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

Rebecca Landy1,*, Li C. Cheung2, Mark Schiffman2, Julia C. Gage2, Noorie Hyun2, Nicolas Wentzensen2, Walter K. Kinney3, Philip E. Castle4, Barbara Fetterman5, Nancy E. Poitras5, Thomas Lorey5, Peter D. Sasieni1, and Hormuzd A. Katki2

1Centre 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

2Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, DHHS, Bethesda, MD, USA

3Division of Gynecologic Oncology, Kaiser Permanente Medical Care Program, Oakland, CA, USA (retired)

4Albert Einstein College of Medicine, Bronx, NY

5Regional Laboratory, Kaiser Permanente Northern California, Berkeley, CA, USA

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 1,037,065 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

*Corresponding author: Rebecca Landy, 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, R.Landy@qmul.ac.uk.

Publisher’s Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Conflicts of interest: The authors declare there is no conflict of interest.
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.

Keywords
Risk estimation; screening; statistical methods; epidemiology; electronic health-records; cervix

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. 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 always immediately diagnosed. In particular, people with negative screening tests generally do not undergo definitive disease ascertainment (known as verification bias), 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, that does account for the above three issues. We have also proposed a modification of the non-parametric Turnbull method 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.

**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 programmes with wide coverage exist, through detection and removal of precancers. There are currently two tests which are widely used in cervical screening programmes. 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 pre-cancerous disease than cytology, but is less specific. 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 programme 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 programmes are routinely collected, containing the date and result of each screening test.

**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, 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. 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. This too causes bias unless the screening interval is homogeneous and short.

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 is non-parametric, and appropriate for prevalent (including undiagnosed) and incident disease. We have adapted Turnbull to account for undiagnosed baseline disease (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. 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. 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 \( S_i(t; (\gamma, \tau)) = \exp\left(-\frac{t}{\exp(Z_i^T \gamma)}\right)^\frac{1}{\tau} \). \( X_i \) and \( Z_i \) are vectors of covariates, \( \beta \) are regression coefficients for prevalent disease effects, \( \gamma \) are regression coefficients for incident disease effects, and \( \tau \) 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. 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 (15–17) 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.
Both Turnbull and the Weibull models account for interval censoring, and are fitted to interval data, the \([L_i, R_i]\) in which disease could have occurred, for all women who were not known to have prevalent disease. As the software for fitting models to interval censored data cannot handle intervals of width 0, for Turnbull we define prevalent disease to occur in the interval \([0,0.01 \text{ years})\), to avoid overlapping with incident disease. In the logistic-Weibull model the intervals for participants in whom baseline disease status was not ascertained, but disease was present in follow-up, is set to begin at exactly 0, whereas incident disease intervals that begin at zero are pushed to 0.01 years.

The Weibull model smooths over flats and sudden peaks which could result from Kaplan-Meier or Turnbull models, producing estimates at all time points. The Weibull model was chosen based on the Armitage and Doll model of carcinogenesis,\(^{(19)}\) which hypothesised that the accumulation of a few critical steps in pathogenesis (e.g., mutations) was required to transform a normal cell to a cancerous cell. We have recently generalized logistic-Weibull to a semiparametric logistic-Cox model that makes fewer distributional assumptions, and a weakly-parametric model using integrated B-splines to model incident disease. This model can be useful when no parametric assumptions are a good fit, but a smooth risk estimate is preferred.\(^{(20)}\)

While a testing schedule with shorter screening intervals will result in more accurate estimation, the results of Turnbull and logistic-Weibull analyses are not biased by the testing schedule. Both Turnbull and logistic-Weibull assume baseline disease ascertainment occurs at random, though for logistic-Weibull this can be at random conditional on covariates which can be adjusted for, such as prior screening history or age. We note that whilst these methods are generalizable to other screen-detected diseases, Weibull may not always be an appropriate parametric model.

We compare risk estimates from Kaplan-Meier, Kaplan-Meier using the interval midpoint as date of diagnosis, adapted Turnbull and Weibull/logistic-Weibull, to demonstrate the impact the choice of statistical method can have on the results.

**Cervical Screening Data**

Data were available on 1,037,065 women aged 30–64 who had at least one cervical screening co-test (HPV test and cervical cytology) between 1\(^{st}\) January 2003 and 30\(^{th}\) June 2013 at Kaiser Permanente Northern California (KPNC). We estimate risk of precancer or worse (CIN2+, including cervical intraepithelial neoplasia grade 2, 3, adenocarcinoma in situ, or cancer) following an HPV-positive, cytology-negative co-test using data from the 41,067 women who had this screening result. This is a relatively common, non-normal screening result, the management of which remains controversial.

To investigate the impact of the choice of interval in which disease is defined to have occurred, we reanalyse these data using three different definitions for intervals, defined in Supplementary Material 2. Different interval definitions allow analyses to account for test insensitivity. As this example uses observed data, the true risk is unknown.
Simulations

We carried out a series of simulations to examine critically the properties of Kaplan-Meier using the diagnosed time and the interval midpoint as the event time, adapted Turnbull and Weibull/logistic-Weibull models. Unless otherwise specified, screening was simulated to occur independently of disease risk at intervals drawn from an exponential distribution with mean 3.5 years, variance 9 years, up to a maximum of 10 years of follow up. The screening test was presumed to always detect disease present at the time of testing.

In simulation 1, event times were simulated assuming a Weibull distribution (Figure 1a), with disease present in 1.8% by 3 years and 7.0% by 6 years (corresponding to shape parameter 2 and log-scale parameter 3.1) with no prevalent disease at time 0. In total, 1000 datasets were simulated from this distribution with population sizes of (i) 250, (ii) 2500 and (iii) 10,000 people.

In simulation 2, we assumed a prevalent disease risk of 3%, and thereafter event times were simulated under the Weibull distribution used in simulation 1. Baseline disease status was ascertained in 20% of the population selected at random. 1000 datasets of size 10,000 were simulated from this distribution. We then repeated the 1000 simulations, allowing for non-random baseline disease ascertainment: 52% of the population with prevalent disease had their baseline disease status ascertained, as did 19% of the population without prevalent disease, corresponding to 20% of the population overall.

In simulation 3, event times were simulated under a gamma distribution (Figure 1b), with disease present in 2.1% by 2 years and 35.2% by 4 years (corresponding to shape parameter 8 and scale parameter 0.6). 1000 datasets of size 10,000 were simulated.

Simulating data from known distributions allowed the true risk to be known, and compared to the estimated risks. We calculated the sum of the squared difference between the estimates from each model and the true cumulative risk calculated each day up to 10 years to compare the goodness-of-fit between the models.

Results

When analysing the observed KPNC cervical screening data, we see that the risk estimates vary considerably over time between the four methods of analysis (Figure 2). Naïve Kaplan-Meier provided the lowest risk estimates at earlier times, and the highest estimates at later times. The largest difference in risk estimates is at baseline, where Kaplan-Meier estimates are close to zero due to low levels of prevalent disease ascertainment, whereas Turnbull and logistic-Weibull models provide higher estimates of prevalent (including undiagnosed) disease. Other than for prevalent disease, Kaplan-Meier using the midpoint provides similar estimates to the Turnbull and logistic-Weibull models in this example.

In the simulations, we can compare the risks estimated by each method to the true risk. For simulation 1, where event times were simulated from a Weibull distribution with no prevalent disease, we show the risk in a single simulation with sample sizes 250, 2,500 and 10,000, and the mean risk over 1,000 simulations with sample size 2,500 for each of the four
methods considered, (Figure 3). Kaplan-Meier significantly underestimates risk until around 8.5 years, after which it overestimates risk. Although Kaplan-Meier using the interval midpoint provides reasonable risk estimates at some time points, it is inaccurate at other points (years 5–7). In general, it is impossible to know when these estimates will be accurate. The Turnbull estimates are generally reasonable; although the mean over 1000 simulations yields unbiased risks (Figure 3d), the risks in any given analysis (Figures 3a, 3b, 3c) remain choppy as additional jump points accrue slowly as the sample size increases. Increasing sample size reduces the choppiness of Kaplan-Meier, Kaplan-Meier using the interval midpoint and Turnbull estimates. In contrast to Kaplan-Meier and Turnbull, the Weibull model provides smooth estimates of risk at all times. The sum of squared errors (a measure of goodness of fit, where low values indicate better fit) was lowest for Weibull (sample size 10,000: 9.1×10⁻⁶), then Turnbull (0.059).

Using data from the same simulation, we present the 90% empirical confidence intervals for the Turnbull and Weibull models, as well as the true risk, in Figure 4. The confidence intervals are wider for the Turnbull than Weibull model, especially for smaller sample sizes. In our simulations a sample of approximately 20,000 gave confidence intervals the same width as from a Weibull model using a sample of 2,500, i.e. the sample required was eight times bigger (not shown).

Although Kaplan-Meier can be adapted to handle disease diagnosed at baseline, it cannot be adapted to handle undiagnosed prevalent disease, which results from incomplete disease ascertainment at baseline. Using simulation 2, where 3% of the sample had prevalent disease, with 20% random baseline disease ascertainment, Figure 5 shows that Kaplan-Meier underestimated baseline disease risk by 80%, as disease status was ascertained in only 20% of the sample in this simulation. The (unadapted) Turnbull model also underestimated prevalent disease risk. The pure Weibull model does not allow prevalent disease, resulting in a poor fit to the data. In contrast, the adapted Turnbull and logistic-Weibull provide accurate estimates of prevalent disease risk. Table 1 shows the prevalent disease estimates for sample sizes 250, 2500 and 10,000, for random and non-random baseline disease ascertainment. The Kaplan-Meier estimates are as expected (0.52*3% + 0.19*0%), estimating the risk as the amount of disease that was detected at baseline. Although the Turnbull and logistic-Weibull estimates were quite biased with a sample size of 250, in different directions, they were less biased than the Kaplan-Meier estimates. Turnbull provided the least biased results with the larger sample sizes.

In Figure 6, where data were simulated from a gamma distribution in simulation 3, the naïve Kaplan-Meier estimates are much lower than the true risks, Kaplan-Meier using the midpoint estimates are high, but getting closer to the truth from 4 years. The Weibull model produces reasonable estimates at most time points, but from years 1.5–3 the estimates are poor; for example at 2 years the true risk is 2.1%, but the Weibull model estimate is 4.6%. The Weibull estimates are also diverging from the truth from 4 years. The Turnbull estimates are unbiased.

Using the observed KPNC cervical screening data, we estimated disease risks using three sets of rules to define the start and end of the interval in which disease occurred. Table 2
demonstrates that, although there are differences in risk estimates between the risk-estimation methods, the results are very similar (the maximum absolute difference is 2%) within each method for the 3 interval definitions. In this example, 30% of the intervals are affected by the choice of when to start and end the interval. The biggest difference occurs in the prevalent disease risk. More details on the results in Table 2 can be found in Supplementary Material 2.

**Discussion**

It is well known that the choice of analysis model affects the validity of the results. For electronic health-records of screen-detected disease, where the data are interval-censored, standard methods for right-censored data, such as naïve Kaplan-Meier, are not appropriate, as they produce biased results. The Turnbull estimate is unbiased, though is choppy and has wide confidence intervals, whereas parametric estimators are smooth and require much smaller sample sizes to produce the same width confidence intervals. It is important to handle prevalent disease appropriately, particularly when using parametric models, and to check the parametric logistic-Weibull model versus the non-parametric adapted Turnbull, as the underlying distribution of the data is rarely known, and the Weibull model may not provide a good fit to the data.

Kaplan-Meier methods may be appropriate in the absence of undiagnosed baseline disease, and when the intervals between screens are short or everyone returns at regular intervals. Such conditions might hold for designed studies, but are unlikely to hold for EHR data from routine clinical practice.

The risk of detectable prevalent disease is of interest as it determines whether to refer patients immediately for further (potentially invasive) tests. Outside of trials, it is difficult to know what proportion of women with a negative test actually have disease at baseline. This is because the women with negative tests who are referred for further tests are not a random sample of the women with negative tests; rather they are the women believed to be at higher risk, for example due to symptoms. This is also true for the situation modelled here, of risk following an HPV-positive cytology-negative screening result, where guidelines suggest women are invited for a repeat test at 12 months. Unlike Kaplan-Meier, both the adapted Turnbull (14) and logistic-Weibull models allow some disease found in follow-up to be undiagnosed prevalent disease, though both assume baseline disease ascertainment occurs at random (missing completely at random in Turnbull; missing at random, given covariates which can be adjusted for, in logistic-Weibull), which may not be true. When baseline disease ascertainment depends on baseline disease status, this is a form of informative censoring, which none of these methods are designed to handle; therefore the prevalent disease risks were overestimated by both Turnbull and logistic-Weibull for the larger sample sizes, as women at higher risk were preferentially sampled. Actual referral data is required to validate whether estimated prevalent disease risks are correct; additionally, if we are to refer immediately for uncomfortable, risky, and/or costly tests it is especially important for prevalent disease that the risk estimate is for diagnosable disease. If prevalent disease status is never verified then the estimated prevalent disease risk is purely model driven. Therefore
prevalent disease risk is best estimated when prevalent disease status is verified for everyone, or at least a representative subsample of participants.

Although the logistic-Weibull makes assumptions, it has advantages over the non-parametric Turnbull estimates. Logistic-Weibull accounts for covariates, and allows differences in risk between groups to be tested. The Turnbull estimates can be choppy and estimates may not be generated for all time points, which occurs even in our dataset of 1 million women. Cumulative risk curves can be flat for a length of time before implausibly suddenly increasing. On the contrary, logistic-Weibull estimates are smooth and defined at all times. As screening guidelines are based on risk of disease at a given time point, Turnbull estimates may be less reliable if there are jumps in the estimated risk around the time of interest.

Parametric assumptions can shrink confidence intervals, particularly with smaller sample sizes, but we reiterate the importance of checking distributional assumptions and the parametric estimates against the Turnbull estimates. In Figure 6, the Weibull model provided a good estimate at some time points, but did not provide a good fit overall. In principle any parametric model can be fit, but the fit should be checked. We have developed two more flexible models; first the logistic-Cox model, which relaxes Weibull assumptions to require only proportional hazards for incident disease. The second is a weakly parametric model using integrated B splines, which is smooth and robust to different distributions, though has some loss of power as more parameters must be estimated. However, these models are computationally intensive and thus we promote parametric models, such as the logistic-Weibull, for routine use as long as the fit is checked. Our package for fitting the logistic-Weibull, logistic-Cox and weakly-parametric models, as well as adapted Turnbull, in R, PIMixture, is available at https://dceg.cancer.gov/tools/analysis/PIMixture.

Although the interval definitions for when disease could have occurred are subjective, they can incorporate scientific knowledge about the disease and screening tests. Wider intervals in which disease could occur imply less efficiency and larger confidence intervals around risk estimates, but are appropriate if tests are known to lack sensitivity or specificity. For different choices of reasonable intervals in which disease could occur, we find similar results for each analysis method. Future work could consider how best to use information from individuals with positive screening results, but no confirmed diagnosis. These individuals are at higher risk of disease, but are currently right-censored identically to individuals censored following a negative screening test. Reasonable intervals in which disease could occur should reflect a scientific understanding of disease pathogenesis and test quality, to narrow down when disease may have occurred/could occur.

It is important to consider the difference between undetectable disease that exists, detectable disease, and clinically relevant disease. Detectable disease that is not clinically relevant, although of potential etiologic interest, is not usually ascertained in routine clinical practice. Finally, undetectable latent disease is conceptually important, but difficult to account for.

When estimating cumulative risks over time from electronic health record data, it is important to consider whether interval censoring and the possibility of undiagnosed
prevalent disease are present. If so, appropriate statistical methods should be applied, such as logistic-Weibull. However the model fit must be checked against Turnbull non-parametric risk estimates.

**Supplementary Material**

Refer to Web version on PubMed Central for supplementary material.

**Acknowledgments**

Many thanks to Peter Windridge and Adam Brentnall for their assistance. This research was conducted, in part, by the Intramural Research Program of the NCI/NIH/DHHS.

**Funding:** This work was supported by a Union for International Cancer Control Yamagiwa-Yoshida travel grant [grant number 2015/YY1/329257 to RL]; Cancer Research UK [grant number A16892 to PS]; and the Intramural Research Program of the National Cancer Institute, National Institutes of Health, Department of Health and Human Services, United States.

**References**

1. Huang, J., Wellner, JA., editors. Proceedings of the First Seattle Symposium in Biostatistics. Springer; 1997. Interval censored survival data: a review of recent progress.

2. Schmidt RL, Factor RE. Understanding sources of bias in diagnostic accuracy studies. Archives of pathology & laboratory medicine. 2013; 137(4):558–65. [PubMed: 23544945]

3. Cheung LC, Pan Q, Hyun N, Schiffman M, Fetterman B, Castle PE, et al. Mixture models for undiagnosed prevalent disease and interval-censored incident disease: applications to a cohort assembled from electronic health records. Statistics in Medicine. 2017

4. Cuzick J, Clavel C, Petry KU, Meijer CJ, Hoyer H, Ratnam S, et al. Overview of the European and North American studies on HPV testing in primary cervical cancer screening. International Journal of Cancer. 2006; 119(5):1095–101. [PubMed: 16586444]

5. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. Journal of the American statistical association. 1958; 53(282):457–81.

6. Khan MJ, Castle PE, Lorincz AT, Wacholder S, Sherman M, Scott DR, et al. The elevated 10-year risk of cervical precancer and cancer in women with human papillomavirus (HPV) type 16 or 18 and the possible utility of type-specific HPV testing in clinical practice. Journal of the National Cancer Institute. 2005; 97(14):1072–9. [PubMed: 16030305]

7. Ronco G, Dillner J, Elfström KM, Tunesi S, Snijders PJ, Arbyn M, et al. Efficacy of HPV-based screening for prevention of invasive cervical cancer: follow-up of four European randomised controlled trials. The lancet. 2014; 383(9916):524–32.

8. Dillner J, Rebolj M, Birembaut P, Petry K-U, Szarewski A, Munk C, et al. Long term predictive values of cytology and human papillomavirus testing in cervical cancer screening: joint European cohort study. Bmj. 2008; 337

9. Cuzick J, Szarewski A, Mesher D, Cadman L, Austin J, Perryman K, et al. Long-term follow-up of cervical abnormalities among women screened by HPV testing and cytology.—Results from the Hammersmith study. International Journal of Cancer. 2008; 122(10):2294–300. [PubMed: 18240149]

10. Nobbenhuis MA, Walboomers JM, Helmerhorst TJ, Rozendaal L, Remmink AJ, Risse EK, et al. Relation of human papilloma virus status to cervical lesions and consequences for cervical-cancer screening: a prospective study. The Lancet. 1999; 354(9172):20–5.

11. Law CG, Brookmeyer R. Effects of mid-point imputation on the analysis of doubly censored data. Statistics in medicine. 1992; 11(12):1569–78. [PubMed: 1439361]

12. Turnbull BW. The empirical distribution function with arbitrarily grouped, censored and truncated data. Journal of the Royal Statistical Society Series B (Methodological). 1976:290–5.
13. Katki HA, Schiffman M, Castle PE, Fetterman B, Poitras NE, Lorey T, et al. Benchmarking CIN3+ risk as the basis for incorporating HPV and Pap cotesting into cervical screening and management guidelines. Journal of lower genital tract disease. 2013; 17(5 0 1):S28. [PubMed: 23519302]

14. Cheung LC, Pan Q, Hyun N, Schiffman M, Fetterman B, Castle PE, et al. Mixture models for leftcensored and irregularly interval-censored data: Applications to a cancer screening cohort assembled from electronic health-records. Statistics in medicine. (in press).

15. Katki HA, Schiffman M, Castle PE, Fetterman B, Poitras NE, Lorey T, et al. Five-year risk of CIN3+ and cervical cancer for women with HPV testing of ASC-US Pap results. Journal of lower genital tract disease. 2013; 17(5 0 1):S36. [PubMed: 23519303]

16. Katki HA, Schiffman M, Castle PE, Fetterman B, Poitras NE, Lorey T, et al. Five-year risk of CIN3+ and cervical cancer for women who test Pap-negative, but are HPV-positive. Journal of lower genital tract disease. 2013; 17(5 0 1):S56. [PubMed: 23519306]

17. Katki HA, Schiffman M, Castle PE, Fetterman B, Poitras NE, Lorey T, et al. Five-year risks of CIN2+ and CIN3+ for women with HPV-positive and HPV-negative LSIL Pap results. Journal of lower genital tract disease. 2013; 17(5 0 1):S43. [PubMed: 23519304]

18. Massad LS, Einstein MH, Huh WK, Katki HA, Kinney WK, Schiffman M, et al. 2012 updated consensus guidelines for the management of abnormal cervical cancer screening tests and cancer precursors. Obstetrics & Gynecology. 2013; 121(4):829–46. [PubMed: 23635684]

19. Armitage P, Doll R. The age distribution of cancer and a multi-stage theory of carcinogenesis. British journal of cancer. 1954; 8(1):1. [PubMed: 13172380]

20. Hyun N, Cheung L, Pan Q, Schiffman M, Katki H. Flexible risk prediction models for left or interval-censored data from electronic health records. Annals of Applied Statistics. 2017; 11(2): 1063–84.
Highlights

- Onset of screened disease is unobserved and known only to occur between screen visits
- Additionally, disease may be prevalent, but is not always diagnosed immediately
- Kaplan-Meier risk estimates from screening data are biased
- Logistic-Weibull and Turnbull models can produce unbiased estimates
- Logistic-Weibull estimates are smoother with smaller confidence intervals
Figure 1.
Cumulative risks over time under a) Weibull b) Weibull with 3% prevalent disease and c) gamma distributions used in the simulation analyses
Figure 2.
Risk estimates of CIN2+ following a positive HPV test and negative cytology test among women in the cervical cancer screening program at Kaiser Permanente Northern California, 2003–2013, using naïve Kaplan-Meier, Kaplan-Meier using the midpoint, Turnbull and logistic-Weibull models.
Figure 3.
Risk estimates using naïve Kaplan-Meier, Kaplan-Meier using the midpoint, Turnbull and Weibull models, and the true risks, from one simulation, with event times simulated from a Weibull distribution: a) sample size 250, from one simulation, b) sample size 2,500, from one simulation c) sample size 10,000, from one simulation d) sample size 2,500, mean risk estimates from 1,000 simulations
Note that the Turnbull, Weibull and truth lines overlie one another.
Figure 4.
90% intervals of risk estimates using Turnbull and Weibull models, and the true risks, from 1,000 simulations, with event times simulated from a Weibull distribution with sample sizes 250 and 10,000.
Figure 5.
Mean risk estimates using naïve Kaplan-Meier, Kaplan-Meier using the midpoint, Turnbull, adapted Turnbull, Weibull and logistic-Weibull models, and the true risks, from 1,000 simulations, sample size 10,000, with event times simulated from a Weibull distribution, and 3% prevalent disease.
Note that the adapted Turnbull, logistic-Weibull and truth lines overlie one another.
Figure 6.
Mean risk estimates using naïve Kaplan-Meier, Kaplan-Meier using the midpoint, Turnbull and Weibull models, and the true risks, from 1,000 simulations, sample size 10,000, with event times simulated from a gamma distribution.
Note that the Turnbull and truth lines overlie one another.
Table 1

Mean prevalent disease estimates from 1,000 simulations for random and non-random baseline disease ascertainment*, when event times were simulated from a Weibull distribution with 3% prevalent disease.

|                          | random baseline disease ascertainment | non-random baseline disease ascertainment |
|--------------------------|--------------------------------------|------------------------------------------|
|                          | Sample size                          | Sample size                              |
|                          | 250        | 2500       | 10,000      | 250         | 2500       | 10,000      |
| Truth                    | 3%         | 3%         | 3%          | 3%          | 3%         | 3%          |
| naïve Kaplan-Meier       | 0.6%       | 0.6%       | 0.6%        | 1.5%        | 1.6%       | 1.6%        |
| Kaplan-Meier using the midpoint | 0.6%       | 0.6%       | 0.6%        | 1.5%        | 1.6%       | 1.6%        |
| Turnbull                 | 2.3%       | 2.9%       | 3.0%        | 3.6%        | 3.7%       | 3.7%        |
| logistic Weibull         | 2.3%       | 2.9%       | 3.0%        | 4.0%        | 3.6%       | 3.7%        |

*random baseline disease ascertainment: 20% of the population had baseline disease status ascertained, regardless of baseline disease status. Non-random baseline disease ascertainment: 52% of the population with baseline disease had baseline disease status ascertained, 19% of the population with without baseline disease had baseline disease status ascertained.
Table 2

Estimated risk of CIN2+ following an HPV-positive, cytology-negative screening result using Kaplan-Meier, Turnbull and logistic-Weibull, among 41,067 women aged 30–64, using three definitions of intervals

| Interval                          | 0 years A | 0 years B | 0 years C | 1 year A | 1 year B | 1 year C | 3 years A | 3 years B | 3 years C | 5 years A | 5 years B | 5 years C | 7 years A | 7 years B | 7 years C |
|----------------------------------|-----------|-----------|-----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|
| Naïve KM cumulative risk, prevalent disease (%) | 0.59      | 0.48      | 0.56      | 0.92     | 1.43     | 0.87     | 8.00     | 8.03     | 7.64     | 12.18    | 12.07    | 11.77    | 15.48    | 15.06    | 14.85    |
| Event at midpoint, KM cumulative risk, prevalent disease (%) | 0.59      | 0.48      | 0.56      | 4.47     | 4.80     | 4.06     | 9.14     | 9.16     | 8.15     | 11.34    | 11.22    | 11.08    | 12.70    | 12.51    | 12.94    |
| Adapted Turnbull cumulative risk estimates | 4.71      | 5.15      | 3.18      | 5.02     | 6.11     | 3.25     | 8.64     | 8.75     | 7.94     | 11.11    | 11.05    | 11.24    | 12.28    | 12.12    | 12.60    |
| Logistic Weibull cumulative risk (%) | 4.47      | 4.79      | 2.94      | 5.91     | 6.13     | 4.45     | 8.34     | 8.40     | 7.47     | 10.56    | 10.48    | 10.42    | 12.64    | 12.44    | 13.30    |

Interval A: Start date: The latest recorded negative co-test, negative (i.e. showing disease less severe than the outcome) biopsy or second consecutive negative cytology with no positive HPV test, end date: the date of the biopsy on which disease was diagnosed. Interval B: Start date: same as interval A, end: the date of the cytology test which referred the woman to the biopsy on which disease was diagnosed, if there was a cytology test less than 12 months before the biopsy; otherwise the date of the biopsy which diagnosed disease was used. Interval C: Start date: the date of cytology prior to the screen which referred the woman to the diagnosing biopsy, end date: same as interval A. KM: Kaplan-Meier