Preventive Medicine xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Preventive Medicine

journal homepage: www.elsevier.com/locate/ypmed

Challenges in risk estimation using routinely collected clinical data: The example of estimating cervical cancer risks from electronic health-records

Rebecca Landy[a,*], Li C. Cheung[b], Mark Schiffman[b], Julia C. Gage[b], Noorie Hyun[b], Nicolas Wentzensen[b], Walter K. Kinney[c,1], Philip E. Castle[d], Barbara Fetterman[e], Nancy E. Poitras[e], Thomas Lorey[e], Peter D. Sasieni[a], Hormuzd A. Katki[b]

[a] *Centre for Cancer Prevention, Wolfson Institute of Preventive Medicine, Barts and the London School of Medicine and Dentistry, Queen Mary, University of London, Charterhouse Square, London EC1M 6BQ, UK*

[b] *Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, DHHS, Bethesda, MD, USA*

[c] *Division of Gynecologic Oncology, Kaiser Permanente Medical Care Program, Oakland, CA, USA*

[d] *Albert Einstein College of Medicine, Bronx, NY, USA*

[e] *Regional Laboratory, Kaiser Permanente Northern California, Berkeley, CA, USA*

ARTICLE INFO

Keywords:
Risk estimation
Screening
Statistical methods
Epidemiology
Electronic health-records
Cervix

ABSTRACT

Electronic health-records (EHR) are increasingly used by epidemiologists studying disease following surveillance testing to provide evidence for screening intervals and referral guidelines. Although cost-effective, undiagnosed prevalent disease and interval censoring (in which asymptomatic disease is only observed at the time of testing) raise substantial analytic issues when estimating risk that cannot be addressed using Kaplan-Meier methods. Based on our experience analysing EHR from cervical cancer screening, we previously proposed the logistic-Weibull model to address these issues. Here we demonstrate how the choice of statistical method can impact risk estimates. We use observed data on 41,067 women in the cervical cancer screening program at Kaiser Permanente Northern California, 2003–2013, as well as simulations to evaluate the ability of different methods (Kaplan-Meier, Turnbull, Weibull and logistic-Weibull) to accurately estimate risk within a screening program. Cumulative risk estimates from the statistical methods varied considerably, with the largest differences occurring for prevalent disease risk when baseline disease ascertainment was random but incomplete. Kaplan-Meier underestimated risk at earlier times and overestimated risk at later times in the presence of interval censoring or undiagnosed prevalent disease. Turnbull performed well, though was inefficient and not smooth. The logistic-Weibull model performed well, except when event times didn't follow a Weibull distribution. We have demonstrated that methods for right-censored data, such as Kaplan-Meier, result in biased estimates of disease risks when applied to interval-censored data, such as screening programs using EHR data. The logistic-Weibull model is attractive, but the model fit must be checked against Turnbull non-parametric risk estimates.

1. Introduction

Large-scale epidemiologic research is shifting from formally designed trials and observational studies to increasing use of electronic health-records (EHR) that contain longitudinal data for millions of people in routine clinical practice. In particular, EHR facilitate studies of screen-detected disease and precursors in population screening programs, where risks of disease are typically low and thus very large studies are required. The longitudinal nature of these records is of particular importance when estimating risk over time, for example to provide evidence on which to base screening intervals.

Three key features of electronic health record data should be accounted for when disease is asymptomatic and only detected through screening or clinical testing. First, asymptomatic disease onset is only known to occur between the last time considered 'disease-free' and the time of diagnosis; this is called interval censoring (Huang and Wellner, 1997). Unlike designed observational studies, where researchers control when participants return, patients are encouraged to comply with doctors' suggestions, but can return at any time for unknown reasons, and thus the interval censoring is irregular and may be informative (i.e., linked to risk of disease). Second, disease may be present at first screen (prevalent disease), called left-censoring. Third, prevalent disease is not

[*] Corresponding author.
E-mail address: R.Landy@qmul.ac.uk (R. Landy).
[1] Retired.

https://doi.org/10.1016/j.ypmed.2017.12.004
Received 27 July 2017; Received in revised form 27 November 2017; Accepted 4 December 2017

always immediately diagnosed. In particular, people with negative screening tests generally do not undergo definitive disease ascertainment (known as verification bias (Schmidt and Factor, 2013)), and in the real world many individuals with a positive screen will not have a diagnostic test, or postpone such testing. Consequently, when prevalent disease was not ascertained but disease is diagnosed at future screens, some of that ‘incident’ disease is actually undiagnosed prevalent disease. In our experience, designed observational studies either exclude prevalent disease (and even incident disease diagnosed too close to baseline), or assume all detected disease is incident; however prevalent disease is often relevant, not merely a nuisance factor to be removed. These three issues also apply to designed observational studies of screening, if baseline disease ascertainment is not performed on everyone (or at least a representative sample) or if there was substantial variation in visit times between participants. As risk estimates from routine clinical practice are crucial to inform screening guidelines, risk must be estimated accurately using appropriate methodology.

We critically examine the performance of different methods for the analysis of EHR to estimate risk, providing examples to show the magnitude of bias when using off-the-shelf methods in realistic settings. The standard Kaplan-Meier method cannot account for the above three features, yet is commonly used. We have proposed a new model, the logistic-Weibull model, (Cheung et al., 2017) that does account for the above three issues. We have also proposed a modification of the non-parametric Turnbull method (Cheung et al., 2017) to check the fit of the logistic-Weibull model. We use simulations to examine critically the assumptions underlying all methods, and sensitivity to these assumptions, to recommend a robust methodology and practical advice for epidemiologists calculating risk from EHR. In addition to the simulations, we provide an example using observed cervical screening data.

2. Cervical cancer and cervical screening

Cervical cancer originates from a persistent high-risk human papilloma virus (HPV) infection, which may progress to asymptomatic precancer, occult cancer and symptomatic invasive cancer. There is also natural regression within this process, when the body naturally clears/suppresses HPV or even apparent precancer, immunologically. Cervical screening has been highly successful at preventing cervical cancer where good programs with wide coverage exist, through detection and removal of precancers. There are currently two tests which are widely used in cervical screening programs. Traditionally cytology was used, where the cells are examined and abnormal cells are identified. More recently HPV testing was introduced, which tests for the presence of carcinogenic HPV. HPV testing has higher sensitivity to detect precancerous disease than cytology, but is less specific (Cuzick et al., 2006). Depending on the result of the tests, women are invited back for their next test at a routine screening interval, invited back earlier for more intensive surveillance, or referred for magnified examination (colposcopy), biopsy and possible treatment. Disease ascertainment only occurs at colposcopy. Cervical cancer is an ideal disease for a screening program because detectable and highly treatable asymptomatic pre-cancerous lesions can be targeted, and their rate of growth and invasion is typically slow. As there are harms associated with screening as well as benefits, it is important that screening intervals are an appropriate length; short enough not to miss the passage through precancer to cancer, but long enough that the risk of detecting precancer is not negligible. It is therefore useful to know how long after a given screening result precancer first becomes detectable, to determine appropriate screening intervals. It is also important to estimate prevalent risk accurately, as this informs whether to intervene when the initial test results are known. Data from screening programs are routinely collected, containing the date and result of each screening test.

3. Methods for estimating risk

There are three main classes of models available to analyse data: non-parametric, semi-parametric and parametric models. Statistical models are simplifications of reality, embodying numerous assumptions. Non-parametric models for a distribution function make no distributional assumptions, unlike parametric models, which specify the distribution in terms of parameter(s). Semi-parametric models have both parametric and non-parametric components. More details on the methods described below are available in Supplementary material 1.

The *Kaplan-Meier method* (Kaplan and Meier, 1958) which is non-parametric, is the most widely used method of estimating risk of precancerous cervical disease as a function of time from an ‘entry’ screen (Khan et al., 2005; Ronco et al., 2014; Dillner et al., 2008; Cuzick et al., 2008; Nobbenhuis et al., 1999). However, Kaplan-Meier is not appropriate when disease is screen-detected. The most popular adaptation of the Kaplan-Meier estimator equates the time of onset with time of diagnosis, which consistently overestimates the time of onset. To improve this, the midpoint of the interval in which disease could have occurred can be imputed as the time of onset for Kaplan-Meier (Nobbenhuis et al., 1999). This too causes bias unless the screening interval is homogeneous and short (Law and Brookmeyer, 1992).

The Kaplan-Meier estimator only correctly estimates time of screen-detected disease diagnosis (not time of disease onset) at times when all participants have their disease status ascertained. However, risk of diagnosis for asymptomatic conditions is largely determined by the testing schedule, limiting consideration to the chosen screening intervals underlying the data. Consideration of disease onset is required to calculate risks for any possible screening interval, and requires consideration of prevalent and incident disease.

The *Turnbull estimator* (Turnbull, 1976) is non-parametric, and appropriate for prevalent (including undiagnosed) and incident disease. We have adapted Turnbull to account for undiagnosed baseline disease (Cheung et al., 2017) (see Supplementary material 1). However, the adapted Turnbull method cannot account for covariates and can result in survival curves with big steps and wide confidence intervals even in a dataset of 1 million women.

To account for covariates and improve statistical efficiency, we recently developed the *logistic-Weibull model*, a parametric model that jointly models prevalent and incident disease (Cheung et al., 2017). The absence/presence of prevalent disease is modelled using logistic regression, and cumulative risk of incident disease among women who did not have disease diagnosed at baseline is modelled using a Weibull survival model. Weibull models account for interval censoring, and are reasonable when follow-up times are short relative to a woman's typical time to event (Katki et al., 2013a). The cumulative risk is a weighted sum of prevalent logistic-regression disease risk $\pi_i(\beta)$ and incident Weibull-model disease risk $(1 - S_i(t;(\gamma,\tau)))$ at time t:

$$F_i(t;\beta,\gamma,\tau) = \pi_i(\beta) + (1 - \pi_i(\beta))(1 - S_i(t;(\gamma,\tau))),$$

where $\pi_i(\beta) = \frac{\exp(X_i^T\beta)}{1+\exp(X_i^T\beta)}$ and $S_i(t;(\gamma,\tau)) = \exp\left\{-\left(\frac{t}{\exp(Z_i^T\gamma)}\right)^{\frac{1}{\tau}}\right\}$, X_i and Z_i are vectors of covariates, β are regression coefficients for prevalent disease effects, γ are regression coefficients for incident disease effects, and τ governs the shape of the Weibull distribution. Because undiagnosed baseline disease can be considered as “missing data”, we developed an EM algorithm to estimate model parameters (Cheung et al., 2017). In the special case of no undiagnosed baseline disease (i.e. everyone at baseline undergoes definitive disease ascertainment), the logistic regression and Weibull regression can be conducted separately to obtain parameter estimates. Logistic-Weibull models were recently used by Katki et al. (2013b, 2013c, 2013d) to estimate cervical cancer and pre-cancer risks among 1 million women undergoing cervical screening to inform U.S. risk-based screening guidelines for cervical cancer (Massad et al., 2013).