieved by taking the DRS as an estimate of  $h^0(j|\mathbf{Z})$  and using this quantity to generate potential times to event among the exposed had they been unexposed. Prognostic balance hinges upon our estimated DRS function  $g(j|\mathbf{Z})$  being sufficient to serve this purpose. Further work, formally demonstrating the properties of prognostic balance in the setting of discrete-time hazards, is beyond the scope of the current paper but would make a useful contribution to the literature on the DRS. Other investigators have described related approaches to using a DRS in the context of proportional hazards regression, implemented through matching, stratification, or covariate adjustment (e.g., entering the score by means of indicator terms for categories) (22, 23). In this paper we describe an approach to estimation of a ratio of standardized discrete-time hazard ratios that avoids coarsening (as might occur when matching, stratifying, or adjusting for categories defined by the DRS) or the need to specify a parametric form when including the DRS as a continuous explanatory variable in a regression model.

The approach described here focuses on a setting where  $Z_i$  denotes baseline covariates. We have not addressed the setting of time-varying covariates, and notably have not addressed settings of time-varying covariates affected by prior exposure. Extension of the DRS to these more analytically complex settings is beyond the scope of the current paper but is an important direction for future work.

We have focused on a setting involving a contrast defined by a binary point exposure of primary interest. In recent work, we have discussed conditions and approaches to analysis of standardized risk ratios using the DRS when the exposure variable has more than 2 levels (24). Extension of the present approach to a contrast involving more than 2 categories may be considered in future work involving estimation of standardized discrete-time hazard ratios when the exposure variable has more than 2 levels.

The proposed approach yields a standardized discrete-time hazard ratio that may be useful for summarization of research findings. This standardized ratio corresponds to an estimate of a causal contrast, the exposure effect among the exposed. This has the appeal of conveying an estimate that pertains to a well-defined target population, and it is a useful summary measure of potential interest in a range of settings. In a study of the safety or effectiveness of a therapy, the treatment effect among the treated is often what is desired; similarly, in occupational settings, a manager or regulator considering a policy or intervention designed to reduce exposure to an agent often desires an estimate of the exposure’s effect among the exposed. In such settings, a marginal estimate of the exposure’s effect, in the form of the proposed standardized ratio, is one potentially appealing quantitative summary. As illustrated in the Web Appendix, calculation of the proposed DRS-based standardized discrete-time hazard ratio is readily performed.

## ACKNOWLEDGMENTS

Author affiliations: Department of Epidemiology, School of Public Health, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina (David B. Richardson, Alexander P. Keil, Jessie K. Edwards, Stephen R. Cole); and Division of Pharmaceutical Outcomes and Policy, Eshelman School of Pharmacy, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina (Alan C. Kinlaw).

D.B.R. was supported by the National Institute for Occupational Safety and Health (grant R01 OH011409). A.P.K. received funding from the Eunice Kennedy Shriver National Institute of Child Health and Human Development (grant DP2-HD-08-4070). This study was sponsored in part by the Alpha Foundation for the Improvement of Mine Safety and Health, Inc. (Philadelphia, Pennsylvania).

The views, opinions, and recommendations expressed herein are solely those of the authors and do not imply any endorsement by the Alpha Foundation or its directors or staff.

Conflict of interest: none declared.

## REFERENCES

1. Kurth T, Walker AM, Glynn RJ, et al. Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of nonuniform effect. *Am J Epidemiol*. 2006;163(3):262–270.
2. Brookhart MA, Wyss R, Layton JB, et al. Propensity score methods for confounding control in nonexperimental research. *Circ Cardiovasc Qual Outcomes*. 2013;6(5):604–611.
3. Robins JM, Hernán MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology*. 2000;11(5):550–560.
4. Tadrous M, Gagne JJ, Stürmer T, et al. Disease risk score as a confounder summary method: systematic review and recommendations. *Pharmacoepidemiol Drug Saf*. 2013;22(2):122–129.
5. Stürmer T, Schneeweiss S, Brookhart MA, et al. 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.
6. Hansen BB. The prognostic analogue of the propensity score. *Biometrika*. 2008;95(2):481–488.
7. Richardson DB, Keil AP, Kinlaw AC, et al. Marginal structural models for risk or prevalence ratios for a point exposure using a disease risk score. *Am J Epidemiol*. 2019;188(5):960–966.
8. Cole SR, Hernán MA. Constructing inverse probability weights for marginal structural models. *Am J Epidemiol*. 2008;168(6):656–664.
9. Hernán MA, Robins JM. Estimating causal effects from epidemiological data. *J Epidemiol Community Health*. 2006;60(7):578–586.
10. Naimi AI, Moodie EE, Auger N, et al. Constructing inverse probability weights for continuous exposures: a comparison of methods. *Epidemiology*. 2014;25(2):292–299.
11. Hajage D, Tubach F, Steg PG, et al. On the use of propensity scores in case of rare exposure. *BMC Med Res Methodol*. 2016;16:Article 38.
12. Arbogast PG, Ray WA. Use of disease risk scores in pharmacoepidemiologic studies. *Stat Methods Med Res*. 2009;18(1):67–80.
13. Glynn RJ, Gagne JJ, Schneeweiss S. Role of disease risk scores in comparative effectiveness research with emerging therapies. *Pharmacoepidemiol Drug Saf*. 2012;21(suppl 2):138–147.
14. Huber PJ. The behavior of maximum likelihood estimates under nonstandard conditions. In: Le Cam LM, Neyman J, eds. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*. Berkeley, CA: University of California Press; 1967:221–233.
15. Lee BK, Lessler J, Stuart EA. Improving propensity score weighting using machine learning. *Stat Med*. 2010;29(3):337–346.
16. Richardson DB, Kinlaw AC, MacLehose RF, et al. Standardized binomial models for risk or prevalence ratios and differences. *Int J Epidemiol*. 2015;44(5):1660–1672.
17. Sato T, Matsuyama Y. Marginal structural models as a tool for standardization. *Epidemiology*. 2003;14(6):680–686.
18. Kalbfleisch JD, Prentice RL. *The Statistical Analysis of Failure Time Data*. 2nd ed. Hoboken, NJ: John Wiley & Sons, Inc.; 2002.
19. Greenland S. Absence of confounding does not correspond to collapsibility of the rate ratio or rate difference. *Epidemiology*. 1996;7(5):498–501.
20. Gail MH, Wieand S, Piantadosi S. Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika*. 1984;71(3):431–444.
21. Wyss R, Hansen BB, Ellis AR, et al. The “dry-run” analysis: a method for evaluating risk scores for confounding control. *Am J Epidemiol*. 2017;185(9):842–852.
22. Bohn J, Schneeweiss S, Glynn RJ, et al. Controlling confounding in a study of oral anticoagulants: comparing disease risk scores developed using different follow-up approaches. *EGEMS (Wash DC)*. 2019;7(1):Article 27.
23. Desai RJ, Wyss R, Jin Y, et al. Extension of disease risk score-based confounding adjustments for multiple outcomes of interest: an empirical evaluation. *Am J Epidemiol*. 2018;187(11):2439–2448.
24. Richardson DB, Keil AP, Cole SR, et al. Assessing exposure-response trends using the disease risk score. *Epidemiology*. 2020;31(2):e15–e16.