Statistical Theory and Related Fields

ISSN: 2475-4269 (Print) 2475-4277 (Online) Journal homepage: https://www.tandfonline.com/loi/tstf20

Nutritional epidemiology methods and related statistical challenges and opportunities

Ross L. Prentice & Ying Huang

To cite this article: Ross L. Prentice & Ying Huang (2018) Nutritional epidemiology methods and related statistical challenges and opportunities, Statistical Theory and Related Fields, 2:1, 2-10, DOI: 10.1080/24754269.2018.1466098

To link to this article: https://doi.org/10.1080/24754269.2018.1466098

STATISTICAL THEORY AND RELATED FIELDS
2018, VOL. 2, NO. 1, 2–10
https://doi.org/10.1080/24754269.2018.1466098

Taylor & Francis
Taylor & Francis Group

Nutritional epidemiology methods and related statistical challenges and opportunities

Ross L. Prentice and Ying Huang

Fred Hutchinson Cancer Research Center and University of Washington, Seattle, WA, USA

ABSTRACT

The public health importance of nutritional epidemiology research is discussed, along with methodological challenges to obtaining reliable information on dietary approaches to chronic disease prevention. Measurement issues in assessing dietary intake need to be addressed to obtain reliable disease association information. Self-reported dietary data typically incorporate major random and systematic biases. Intake biomarkers offer potential for more reliable analyses, but biomarkers have been established only for a few dietary variables, and these may be too expensive to apply to all participants in large epidemiologic cohorts. A possible way forward involves additional nutritional biomarker development using high-dimensional metabolomic profiling, using blood and urine specimens, in conjunction with further development of statistical approaches for accommodating measurement error with failure time response data. Statisticians have the opportunity to contribute greatly to worldwide public health through the development of statistical methods to address these nutritional epidemiology research challenges, as is elaborated in this contribution.

ARTICLE HISTORY
Received 7 September 2017
Accepted 14 April 2018

KEYWORDS
Chronic disease; epidemiology; failure time data; hazard ratio; measurement error; metabolomics; nutritional biomarker; regression calibration

1. Introduction

Chronic diseases constitute the major cause of morbidity and mortality in many countries worldwide, especially in countries that are more economically developed. In fact, the incidence of cardiovascular diseases, major cancers and diabetes tends to be several times higher in economically developed populations than in other parts of the world (e.g., Forman et al., 2014). Much of this elevated incidence appears to be driven by modifiable exposures, since migrant populations tend to develop disease rates similar to those in their new environment within a generation or two of migration, even though the acculturation process may span some decades. However, the primary drivers for the observed risk elevations for specific chronic diseases are not well understood.

Diet and physical activity patterns over the lifespan provide natural candidate exposures to explain chronic disease risk variations among populations, as well as chronic disease risk changes over time in specific populations. However, when expert committees have been assembled to review the analytic epidemiology literature on these patterns and exposures they have mostly concluded that there are few nutrition and chronic disease associations that can be viewed as established, or even as probable (World Cancer Research Fund and American Institute for Cancer Research, 1997, 2007; World Health Organization, 2003). In contrast, ecological analyses tend to exhibit strong correlations between national incidence rates and per capita food 'disappearance' measures, especially for such food components as total energy and total fat (Armstrong & Doll, 1975; Prentice & Sheppard, 1990).

Much of the explanation for these apparently discrepant findings likely resides in the properties and quality of available dietary data. Analytic epidemiology studies mostly rely on self-reported dietary intake data, with prominent assessment methodologies involving food frequencies, food records or dietary recalls. At best these measurement approaches yield noisy estimates of targeted intakes, which are usually expressed as daily average intakes over a short period of a few days to a few months. The noise feature alone typically leads to greatly attenuated associations, necessitating studies having a large number of incident cases of disease for associations to be evident. A larger issue is systematic bias in the self-report assessments, corresponding to differential reporting by study subjects according to such personal characteristics as body mass index (BMI) defined as weight in kilograms divided by the square of height in metres, age and ethnicity. These random and systematic biases may combine to thoroughly distort, or possibly even reverse, disease association estimates. In comparison, ecological analyses that compare dietary intakes among population groups (e.g., countries) can be expected to be relatively free from influences due

CONTACT Ross L. Prentice rprentice@fredhutch.org, rprentic@whi.org Fred Hutchinson Cancer Research Center and University of Washington, Seattle, WA, USA

to the noise component of measurement error, even if based on individual self-report data, but systematic bias along with potential ecological confounding, preclude a strong reliance on related disease association analyses.

What then are the study designs that can help to develop reliable information on dietary intakes and patterns for chronic disease risk reduction? Certainly randomised, controlled dietary intervention trials have the potential to be informative. However, there are substantial challenges related also to this research strategy. Such trials typically need to be quite large for change to a new dietary pattern to appreciably offset preceding years or decades of the study participant's usual diet, and usually need to be of long duration for the same reason. Hence, randomised dietary intervention trials with chronic disease outcomes can be quite expensive and logistically challenging, while only evaluating one, or a few, specific dietary patterns. Furthermore, the long trial duration can pose challenges for adherence to the assigned dietary intervention and may open up the possibility of post-randomisation confounding if participants adopt medications and other approaches to chronic disease risk reduction in a differential manner among randomised groups.

In comparison, the use of nutritional biomarkers provides a practical and potentially comprehensive approach to strengthening nutritional epidemiology observational research. If such biomarkers can be obtained from biospecimens, typically blood or urine specimens, stored on the members of large epidemiology cohorts, then these objective intake measures can be directly associated with subsequent disease risk, perhaps using nested case–control (Prentice & Breslow, 1978; Thomas, 1977) or case–cohort (Prentice, 1986; Self & Prentice 1988) sampling within study cohorts.

Otherwise, biomarker determinations can be made in a cohort subsample, and used to correct self-report data for random and systematic biases, with corrected estimates subsequently associated with disease risk. However, only a few dietary components have established intake biomarkers, and there is a strong research need for the development of additional biomarkers. To be useful, such biomarkers should plausibly adhere to a classical measurement model. Even when such biomarkers can be identified, there is a need for further development of statistical methods and theory for estimating key disease association parameters, such as parameters relating disease hazard ratios to preceding (unobserved) dietary intake histories (Carroll, Ruppert, Stefanski, & Crainiceanu, 2006; Prentice, 1982), in cohort study contexts.

These nutritional epidemiology methodology needs and opportunities have a strong statistical component. In fact, statistical input in this multidisciplinary nutritional epidemiology research area is as crucial to the development of useful and interpretable disease prevention information as is input from any other disciplinary group. Furthermore, the needed research includes most interesting statistical methodology issues, including issues in the use of high-dimensional metabolomic data for nutritional biomarker development, and issues related to estimating disease association parameters in non-linear models when predictor variables include considerable measurement error.

These issues will be elaborated below, in an attempt to encourage additional statistical theoreticians to consider research goals in this important public health research area, especially during this time of national and international crises in diabetes, obesity, major cancers and cardiovascular diseases, in affluent populations.

2. Nutritional biomarker development methods

The principal requirement for a useful nutritional biomarker, w, is adherence to a classical measurement model,

$$w = z + e, \tag{1}$$

where z is the targeted nutritional variable and e is a pure noise error component that is independent of z and other study subject characteristics (e.g., age, ethnicity, BMI, ...) that may be pertinent to the risk of the chronic disease under study. The hallmark of the biomarker is then freedom from systematic bias relative to the targeted dietary variable and relative to risk factors for the outcome under study. Additionally, the variance of e should not be too large compared to the variance of z, so that w provides an efficient biomarker.

Usually z is defined as log-transformed usual daily intake over a certain time period, such as a few weeks or months, while w typically arises as a corresponding log-transformed intake assessment from biospecimens collected at a single point, or a few points, in time. Prominent examples of nutritional biomarkers include a doubly labelled water (DLW) biomarker of energy intake (Schoeller, 1999), a urinary nitrogen biomarker of protein intake (Bingham, 2003), and 24-hour urine-based measures of sodium and potassium intake (Luft, Fineberg, & Sloan, 1982; Rakova et al., 2013). Additionally, our research group, using data from a 153-woman human feeding study, has recently proposed novel biomarkers for the intake of several carotenoids and tocopherols using blood serum concentration measurements (Lampe et al., 2017), and a carbohydrate biomarker using plasma fatty acid profiles (Song et al., 2017). These latter biomarkers required the inclusion of certain study subject variables for (1) to be plausible.

Only a few research groups have engaged in nutritional biomarker identification and development, and

the brevity of the nutritional biomarker list described above strongly suggests that nutrient metabolite recovery in urine along with blood nutrient concentrations will not provide sufficiently comprehensive sources of data for biomarker development. However, urine and blood metabolomic profiling (i.e., studies of small molecule concentrations) provide an intriguing possibility for additional nutritional biomarker development.

Over the past 15 years, high-dimensional genotype data for disease association analyses, and for other purposes, have provided a considerable stimulus to statistical theory development, with methods based on the notion of only a few real associations among many examined, or sparsity, coming to play an influential role (Hastie, Tibshirani, & Wainwright, 2015). These studies have generated lengthy lists of chronic disease-associated genetic variants for many chronic diseases and conditions. Most such associations, however, are very weak and collectively may not explain as much response variation as do simply collected data on family history for the outcomes in question. The difference between the outcome variation explained by family history compared to that explained by measured genetic variates is sometimes referred to as the 'missing heritability '. Another explanation, however, is that much of the observed familial association is attributable to shared environment, including similar diet and activity patterns among family members, rather than to shared genotype.

High-dimensional exposure history data are more complicated to model and analyse than are high-dimensional genotype data for at least two reasons. Unlike time-invariant germline genetic variants that can be assessed or imputed with great precision, environmental exposure data often are assessed with substantial measurement error, as with dietary and activity pattern assessments. Second, exposure patterns for individuals may change in a noteworthy fashion over the years and decades that are relevant to chronic disease risk. Hence, the statistical challenges in using high-dimensional exposure data are substantial in the nutritional epidemiology area, and require the input of theoreticians who are knowledgeable in the application of both high-dimensional data and exposure measurement error methodologies.

The two 'exposome' complexities just mentioned are separable to some extent. Blood and urine metabolomic profiles typically provide measurements that are responsive to recent dietary exposures, for example over the past few days. In that the diets of free living individuals tend to track over time, much may be learned by studying disease risk in relation to dietary exposures over short preceding time periods (e.g., most recent year). The incorporation of dietary changes over an extended period of time may be able to be accomplished by obtaining biospecimens periodically during a lengthy cohort follow-up period, and by relating disease risk at specific follow-up times to a preceding biomarker-based dietary intake history.

Our research group has been developing metabolomic profile data in the context of the human feeding study mentioned above (Lampe et al., 2017) among 153 participants in the U.S. Women's Health Initiative. Profiles developed in the laboratory of Dr. Dan Raftery involve both targeted platforms, typically with 100–200 pre-specified metabolites, and global platforms with a much larger number of metabolites, many of which lack biological identification. Especially, the global platforms, which require peak identifications in mass spectra (e.g., liquid chromatography/mass spectrometry (LC/MS) or gas chromatography/mass spectrometry (GC/MS)), include complex missing data features and a non-ignorable noise component for quantitative measurements. Higher dimensional statistical methods that have proven to be successful in genetic association applications need to be extended to allow for the measurement properties of these types for metabolomic profile data. Without such extension, it seems likely that global platform measurements will be systematically excluded from potential biomarker specifications based on their weak performance in cross-validation components of model building activities, even if the underlying metabolites are highly relevant to the targeted intake.

3. Disease association analysis methods using nutritional biomarkers

Suppose now that data from a study cohort are available as $S = T \wedge C, \delta = I[S = T]$ and $Z(S) = \{z(u); 0 \leq u < S\}$, where S is the smaller of time from cohort enrolment to chronic disease diagnosis (T) or to right censoring (C), δ is a non-censoring indicator, and $Z(S)$ is the history of actual dietary intakes, as well as dietary self-report and potential confounding factors for the study subject up to time S. Cox regression (Cox, 1972, 1975; Kalbfleisch & Prentice, 2002) provides a major tool for studying the association between Z and disease risk, under the usual assumption that the hazard rate for T at following time t does not depend on censoring conditional on $Z(t)$, for any $t > 0$. Under the Cox model, the hazard rate

$$\lambda\{t; Z(t)\} = \lim_{\Delta t \downarrow 0} \text{pr}\{t \leq T < t + \Delta t; T \geq t, Z(t)\}/\Delta t$$

is modelled as

$$\lambda\{t; Z(t)\} = \lambda_{0s}(t) \exp\{x(t)\beta\}, \tag{2}$$

where $x(t) = \{x_1(t), \ldots, x_p(t)\}$ is a data analyst-defined regression vector formed from $\{Z(t), t\}$ with corresponding hazard ratio parameter $\beta = (\beta_1, \ldots, \beta_p)'$ to be estimated, while λ_{0s} is an unspecified 'baseline' hazard rate function at $x(t) \equiv 0$ in stratum s, where the

stratification $s = s\{t; Z(t)\} \in \{1, 2, \ldots\}$ is also defined by the data analyst. Estimation of the association parameter β is based on applying usual maximum likelihood formulae to the 'partial likelihood' function (Cox, 1975)

$$L(\beta) = \prod_{s>0} \prod_{i=1}^{d_s} \left\{ \prod_{k \in D_s(\Delta t_{si})} e^{x_k(t_{si})\beta} \Big/ \sum_{\ell \in R_s(t_{si})} e^{x_\ell(t_{si})\beta} \right\}, \tag{3}$$

where $t_{s1} < t_{s2} < \cdots < t_{sd_s}$ are the uncensored disease incidence times in stratum s, $D_s(\Delta t_{si})$ is the set of individuals failing at time t_{si} in stratum s and $R_s(t_{si})$ is the set of study subjects 'at risk' (i.e., without prior disease diagnosis or censoring) for disease occurrence in stratum s at time t_{si}. The Cox model incorporates substantial flexibility as a result of its nonparametric baseline disease rates and its stratification features, and it is well suited to estimation problems for exposures that may vary over time, and for confounding factors that may also need to be allowed vary over a study follow-up time for an independent censoring assumption to be plausible. Note that the hazard ratio interpretation for β is natural and convenient in many biomedical research contexts, including nutritional epidemiology studies.

Expression (3) also provides a basis for the estimation of β in (2) when data are available only for cases developing disease during cohort follow-up and time-matched 'controls' without disease at the time of corresponding case occurrence simply by regarding each matched case–control set as a distinct stratum (Prentice & Breslow, 1978; Thomas, 1977). Similarly, maximisation of (3) is also appropriate if data are available only on cases and a random sample, or stratified random sample, of the study cohort, with $R_s(t_{si})$ redefined to include only cases occurring in stratum s at time t_{si} and subcohort controls at risk in stratum s at that time. Note that a variance estimator more complex than that from the negative second derivative of $\log L(\beta)$ is required with case–cohort sampling (Prentice, 1986; Self & Prentice, 1988). These ad hoc sampling designs do not have established optimality properties, though efficiency can be expected to be good if case and comparison groups are well matched on potential confounding variables. The corresponding hazard ratio parameter estimates cited above are also suboptimal, with efficiency loss that tends to be larger when some of the covariate components are available for all cohort members. Estimating efficiency can be improved by including inverse probability weights in these estimating equations (Breslow, McNeney, & Wellner, 2003; Breslow & Wellner, 2007), but resulting estimators have not been shown to be semiparametric efficient.

Now consider the estimation of the hazard ratio parameter in (2) when the 'covariate history' $Z(t)$ incorporates measurement error. More specifically suppose that the targeted $x(t)$ in (2) can be written as

$$x(t) = \tilde{x}(t) + \tilde{e}(t), \tag{4}$$

where $\tilde{x}(t)$ values are obtained from available measurements, and $\tilde{e}(t)$ is a measurement error component that is independent of $\tilde{x}(t)$ and potential confounding factors. Also suppose that the stratification variable $s = s\{t; Z(t)\}$ relies only on elements of Z that are free of measurement error. The induced hazard rate model that specifies disease risk given measured data only, at each follow-up time t, can be written (Prentice, 1982) as

$$\lambda_{0s}(t) E\{e^{x(t)\beta}; T \geq t, \tilde{X}(t)\},$$

where $\tilde{X}(t) = \{\tilde{x}(u); 0 \leq u < t\}$ and E denotes expectation. In general, these induced hazard rates involve an expectation factor that is a complicated function of the baseline hazard rates in (2). However, if the disease outcome is rare during the cohort follow-up period, then the conditioning event $T \geq t$ can be ignored to a good approximation. Doing so leads to a hazard rate model under the specialised Berkson-type measurement model (4) of

$$\begin{aligned} &\lambda_{0s}(t) E\{e^{x(t)\beta}; \tilde{X}(t)\} \\ &\quad = \lambda_{0s}(t)\, e^{\tilde{x}(t)\beta} E\{e^{\tilde{e}(t)\beta}; \tilde{X}(t)\} \\ &\quad = \tilde{\lambda}_{0s}(t)\, e^{\tilde{x}(t)\beta} \end{aligned}$$

with the last equality following from the independence of $\tilde{e}(t)$ and $\tilde{x}(t)$ and normality assumptions. Hence if one could identify a data construct $\tilde{x}(t)$ that adheres to (4), one could regress the hazard rate on $\tilde{x}(t)$ in a standard Cox model fashion to estimate the regression coefficient β in (2).

Suppose that an assessment $q(t)$ of $x(t)$ is available for all members of a study cohort, while a biomarker assessment $w(t)$ of $x(t)$ is also available on a random sample from the same population, at all follow-up times $t \geq 0$. If $q(t)$ is a self-report assessment of $x(t)$, then a measurement model

$$q(t) = a_0 + a_1 x(t) + a_2' v(t) + \varepsilon(t) \tag{5}$$

may be appropriate where a_0, a_1 and $a_2 = (a_{21}, a_{22}, \ldots)'$ are constants, $v(t)' = (v_1(t), v_2(t), \ldots)$ are study subject characteristics that may be associated with the measurement properties of $q(t)$ or that may be needed to control confounding in (2), and $\varepsilon(t)$ is a random noise component that is independent of $x(t)$, given $v(t)$. In the biomarker sample, one will have measurements

$$w(t) = x(t) + e(t),$$

where the error $e(t)$ is independent of $x(t)$ and is also independent of study subject characteristics that determine $v(t)$, an assumption that will often be plausible if $Z(t)$ incorporates dietary intake data over a short

time period (e.g., a few months) prior to t. Also, importantly, suppose that the error terms $e(t)$ and $\varepsilon(t)$ are independent given $v(t)$. Then under joint normality assumptions for $\{x(t), \varepsilon(t)\}$ given $v(t)$, one has

$$E\{x(t); q(t), v(t)\} = b_0 + b_1 q(t) + b_2' v(t)$$

for some constants b_0, b_1 and b_2, and $\tilde{x}(t) = \hat{E}\{w(t); q(t), v(t)\}$ satisfies (4) where $\hat{E}$ denotes an estimator of $x(t)$ arising from linear regression of $w(t)$ on $q(t)$ and $v(t)$ in the biomarker sample. In this context, $\tilde{x}(t)$ is referred to as a biomarker calibrated estimate of $x(t)$. Values of $\tilde{x}(t)$ can be calculated for each of the members of the study cohort and the regression parameter β can be estimated by standard Cox regression (Cox, 1975) of the disease outcome data on $\tilde{x}(t)$. A non-standard variance estimator is required for the regression parameter estimator to acknowledge the randomness in calibration equation coefficient estimates. A bootstrap procedure typically works well for variance estimation. The estimation procedure just described simply generalises the regression calibration procedure for failure time data (Carroll et al., 2006; Prentice, 1982) to a broader class of measurement error models.

In some applications, $q(t)$ may also be a biomarker measurement that is available on the entire cohort, or on a suitable set of cases and controls drawn from the cohort. The calibration procedures may be applied as above, though the $v(t)$ term can then be dropped from the calibration equation. Note that the error terms for the two biomarker assessments in the biomarker sample need to be statistically independent in this context, with implications for the exposure time period used in the definition of $Z(t)$ in (2).

The above procedures depend on the biomarker adhering to a classical measurement model, the disease under study being infrequent (e.g., $< 10\%$) during cohort follow-up, and the so-called instrumental variable $q(t)$ adhering to (5) with error term $\varepsilon(t)$ that is independent of the error term $e(t)$ for the biomarker given $v(t)$. These assumptions will often be appropriate in nutritional epidemiology contexts for dietary exposure variables having an established biomarker. The regression calibration procedure outlined above also assumed the log-hazard rate to depend linearly on the modelled exposure variable $x(t)$. Additional hazard ratio regression modelling choices will also be of interest for the exploration and presentation of nutritional epidemiology data. However, estimation procedures for such other modelling choices have received little attention to date, when the measured exposure variables incorporate substantial measurement error.

For example, it is common to display epidemiological data by showing estimated hazard ratios, or closely related odds ratios, across quartiles or quintiles of the modelled exposure variable. One possibility for the estimation of hazard ratios across such quantiles, assuming model (2), is to calibrate the exposure variable, then estimate hazard ratios based on quantiles of the calibrated exposure. Another possibility is to define $x(t)$ in (2) to be a set of quantile indicator variables, typically taking the smallest or largest quantile as the base value for hazard ratio comparison. One can then consider a regression calibration procedure of the type outlined above with $\tilde{x}(t)$ defined as a set of calibrated quantile indicators for each quantile except the comparator. Simulation studies described in the following sections show, perhaps surprisingly, that the second approach has better performance than the first and even enjoys some robustness to departure from the rare disease assumption used in the calibration procedure. The main point here, however, is that hazard ratio estimation procedures are needed to handle a variety of regression model forms in (2), as an integral component of nutritional epidemiology association analysis methods when biomarker data are available in a study cohort, or in appropriate subsamples thereof.

4. Hazard ratio estimation for exposure quantiles

It is commonplace in epidemiological reporting to show estimated hazard ratios across quantiles of key univariate exposure variables. The regression calibration approach outlined above has not previously been adapted to this estimation problem.

To do so consider a time-independent targeted variable $x^* = I\{x \in (x_0, x_1)\}$ for some fixed x_0 and x_1 values, when I again denotes an indicator function, and suppose that

$$\lambda(t; x, q, v) = \lambda_0(t) \exp\{\beta_1 x^* + \beta_2' v\}.$$

Under a rare outcome specification, the induced hazard rate given observable variates is to a good approximation

$$\lambda(t; q, v) = \tilde{\lambda}_0(t) \exp\{\beta_1 E(x^*; q, v) + \beta_2' v\}.$$

Under the multivariate normality assumptions of the previous section, x given (q, v) is normally distributed with mean that can be estimated by regressing biomarker values w on q and v, and with variance that can be estimated using repeat biomarker determinations in a biomarker substudy. From these estimators, one can compute a corresponding estimator of the expectation of x^* given q and v by integrating this estimated normal density from x_0 to x_1. Simultaneous calibrated hazard ratio estimators can be calculated by corresponding integration over the elements of a partition formed by quantile cutpoints of this same estimated normal distribution for x^*.

To test this approach, we simulated data from a hazard rate model

$$\lambda(t; x, q, v) = \lambda_0(t) \exp\{\beta_1 x_1^* + \beta_2 x_2^* + \beta_3' v\},$$

Table 1. Simulation[a] summary statistics for regression calibration estimates of tertile hazard ratios.

		Hazard ratio regression coefficients		
Estimation	Statistic	$\beta_1 = 0.405$	$\beta_2 = 0.811$	$\beta_3 = 0$
RC1[b]	Sample mean	0.416	0.805	0
	Sample standard deviation	0.217	0.111	0.031
	95% CI coverage	96.6	94.9	95.0
True[b]	Sample mean	0.406	0.813	0
	Sample standard deviation	0.072	0.069	0.028
	95% CI coverage	95.1	95.5	95.0
Naive[b]	Sample mean	0.259	0.510	−0.081
	Sample standard deviation	0.074	0.072	0.03
	95% CI coverage	49.3	1.3	22.7
RC2[b]	Sample mean	0.281	0.553	0
	Sample standard deviation	0.071	0.075	0.03
	95% CI coverage	62.4	9.1	94.8

[a]Simulation based on 5000 cohorts each of size 2000, with an external biomarker subsample of size 500 in which both biomarker (w) and self-report (q) are measured along with a 20% random subsample in which a second biomarker measurement (w) is available.

[b] RC1 is proposed regression calibration procedure; True is from Cox regression using x-value without measurement error; Naive is based on tertiles for measured q-values; and RC2 arises from forming tertiles of calibrated x-values.

where x_1^* and x_2^* are indicators corresponding to the second and third tertiles of x, which followed a standard normal distribution. Also the univariate covariate v was taken to be independent of x and to adhere to a standard normal distribution, while sampling errors e and ε were also normally distributed with mean zero and variance 0.5, and were independent of each other and of the other modelled variates, while the measured exposure q derived from $q = 0.8x + 0.5v + \varepsilon$. Also, terminal censoring was imposed at a fixed value c. Data were generated from a cohorts of size 2000 with (q, v) measurements, along with an external biomarker sample of size 500 with both w and q values available and a 20% reliability subsample with a second w value having measurement error that is independent of the first.

Multiple simulation scenarios were considered, each giving very similar results. Table 1 shows summary statistics from 5000 generated cohort samples with $\lambda_0(t) \equiv 0.7, \beta_1 = \log(1.5), \beta_2 = 2\log(1.5)$ and $c = 1$, giving a censoring probability of about 35%. Even though one does not expect a rare disease approximation to be accurate with censoring rates as low as 35% the calibrated hazard ratio estimators (RC1) for the second and third tertiles show very little bias relative to their generating values. Sample standard deviation estimates and coverage rates for estimated 95% confidence intervals are also shown, the latter being close to nominal values. Also shown in Table 1 are corresponding summary statistics (i) if one had available the actual generated x-values and used these in standard Cox regression (true); (ii) if one used tertiles from the measured q-values in Cox regression (naive) and (iii) if one used tertiles from the calibrated X (RC2). Clearly the naive and RC2 'estimators' do not perform adequately in this simulation setting.

Our proposed tertile hazard ratio estimators (RC1) seem eminently usable though, of course, they incorporate considerable additional random variation, compared to analyses based on true x-values, as is to be expected with this amount of measurement error contamination.

5. Example of sodium intake and cardiovascular disease risk

To further illustrate the importance of needed hazard ratio estimation developments, consider the association between dietary sodium and cardiovascular disease risks. Even though a high intake of sodium, or a high intake ratio of sodium to potassium, is associated with elevated blood pressure in observational studies and randomised trials (Stamler et al., 1988; Tzoulaki et al., 2012; Whelton et al., 1997), evidence for these dietary associations with cardiovascular diseases has been inconclusive (Bibbins-Domingo et al., 2010; Strazzullo, D'Elia, Kandala, & Cappuccio, 2009; Yang et al., 2011) in spite of considerable public health interest and importance (Mozaffarian et al., 2014; Oria, Yaktine, & Strom, 2013). Uncertainty concerning these associations was enhanced when the large international Prospective Urban Rural Epidemiology (PURE) reported a J-shaped relationship between sodium excretion and major cardiovascular disease outcomes, with higher disease risk at intakes that were relatively low as well as relatively high (O'Donnell et al., 2014) with risk elevations at the low end at values well below recommended maximal intakes (US Department of Health and Human Services, 2015). This led to questions concerning the wisdom of sodium reduction as an isolated public health recommendation (Oparil, 2014).

While most reports of sodium intake in relation to chronic disease outcomes have relied on dietary self-report, the PURE study can be commended for using a biomarker assessment of sodium intake. Specifically morning spot urine sodium excretion was adjusted using a formula (Kawasaki, Itoh, Uezono, & Sasaki, 1993) to provide an estimate of 24-hour urinary excretion. However, in other studies spot urine excretion has been found to not correlate well with 24-hour sodium excretion (Cogswell et al., 2013; Ji, Miller, Venezia, Strazzullo & Cappuccio, 2014; Ji et al., 2012), implying that even if the adjusted intake estimates adhere to (1) the error variance may be quite large relative to the variance for the targeted intake Z. This suggests that the spot urine derived intake estimates may be inefficient, at best, as a biomarker of usual daily sodium intake. Even sodium excretion from 24-hour urine specimens is somewhat noisy as a usual intake biomarker, with average excretion over multiple days

able to usefully reduce the measurement error variance in (1).

The PURE study authors presented associations between estimated usual sodium intake and cardiovascular disease hazard ratios by fitting a cubic spline model in (2) without making any provision for measurement error in their sodium intake estimates. Methods for fitting this type of model while allowing measurement error in (1) to constitute a major fraction of the biomarker variations are needed to interpret, and to correct, the PURE study associations for measurement error.

Recently the authors have used the regression calibration procedure described above with 24-hour sodium excretion as a biomarker, in conjunction with food frequency estimated sodium intake and a range of study subject characteristics to develop calibration equations to estimate short-term sodium intake (Huang et al., 2014). These developments used data from a biomarker substudy of the Women's Health Initiative (WHI) cohorts (Prentice et al., 2011). The calibration equations were used to produce usual daily intake estimates for individuals in WHI cohorts of postmenopausal women in the United States. Calibrated estimates of log-sodium intake were then associated with hazard ratios over cohort follow-up for various cardiovascular disease outcomes. Positive associations were found between calibrated sodium intake, and calibrated ratios of sodium to potassium intake, with major cardiovascular diseases, including coronary heart disease, and heart failure (Prentice et al., 2017). In contrast to the PURE Study, these analyses do not suggest higher risk for these major cardiovascular disease outcomes at relatively low sodium intakes, but a careful study of hazard ratio function shape, while allowing for measurement error in intake estimates would require the ability to fit hazard ratio models more general than the linear model in log-intake applied in these analyses.

In some applications, an additive model of the form (1) may be plausible, but the classical measurement model assumption may not hold because of dependence of the variance of the error term e on the value of the targeted nutritional variable x. If the error variance is large compared to the variance of x, then even modest dependencies of the error variance on x could have important implications of the estimated shape of the hazard ratio function, especially if complex hazard ratio dependencies, such as cubic spline models, are entertained. Hence, additional statistical methods and theory development are strongly needed for this important public health question to be addressed using dietary biomarker and self-report data. Such developments are needed not only for full-cohort data analyses but also for the major cohort subsampling designs, including nested case–control and case–cohort samplings.

In summary, even though sodium overconsumption is projected to be responsible for very substantial morbidity and mortality worldwide (Mozaffarian et al., 2014), issues related to sodium intake assessment have prevented definitive quantitative results from emerging on the associations between sodium intake over the lifespan and the incidence and mortality of specific cardiovascular diseases. The further development of statistical methods and theory is a crucial component of related needed research.

6. Summary and conclusion

There have been many important statistical developments over the past 15–20 years as reliable, high-dimensional genotype data on individual study subjects came available. During the same time period, high-dimensional data on gene, protein and metabolite expression profiles, using blood and urine specimens, as well as high-dimensional data from various types of imaging techniques, have been ascertained in a variety of contexts. These latter data types typically target quantities that vary over the lifespan of the study subject, and the ability of assessment platforms to be comprehensive in terms of analytes measured, may be a challenge (e.g., mass spectrometry-based proteomic or metabolomic platforms).

In public health contexts, gene, protein and metabolite profiles may reflect both genotype and prior exposure history, including such exposures as diet and physical activity patterns over the preceding months or years. If these exposure patterns could be well measured by self-report, then the high-dimensional data just mentioned could be used to explain biological pathways and processes whereby these commonplace activities affect chronic disease risk. However, after several decades of development and application of self-report data for these exposures it is evident that they are not sufficiently reliable for many nutritional epidemiology purposes, most notably for the study of associations with total energy intake, or with the absolute intake of the components of energy.

To the extent that measures in urine and blood, including metabolomic platform measurements, directly reflect dietary intake patterns, these measures may be able to provide an objective assessment of the intake of food and nutrients over the recent past. Repeat application of such objective assessments over cohort follow-up periods may then allow an objective dietary exposure histories to be developed with enhancement of the reliability of related nutritional epidemiology association analyses.

While this biomarker approach to nutritional epidemiology study has considerable potential, there is a need for an intensive research effort to develop biomarkers for many additional nutritional variables,

and an equal need to develop flexible statistical measurement error methods for applying such objective exposure assessments. The latter need arises because the biomarker strategy may be able to yield objective exposure assessments, but these assessments are likely to incorporate noise components that cannot be ignored in analyses to relate dietary exposures to chronic disease risk.

This article is written with a goal of enlisting additional strong statistical methodologists and theorists in this important public health research.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This manuscript was written with partial support from National Institutes of Health grants R01 CA210921, R01 CA119171 and P30 CA015704.

Notes on contributors

Ross L. Prentice is a Member and Former Director of the Public Health Sciences Division of the Fred Hutchinson Cancer Research Center and is Professor of Biostatistics at the University of Washington.

Ying Huang is Associate Member in Biostatistics at the Public Health Sciences and Vaccine & Infectious Diseases Divisions of the Fred Hutchinson Cancer Research Center, and Affiliate Associate Professor in the Department of Biostatistics at the University of Washington.

References

Armstrong, B., & Doll, R. (1975). Environmental factors and cancer incidence and mortality in different countries, with special reference to dietary practices. *International Journal of Cancer*, *15*(4), 617–631.

Bibbins-Domingo, K., Chertow, G. M., Coxson, P. G., Moran, A., Lightwood, J. M., Pletcher, M. J., & Goldman, L. (2010). Projected effect of dietary salt reductions on future cardiovascular disease. *New England Journal of Medicine*, *362*(7), 590–599.

Bingham, S. A. (2003). Urine nitrogen as a biomarker for the validation of dietary protein intake. *The Journal of Nutrition*, *133*(3), 921S–924S.

Breslow, N., McNeney, B., & Wellner, J. A. (2003). Large sample theory for semiparametric regression models with two-phase, outcome dependent sampling. *The Annals of Statistics*, *31*(4), 1110–1139.

Breslow, N., & Wellner, J. A. (2007). Weighted likelihood for semiparametric models and two-phase stratified samples, with application to cox regression. *Scandinavian Journal of Statistics*, *34*(1), 86–102.

Carroll, R. J., Ruppert, D., Stefanski, L. A., & Crainiceanu, C. M. (2006). *Measurement error in nonlinear models: A modern perspective* (2nd ed.). Boca Raton, FL: Chapman & Hall/CRC.

Cogswell, M. E., Wang, C. Y., Chen, T. C., Pfeiffer, C. M., Elliott, P., Gillespie, C. D., … Loria, C. M. (2013). Validity of predictive equations for 24-h urinary sodium excretion in adults aged 18–39 y. *The American Journal of Clinical Nutrition*, *98*(6), 1502–1513.

Cox, D. R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society. Series B (Methodological)*, *34*(2), 187–220. Retrieved from http://www.jstor.org/stable/2985181.

Cox, D. R. (1975). Partial likelihood. *Biometrika*, *62*(2), 269–276.

Forman, D., Bray, F., Brewster, D. H., Gombe Mbalawa, C., Kohler, B., Piñeros, M., … Ferlay, J. (Eds.). (2014). Age-standardized and cumulative incidence rates and standard errors. In *Cancer incidence in Five Continents Volume X*. IARC Scientific Publication No. 164 (pp. 917–1252). Lyon: International Agency for Research on Cancer.

Hastie, T., Tibshirani, R., & Wainwright, M. (2015). *Statistical learning with sparsity: The Lasso and generalizations (No. 143)*. Boca Raton: Chapman and Hall, CRC Press.

Huang, Y., Van Horn, L., Tinker, L. F., Neuhouser, M. L., Carbone, L., Mossavar-Rahmani, Y., … Prentice, R. L. (2014). Measurement error corrected sodium and potassium intake estimation using 24-hour urinary excretion. *Hypertension*, *63*(2), 238–244.

Ji, C., Miller, M. A., Venezia, A., Strazzullo, P., & Cappuccio, F. (2014). Comparisons of spot vs 24-h urine samples for estimating population salt intake: Validation study in two independent samples of adults in Britain and Italy. *Nutrition, Metabolism and Cardiovascular Diseases*, *24*(2), 140–147.

Ji, C., Sykes, L., Paul, C., Dary, O., Legetic, B., Campbell, N. R., & Cappuccio, F. P. (2012). Systematic review of studies comparing 24-hour and spot urine collections for estimating population salt intake. *Revista Panamericana de Salud Pública*, *32*(4), 307–315.

Kalbfleisch, J. D., & Prentice, R. L. (2002). *The statistical analysis of failure time data* (2nd ed.). New York: Wiley and Sons.

Kawasaki, T., Itoh, K., Uezono, K., & Sasaki, H. (1993). A simple method for estimating 24 h urinary sodium and potassium excretion from second morning voiding urine specimen in adults. *Clinical and Experimental Pharmacology and Physiology*, *20*(1), 7–14.

Lampe, J. W., Huang, Y., Neuhouser, M. L., Tinker, L. F., Song, X., Schoeller, D. A., … Prentice, R. L. (2017). Dietary biomarker evaluation in a controlled feeding study in women from the women's health initiative cohort. *The American Journal of Clinical Nutrition*, *105*(2), 466–475.

Luft, F., Fineberg, N., & Sloan, R. (1982). Estimating dietary sodium intake in individuals receiving a randomly fluctuating intake. *Hypertension*, *4*(6), 805–808.

Mozaffarian, D., Fahimi, S., Singh, G. M., Micha, R., Khatibzadeh, S., Engell, R. E., & Powles, J. (2014). Global sodium consumption and death from cardiovascular causes. *New England Journal of Medicine*, *371*(7), 624–634.

O'Donnell, M., Mente, A., Rangarajan, S., McQueen, M. J., Wang, X., Liu, L., … Yusuf, S. (2014). Urinary sodium and potassium excretion, mortality, and cardiovascular events. *New England Journal of Medicine*, *371*, 612–623.

Oparil, S. (2014). Low sodium intake-cardiovascular health benefit or risk? *New England Journal of Medicine*, *371*(7), 677–679.

Oria, M., Yaktine, A. L., & Strom, B. L. (2013). *Sodium intake in populations: Assessment of evidence*. Washington, DC: National Academies Press.

Prentice, R. L. (1982). Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika*, *69*(2), 331–342.

Prentice, R. L. (1986). A case–cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika*, *73*(1), 1–11.

Prentice, R. L., & Breslow, N. E. (1978). Retrospective studies and failure time models. *Biometrika*, *65*(1), 153–158.

Prentice, R. L., Huang, Y., Neuhouser, M. L., Manson, J. E., Mossavar-Rahmani, Y., Thomas, F., … Van Horn, L. (2017). Biomarker calibrated sodium and potassium intake and cardiovascular disease risk among postmenopausal women. *American Journal of Epidemiology*, *186*(9), 1035–1043.

Prentice, R. L., Mossavar-Rahmani, Y., Huang, Y., Van Horn, L., Beresford, S. A., Caan, B., … Neuhouser, M. L. (2011). Evaluation and comparison of food records, recalls, and frequencies for energy and protein assessment by using recovery biomarkers. *American Journal of Epidemiology*, *174*(5), 591–603.

Prentice, R. L., & Sheppard, L. (1990). Dietary fat and cancer: Consistency of the epidemiologic data, and disease prevention that may follow from a practical reduction in fat consumption. *Cancer Causes and Control*, *1*(1), 81–97.

Rakova, N., Jüttner, K., Dahlmann, A., Schröder, A., Linz, P., Kopp, C., … Titze, J. (2013). Long-term space flight simulation reveals infradian rhythmicity in human Na+ balance. *Cell Metabolism*, *17*(1), 125–131.

Schoeller, D. A. (1999). Recent advances from application of doubly labeled water to measurement of human energy expenditure. *The Journal of Nutrition*, *129*(10), 1765–1768.

Self, S. G., & Prentice, R. L. (1988). Asymptotic distribution theory and efficiency results for case–cohort studies. *The Annals of Statistics*, *16*(1), 64–81.

Song, X., Huang, Y., Neuhouser, M. L., Tinker, L. F., Vitolins, M. Z., Prentice, R. L., & Lampe, J. W. (2017). Dietary long-chain fatty acids and carbohydrate biomarker evaluation in a controlled feeding study in participants from the women's health initiative cohort. *The American Journal of Clinical Nutrition*, *105*(6), 1272–1282.

Stamler, J., Rose, G., Stamler, R., Elliott, P., Marmot, M., Pyorala, K., … Sans, S. (1988). INTERSALT: an international study of electrolyte excretion and blood pressure. Results for 24 hour urinary sodium and potassium excretion.. *British Medical Journal*, *297*(6644), 319–328.

Strazzullo, P., D'Elia, L., Kandala, N. B., & Cappuccio, F. P. (2009). Salt intake, stroke, and cardiovascular disease: Meta-analysis of prospective studies. *British Medical Journal*, *339*, b4567.

Thomas, D. C. (1977). Addendum to 'Methods for cohort analysis: Appraisal by application of asbestos mining' by F. D. K. Liddell, J. C. McDonald and D. C. Thomas. *Journal of the Royal Statistical Society A*, *140*, 469–491.

Tzoulaki, I., Patel, C. J., Okamura, T., Chan, Q., Brown, I. J., Miura, K., … Elliott, P. (2012). A nutrient-wide association study on blood pressure. *Circulation*, *126*(21), 2456–2464.

US Department of Health and Human Services, et al. (2015). *2015–2020 dietary guidelines for Americans*. Washington, DC: USDA.

Whelton, P. K., He, J., Cutler, J. A., Brancati, F. L., Appel, L. J., Follmann, D., & Klag, M. J. (1997). Effects of oral potassium on blood pressure: Meta-analysis of randomized controlled clinical trials. *Journal of the American Medical Association*, *277*(20), 1624–1632.

World Cancer Research Fund and American Institute for Cancer Research. (1997). Food, nutrition and the prevention of cancer: A global perspective (Tech. Rep.). Washington, DC.

World Cancer Research Fund and American Institute for Cancer Research. (2007). Food, nutrition and the prevention of cancer: A global perspective (Tech. Rep.). Washington, DC.

World Health Organization. (2003). Diet, nutrition and the prevention of chronic diseases: Report of a joint WHO/FAO expert consultation (Tech. Rep. No. 916). Geneva.

Yang, Q., Liu, T., Kuklina, E. V., Flanders, W. D., Hong, Y., Gillespie, C., … Hu, F. B. (2011). Sodium and potassium intake and mortality among US adults: Prospective data from the Third National Health and Nutrition Examination Survey. *Archives of Internal Medicine*, *171*(13), 1183–1191.