ELSEVIER

CrossMark

Journal of Clinical Epidemiology 76 (2016) 175−182

Journal of Clinical Epidemiology

Adequate sample size for developing prediction models is not simply related to events per variable

Emmanuel O. Ogundimu*, Douglas G. Altman, Gary S. Collins

Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology & Musculoskeletal Diseases, Botnar Research Centre, University of Oxford, Windmill Road, Oxford OX3 7LD, UK

Accepted 29 February 2016; Published online 8 March 2016

Abstract

Objectives: The choice of an adequate sample size for a Cox regression analysis is generally based on the rule of thumb derived from simulation studies of a minimum of 10 events per variable (EPV). One simulation study suggested scenarios in which the 10 EPV rule can be relaxed. The effect of a range of binary predictors with varying prevalence, reflecting clinical practice, has not yet been fully investigated.

Study Design and Setting: We conducted an extended resampling study using a large general-practice data set, comprising over 2 million anonymized patient records, to examine the EPV requirements for prediction models with low-prevalence binary predictors developed using Cox regression. The performance of the models was then evaluated using an independent external validation data set. We investigated both fully specified models and models derived using variable selection.

Results: Our results indicated that an EPV rule of thumb should be data driven and that EPV ≥ 20 generally eliminates bias in regression coefficients when many low-prevalence predictors are included in a Cox model.

Conclusion: Higher EPV is needed when low-prevalence predictors are present in a model to eliminate bias in regression coefficients and improve predictive accuracy.

Keywords: Events per variable; Cox model; External validation; Predictive modeling; Sample size; Resampling study

1. Introduction

When multivariable prediction models are developed, the sample size is often based on the ratio of the number of individuals with the outcome event to the number of candidate predictors (more precisely, the number of parameters), referred to as the events per variable (EPV). Models developed from data sets with too few outcome events relative to the number of candidate predictors are likely to yield biased estimates of regression coefficients. They lead to unstable prediction models that are overfit to the development sample and perform poorly on new data. Simulation studies of prediction models developed using both logistic regression and Cox regression have suggested minimum EPV values of between 5 and 20 for reliable results [1−5]. An EPV of 10 is widely advocated as the rule of thumb for multivariable logistic and Cox regression analyses.

Through their influential work, Peduzzi et al. [1,3,4] encouraged the use of the 10 EPV rules for both logistic- and Cox regression−based prediction models. However, there were limitations to the design of their simulation studies, particularly with respect to prediction. They emphasized accuracy and precision of the regression coefficients, rather than the measures of predictive ability. The studies were also based on a relatively small data set of 673 individuals (252 of whom had the outcome event) and only considered one prediction model that contained seven predictors (six binary and one ordinal). Predictors were not selected, either before or during the model building. Although these highly cited simulation studies have raised awareness of the importance of the number of outcome events relative to the number of predictors, the limited scenarios examined cast doubt on the generalizability of their findings.

Subsequent simulation studies have examined more complex scenarios by altering the number of predictors in

Conflict of interest: None.

Funding: EOO and DGA received funding from an MRC Partnership Grant for the PROGnosis RESearch Strategy (PROGRESS) group (grant reference number: G0902393). GSC and DGA received funding from the Medical Research Council (grant number G1100513).

* Corresponding author. Tel./fax: +44 (0) 1865 223460.

E-mail address: emmanuel.ogundimu@csm.ox.ac.uk (E.O. Ogundimu).

http://dx.doi.org/10.1016/j.jclinepi.2016.02.031

What is new?

Key findings

- The use of a rule of thumb for selecting events per variable (EPV) should be study dependent.
- Convergence in Cox models depends more on the severity of low prevalence in binary predictors and much less on low EPV.
- Higher EPV is needed when low-prevalence predictors are present in a model to eliminate bias in regression coefficients and improve predictive accuracy.

What is the implication and what should change now?

- EPV $\geq$ 20 should be considered when a data set includes low-prevalence binary predictors - if EPV $\geq$ 20 cannot be guaranteed, then the use of the penalized likelihood approach should be considered.

fixed regression models. Some have suggested that the 10 EPV rules can be relaxed [5], and others that no single EPV rule of thumb can guarantee accurate estimates of regression coefficients [6]. However, these studies have also focused on establishing a recommended minimum EPV in the context of stable regression coefficients, without considering the predictive ability of the model. They have also not considered the generalizability of the findings to real-life settings, for example, when investigators are confronted with many candidate predictors and must choose a subset to include in their final prediction model [7].

Studies examining the influence of backward elimination for predictor selection have shown that the regression coefficients from a logistic regression model may have considerable bias, particularly in small samples [8]. Studies examining the effect of EPV on the development of regression models have therefore tended to use small single data sets and have focused on accurate parameter estimation of regression coefficients. They have offered limited insights into the effect on the predictive performance of the model (e.g., calibration and discrimination).

The presence of low-prevalence binary predictors can induce the problem of complete (or quasi) separation in logistic regression [9,10] or monotone likelihood in Cox regression [11]. These problems may be noticed in an individual study when parameters and standard errors are too large to be useful. The parameter estimates are not unique and depend on trivial issues like the settings of software used for the analysis. While keeping other design factors constant, the probability of separation or monotone likelihood is lower with higher EPV values.

Heinze and Schempe [11] extended the modified likelihood method of Firth [12] to circumvent monotone likelihood problems in the estimation of parameters from Cox model with low-prevalence predictors. However, applied researchers still typically do not apply Firth's correction when fitting a Cox regression model. We focused on this practice and investigated the EPV requirement for parameter estimates and predictive accuracy in the presence of low-prevalence but highly prognostic binary predictors.

We conducted a resampling study using a large general practice data set, comprising over 2 million anonymized patient records, to examine the relationship between EPV, accuracy of regression coefficients, and predictive ability using Cox regression. We investigated scenarios with both fully prespecified models and models derived from the data using automated variable selection. We examined the stability and precision of the regression coefficients and their effect on the models' predictive performance (e.g., the *c*-index, *D*-statistic, and R^2). We also examined the effect of EPV in the development of a prediction model on the model's subsequent performance using a separate large external validation data set.

2. Data and methods

2.1. Study data: The Health Improvement Network

The Health Improvement Network (THIN) is a large database of anonymized electronic health care records collected from general-practice clinics around the United Kingdom (England, Scotland, Wales, and Northern Ireland). The THIN database currently contains medical records from approximately 4% of the United Kingdom population. We used clinical information from 2,084,445 individuals, aged 30 to 84 years, registered between June 1994 and June 2008 from 365 general practices. The characteristics of the THIN data set are summarized in Table 1. Twelve variables were considered: one categorical [smoking status (SMK); four categories], four continuous [age, systolic blood pressure (SBP), body mass index (BMI), and ratio of total serum cholesterol to high-density lipoprotein (RATIO)], and seven binary [sex, diagnosis of type diabetes (TYPE2), rheumatoid arthritis (BRA), atrial fibrillation (BAF), renal disease (RENAL), treated hypertension (HYPER), and family history of coronary heart disease (FHCVD)]. Because of the low prevalence of some of the SMK categories, we combined nonsmokers and former smokers as "nonsmokers" and the rest as "smokers." The primary outcome was cardiovascular disease (CVD), which was experienced by 93,564 individuals in the THIN data set.

Prediction models were developed using the entire THIN data set, omitting individuals from Scotland ($THIN_d$). The individuals from Scotland ($THIN_v$) were used to validate the prediction models in an external validation setting. The sample sizes of the development and validation data sets were 1,973,511 individuals (88,312 CVD events) and 110,934 individuals (5,252 CVD events), respectively.