RISK PREDICTION WITH IMPERFECT OUTCOME INFORMATION FROM EMR

STEPHANIE CHAN and TIANXI CAI
Harvard T.H. Chan School of Public Health, Boston, MA, USA
e-mail: tcai@hsph.harvard.edu

Electronic medical records (EMRs) are a valuable resource for developing “real-world” risk prediction models for a wide range of diseases. These models, once validated, can potentially be incorporated into clinical practice. However, a major challenge for using EMR to build risk models is that the timing of disease onset is not readily available. Extracting clinical event times for patients requires labor intensive medical chart reviews. Additionally, since a significant proportion of clinical events may occur prior to patients’ first EMR encounter or outside of the specific hospital system, the EMR may only capture partial information on the event time. For example, the domain expert would be able to determine whether a patient has experienced a clinical outcome by the end of EMR follow-up, but the exact timing may be unknown even after chart review. The time to first ICD9 billing code for the clinical condition or the first NLP mention of the condition in the notes can serve as a proxy for the true event time, but is subject to measurement error. In this paper, we propose a robust approach to developing a risk prediction model by synthesizing multiple imperfect sources of information on the event time of interest. Treating the observable outcomes as survival time subject to current status censoring and survival time measured with errors, we construct an optimally combined estimator under a flexible semi-parametric transformation model for the survival time given baseline predictors and unspecified measurement errors. Simulation studies demonstrate that the proposed estimator performs well in finite sample. We illustrate the proposed estimator by assessing the effects of genetic markers on coronary artery disease with an EMR study of rheumatoid arthritis patients performed with the Partners HealthCare EMR.