ICSN abstract: Methods for the analysis of screen-detected outcomes using electronic health records

Rebecca Landy[1], Li C. Cheung[2], Mark Schiffman[2], Julia C. Gage[2], Noorie Hyun[2], Nicolas Wentzensen[2], Peter D. Sasieni[1], Hormuzd A. Katki[2]

[1] 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

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

Background: Researchers studying disease following surveillance testing increasingly use electronic health records to evaluate screening intervals and referral guidelines. Whilst this is a cost-effective way to evaluate screening programmes, utilizing Kaplan-Meier methods may raise substantial analytic issues that can bias risk estimates for screen-detected disease . These issues include diagnosed and undiagnosed prevalent disease, and interval censoring in which asymptomatic diseases are only observed at the time of testing. Based on our experience analysing electronic health-records from cervical cancer screening, we previously proposed the logistic-Weibull, a prevalence-incidence model, in order to address these issues. Here we demonstrate how the choice of statistical method can impact risk estimates.

Methods: We use simulations to demonstrate the impact of the choice of statistical method on risk estimation. We evaluate the ability of non-parametric (Kaplan-Meier, Turnbull) and parametric (Weibull and logistic-Weibull) models to address issues common to the analysis of recurrent screening data with an asymptomatic outcome.

Results: Methods taking into account interval censoring, such as Turnbull (with recent developments) and the logistic-Weibull models, can also handle undiagnosed prevalent disease. In simulations, methods appropriate for right-censored data (Kaplan-Meier models) provide biased estimates of risk in the presence of interval censored data. The logistic-Weibull model is more efficient than Turnbull, however as the logistic-Weibull model makes assumptions regarding the distribution of times at which disease becomes diagnosable, it is important the results are visually checked against non-parametric Turnbull risk estimates.

Conclusions: It is important to be aware of the assumptions required by statistical estimators when using electronic health-records to evaluate screening for an asymptomatic disease. These issues and results also apply to a wider range of scenarios. Although the prevalence-incidence models appear useful, many challenges remain to be addressed to unlock the promise of epidemiologic studies of “big data” from electronic health-records.