Chapter 6
Preparing Data for Predictive Modelling

Sander M. J. van Kuijk, Frank J. W. M. Dankers, Alberto Traverso, and Leonard Wee

6.1 Introduction

Predictive modelling is aimed at developing tools that can be used for individual prediction of the most likely value of a continuous measure, or the probability of the occurrence (or recurrence) of an event. There has been a huge increase in popularity of developing tools for prediction of outcomes at the level of the individual patient. For instance, a recent review identified a total of 363 articles that described the development of prediction models for the risk of cardiovascular disease in the general population alone [1].

Such models are often developed using regression techniques that yield a prediction model in the form of a regression formula (see Chap. 8). Such formulae are generally impractical to use and are therefore often simplified into a simple risk score that can easily computed by hand, or presented in such a way that calculation is made easier (such as the use of a nomogram for predicting survival in breast cancer patients with brain metastasis [2], see Fig. 6.1), incorporated in a web-based application or perhaps as an application on a smartphone.

S. M. J. van Kuijk, PhD (✉)
Department of Clinical Epidemiology and Medical Technology Assessment, Maastricht University Medical Center, Maastricht, The Netherlands
e-mail: sander.van.kuijk@mumc.nl

F. J. W. M. Dankers, MSc
Department of Radiation Oncology (MAASTRO), GROW School for Oncology and Developmental Biology, Maastricht University Medical Center+, Maastricht, The Netherlands

Department of Radiation Oncology, Radboud University Medical Center, Nijmegen, The Netherlands

A. Traverso, PhD · L. Wee, PhD
School of Oncology and Developmental Biology (GROW), Maastricht University Medical Center, Maastricht, The Netherlands

P. Kubben et al. (eds.), *Fundamentals of Clinical Data Science*,
https://doi.org/10.1007/978-3-319-99713-1_6

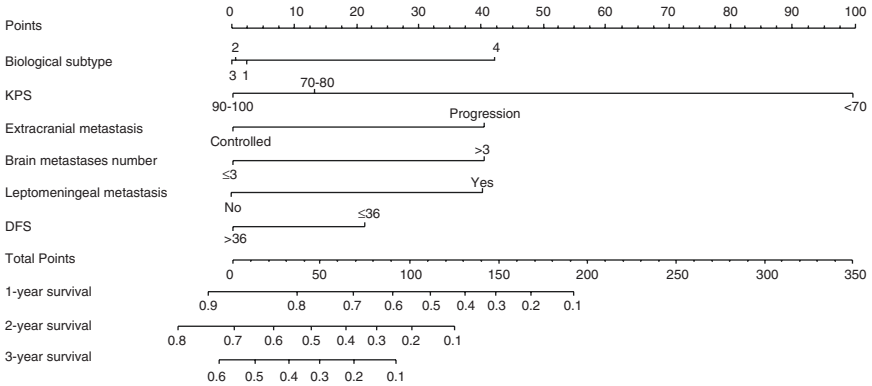

Fig. 6.1 Nomogram for the prediction of overall survival for patients with breast cancer brain metastasis. (Reprinted with permission Huang et al. [2]). Each predictor value corresponds to an amount of points. All points combined (i.e. 'Total points') corresponds to 1, 2, and 3 year survival probability

Two types of prediction tools for binary outcomes can be distinguished: (1) a tool that can be used to predict an individual's probability of the presence of disease *at the moment of prediction* (i.e., a diagnostic prediction model) and (2) one that can be used to predict the probability of the future occurrence of an event (i.e., a prognostic prediction model). An example of the former is a model to estimate the probability of *Chlamydia trachomatis* infection to aid selective screening of youth at high risk of an infection [3]. An example of a prognostic prediction model to estimate an individual's probability of a future event is a model that estimates the probability of a successful vaginal birth after previous caesarean section, which is subsequently included in a decision aid to discuss the intended mode of delivery [4, 5]. Although the application may differ substantially, the methods that are employed to develop such models are similar.

Before any new prediction tool can be developed, patient-level data need to be collected retrospectively or prospectively. Considerations such as choosing the correct study design, determining the necessary sample size for developing a prediction model, transforming variables, and how to deal with incomplete data on potential predictor variables and outcome measures will be covered in this chapter. This chapter does not cover all possible steps that need to be undertaken before a prediction tool can be developed, but focuses on the most important considerations and the most prevalent challenges.

6.2 Study Designs for Prediction Model Development

An important observation to make is that in the development of tools for individual prediction, we are generally not interested in unbiased estimates of causal associations between determinants and the presence of disease or the occurrence of a certain event in the future. In other words, we are not interested to unravel casual associations between predictors and the outcome. We are occupied with selecting

the best set of predictors and include those in a model in such a way that the predictions that the model makes are as accurate as possible. Epidemiological phenomena such as confounding (i.e., bias is introduced in the estimation of coefficients because of a variable associated with both the predictor and the outcome, but is not controlled for) and mediation (i.e., the presence of an intermediate variable that explains the association between the predictor and the outcome) are not relevant in the context of prediction modelling. Interaction terms, which are variables that moderate the association between a predictor and the outcome, can be useful to increase the predictive performance of a model if associations between predictor variables and the outcome differ between subgroups, but are not used to aid causal interpretation. Hence, the estimated regression coefficients that are used for predictions for future patients may not reflect true causal associations but do lead to the best predictions. This is especially true for prediction models for recurrent events, as selecting only participants that experienced a first occurrence may introduce a phenomenon known as index-event bias [6, 7]. This has no effect on the performance of prediction models for future patients as the coefficients are estimated for the purpose of generating predictions, not for aetiological purposes. That being said, models that include predictor variables that show associations that are contradictory to expectations may lack face validity and their introduction in daily clinical practice may be hampered.

6.2.1 Retrospective and Prospective Data

The ideal study design for developing a prognostic prediction model is the prospective cohort study. This way, candidate predictors that are not part of routine clinical care can be added to the patient work up. Additionally, the quality of data collection is in the hands of the researcher, and can be controlled during the course of the study. The retrospective cohort design, efficient as the use of readily available data may be, is often hampered by the fact that some candidate predictors are unmeasured as they are not part of routine clinical care or because the data were collected previously for other purposes than developing a prediction model. As a result, missing data can pose a serious problem in retrospective data. Although valid methods exist to handle missing data, prevention is preferred.

Naturally, when the prediction model is diagnostic in nature as opposed to prognostic (i.e., to predict a state that is already present or absent), a cross-sectional design may suffice. In such a design, both the candidate predictors and the outcome are measured in one go. For diagnostic prediction models, the outcome is often a disease status, confirmed by a gold standard.

6.2.2 Alternative Study Designs

An alternative to the cohort study is making use of data of a randomized controlled trial (RCT). Such a prediction model may serve to identify those patients that have the highest probability of responding to the intervention of interest, or to predict the

probability of experiencing an adverse event, but the data could also be used for predicting other types of events. The benefit of using RCT data is that these data are often of high quality as an RCT is designed to minimize the proportion of missing data and minimize measurement error. Nonetheless, data from an RCT are not without challenges. Often, strict eligibility criteria result in a homogeneous sample hampering generalizability to the population the prediction model will be applied to in the future. For example, many RCT's exclude patients with comorbidities. These comorbidities may be very important prognostic factors that are best included as predictors in the prediction model. Another drawback may be that outcome measures in an RCT may be measured too close in time to the baseline measurement for prediction to be of interest.

Another alternative design is the case-control design. In a case-control design, for each patient who experienced the event (a case), a control patient (or more than one) is recruited for the study. Often, researchers use matching techniques to force the control group to be roughly similar to the group of cases. In case matching has been performed, the distribution of candidate predictors has changed to such an extent that it is unlikely that a useful prediction model can be derived from the data. However, if no matching has been performed, case-control data can be used to develop a prediction model. Regression coefficients (to compute predicted probabilities for future patients) and odds ratios (to express the strength of the association) can be estimated validly as if it were a cohort study. But there remains one major problem associated with case-control data. The prevalence of the event (i.e., the proportion of cases) is defined *by design*. In a case-control study with a 1-1 ratio (i.e., a single control for each case), the prevalence is 50%. As case-control studies are usually performed for rare events, this prevalence may be completely different from the prevalence in the population of patients the model needs to provide predictions for. In this case, the predicted probability is likely to be severely overestimated for future patients. This can be prevented by adjusting the model intercept (i.e., the constant in a logistic regression model) so that the average predicted probability in the data used to train the model is similar to the prevalence of the event in the population of patients the model will be used. This could be done iteratively until similarity is reached, or estimated by including the linear predictor of the model (see Chap. 8) as an offset in a regression model without predictors. If the goal is not providing individual estimates of the probability of an event, but merely to stratify patients into risk-based groups, the actual intercept is of less concern.

6.2.3 Patient Selection

Patients or subjects that are included in the study should reflect the population the model will be applied to in the future, and they should be at risk to develop the outcome of interest. Preferably, the sample is heterogeneous, including a wide range of values on the predictors.

6.3 Sample Size Considerations

6.3.1 Potential Predictor Variables and Model Overfitting

In most cases, the primary aim of predictive modelling is not null-hypothesis testing but determining the structure of a prediction model and estimating indicators of predictive performance (see Chap. 8). As a result, sample size formulas that include the statistical power (say, 80 or 90%) and the type-I error rate alpha (usually 5%) for null-hypothesis testing are generally not applicable. However, there is a limit as to how many candidate predictor variables can be included in the modelling phase. A model that consists of too many predictors is more likely to be overfitted (i.e., the model performs well on the data used to develop or train the model, but performs poor on new patients). One characteristic of the poor external performance of an overfitted model is that it produces too extreme predictions for future patients. Thus, predictions for future patients who are at low risk of the outcome are on average too low, and predictions for patients at high risk of the outcome are on average too high. This can easily be seen in the calibration plot (see Chap. 10). The slope of the calibration plot of a well-calibrated model is close to 1 indicating perfect agreement between predicted probabilities and actual outcomes, but the slope is less than 1 for models that are overfit.

6.3.2 Sample Size Rules-of-thumb

A simulation study has examined the ratio between the number of events that need to be included in the study, and the number of candidate predictor variables that can validly be entered in the modelling step when using logistic regression [8]. They concluded that no major problems occurred for 10 events per variable or more. Note that an event is defined as the outcome that is least prevalent. E.g., if the majority of patients experience the event of interest, the number of patients who do not experience the event determine the minimum sample size (or the maximum number of candidate predictor variables if the sample size is fixed). For example, consider designing a study to develop a prediction model to estimate the probability of lymph node metastases in patients with non-small cell lung cancer. From previous experience you estimate that the outcome will be experienced in 1 in 6 (or in about 17%), and you plan to include 6 predictor variables in the modelling step. According to the rule of thumb, 60 events need to be observed in the data. Hence, 60/0.17 = 353 patients need to be recruited for the study.

Similar rules of thumb exist for different regression models. For the Cox proportional hazards regression model it is suggested to include at least 10 failures for each candidate predictor [9, 10], and for the linear regression model at least 2–10 patients for each candidate predictor [11, 12]. However, there is no guarantee that overfitting does not occur when abiding by these rules of thumb. Other factors

may influence the ratio between candidate predictors and the number of events, such as the frequency of a binary predictor that is relatively rare. For models that include binary predictors that are rare, it is suggested to include at least 20 events per variable [13].

6.4 Pre-processing Your Data

The first step after collecting data is checking for inconsistencies and impossible values in the data. On the patient level, variables that are dependent on each other may be checked several ways. For instance, by computing the difference between systolic and diastolic blood pressure, the differences can be checked with a histogram to rule out impossible values (e.g., values indicating higher diastolic blood pressure). On the variable level, computing ranges provides a first check of whether values beyond an acceptable range were entered in the data. Examine outliers and determine per outlier if this is likely due to an error in the data collection, or whether the outlier represents the true value of the patient. In the latter case, the value(s) should not be removed from the dataset before modelling.

6.4.1 Transforming Predictor Variables

Regression models that are employed to develop prediction models explicitly assume additivity and linearity of the associations between the predictors and the outcome (in linear regression), between the predictors and the log odds of the outcome (in logistic regression), or between the predictors and the log hazard or log cumulative hazard (in Cox proportional hazards regression). The linearity assumption implies that the slope of the regression line (or the estimated coefficient) is the same value over the whole range of the predictor, and the additivity assumption implies that effects of different predictor variables on the outcome are not dependent on the value of other predictors. Regression methods do not place assumptions on the distribution of the predictor variables, but severely skewed continuous variables (e.g, circulating levels of biomarkers) often perform better after transformation to a roughly normal distribution. A frequent transformation of right-skewed predictors that consist of only positive values is taking the natural logarithm. This compresses the long right tail and expands the short left tail. In addition to taking the logarithm of a predictor, other mathematical transformations may be performed as well (e.g., taking the square root). A drawback of including transformed predictors in the model is interpreting the effect of those predictors on the original scale.

There are other methods to account for non-linear associations between the predictor and the outcome, but those are strictly part of the regression modelling phase and do not fall within the scope of preparing data for predictive modelling. Examples of such methods include polynomial regression and spline regression.

6.4.2 Categorizing Predictor Variables

If transforming does not yield the desired effect, or if easy interpretation of coefficients is necessary, continuous predictor variables may be categorized into two or more categories. Keep in mind that when the assumptions of additivity and linearity are met, categorization is likely to result in a decrease of predictive performance compared to using the continuous predictor. Categorizing causes a loss of information and statistical power, but also underestimates the extent of variation in risk [14]. Categorization can be performed using data-driven cut-off values after visualization of the association between the determinant and the outcome, or using well-established cut-off values. For example, evidence suggests that the association between body mass index (BMI) and mortality is U-shaped [15–17]. In this case, choosing cut-off values that are commonly accepted (e.g., below 18.5 kg/m^2 to define underweight and above 25 kg/m^2 to define overweight) may not result in the best performing categories on the data used for development compared to data-driven determination of cut-off values, but it aids interpretation and practical implementation. Bear in mind that the number of categories that are made not only depends on the best fit of the predictor during the modelling phase, but also on the amount of predictors that can be studied using the sample at hand (see sample size considerations). A categorical variable with *n* categories results in the inclusion of *n-1* dummy variables.

6.4.3 Visualizing Data

Associations between continuous predictor variables and the outcome (or log odds etc. of the outcome) can be visualized to check if non-linearity exists and if so, if there are clear indications for certain transformations, polynomials, or categorization. For a continuous outcome, a simple plot can be made consisting of the predictor on the x-axis and the outcome variable on the y-axis with a smooth local regression curve (or LOESS curve) to provide a visual representation of the association. For binary outcomes, graphing the association becomes more tedious as the outcome variable consists only of zeroes and ones. A simple solution is to make groups based on quartiles of the predictor variable, and plot the average of the predictor values against the average of the outcome parameter.

6.5 Missing Data

6.5.1 Why You Should Bother About Missing Data

Most statistical and machine learning packages will omit patients that have one or more missing values on the variables that are used to develop the model. This results in less statistical precision in estimating regression coefficients and other statistics

of interest, reflected by larger standard errors, wider confidence intervals and thus p-values that are less likely to be lower than the alpha that is chosen for testing. Such complete case analysis or listwise deletion not only decreases the sample size, but may also introduce bias if the incomplete patients are not a random sample of all patients recruited for the study. The patients in the sample that are completely observed do not reflect the population of interest anymore. This mechanism that underlies the process of missing values is important for deciding how to handle missing data. Methods such as complete case analysis and proper imputation methods all have assumptions with respect to the mechanism that caused missing data.

When the incomplete patients are a random sample of the complete patients, or in other words when the probability of values to be missing is unrelated to any patient characteristic or response, the missing data are said to be missing completely at random (MCAR). Complete case analysis will provide unbiased estimates, but with less precision compared to a situation where all data are observed. When the probability of values to be missing is associated with the values of other, observed, patient characteristics or responses, the missing data are missing at random (MAR). For instance, if older male patients are less inclined to complete a questionnaire on socio-economic status, but both sex and age are recorded in the dataset. A third mechanism that can be identified is called missing not at random (MNAR). In this case, the probability of values to be missing is associated with the value of the variable itself (such as when a ceiling effect is present), or when the probability is associated with the value of other, unobserved, covariates.

Most methods to handle missing data assume that data are MCAR or MAR. However, there are no methods to discriminate between mechanisms using the data that were collected. Therefore, it is important to think thoroughly about the missing data problem and judge if MCAR or MAR is a likely explanation of the missing data. This makes transparent communication on the missing data problem in a manuscript very important. Sterne et al. have suggested guidelines for reporting analyses that are potentially affected by missing data [18]. Applied to prediction modelling research, the researcher should report the number of missing values per predictor variable and outcome variable, give reasons for missing data if these are known, compute difference in characteristics between patients that are completely observed and patients who are incomplete, and describe the method that was used to account for missing data, including a description of the assumptions that were made.

6.5.2 Handling Missing Data

To prevent a decrease in precision and a high likelihood of biased regression coefficients, missing data can be imputed. Imputing is the replacing of the empty cells in the dataset with actual values. The goal of imputation is not adding new information to the dataset, but to allow all other observations of incomplete patients to be used for the subsequent analysis.

There are numerous methods that can be used to impute missing data. A simple method to impute a continuous variable is to compute the mean of that variable

using data of patients that have an observed value of this variable, and replace every missing data point with this mean value. Simple as it is, imputation with the mean decreases the variance within a variable and distorts the association between the imputed variable and other covariates in the data. Proper imputation methods produce a synthetic part of the data that, when analysed, do not introduce bias in the estimation of regression coefficients (given certain assumptions, usually that data are MAR), and gives a correct estimate of uncertainty, reflected in confidence intervals of parameters estimated in the study.

A very popular imputation method, and for good reasons, is multiple imputation. In multiple imputation, the incomplete variables are imputed using regression models based on other covariates that are used to estimate a likely value for each of the incomplete patients. However, not the estimated value is imputed, but the estimated value to which a random error term (which can be positive or negative) is added to preserve the variance in the dataset. This is performed multiple times so that the analyst ends up with more than 1 imputed dataset. Because of the randomness associated with the error term that is added to the imputation, imputations differ between the imputed datasets. Analyses are performed on each of the imputed datasets, and regression coefficients are averaged to produce a pooled estimate, and the variance is computed using a combination of the *within*-dataset variance and the *between*-dataset variance. This way, the uncertainty introduced by having to impute the data is correctly accounted for. This method of producing pooled estimates after multiple imputation is called Rubin's Rules [19]. Although multiple imputation works well when the MAR assumption is met, it is likely to introduce bias in case the assumption is violated [20, 21]. In case data are known to be MNAR, the analyst needs to specifically define the mechanism that caused missing data to produce unbiased estimates. However, the alternative to imputing data (i.e., complete case analysis) assumes data are MCAR, which may be unrealistic for many incomplete medical datasets.

References

1. Damen JA, Hooft L, Schuit E, Debray TP, Collins GS, Tzoulaki I, et al. Prediction models for cardiovascular disease risk in the general population: systematic review. BMJ (Clin Res Ed). 2016;353:i2416.
2. Huang Z, Sun B, Wu S, Meng X, Cong Y, Shen G, et al. A nomogram for predicting survival in patients with breast cancer brain metastasis. Oncol Lett. 2018;15(5):7090–6.
3. van Klaveren D, Gotz HM, Op de Coul EL, Steyerberg EW, Vergouwe Y. Prediction of chlamydia trachomatis infection to facilitate selective screening on population and individual level: a cross-sectional study of a population-based screening programme. Sex Transm Infect. 2016;92(6):433–40.
4. Schoorel EN, van Kuijk SM, Melman S, Nijhuis JG, Smits LJ, Aardenburg R, et al. Vaginal birth after a caesarean section: the development of a Western European population-based prediction model for deliveries at term. BJOG. 2014;121(2):194–201; discussion
5. Schoorel EN, Vankan E, Scheepers HC, Augustijn BC, Dirksen CD, de Koning M, et al. Involving women in personalised decision-making on mode of delivery after caesarean section: the development and pilot testing of a patient decision aid. BJOG. 2014;121(2):202–9.

6. Sep SJ, van Kuijk SM, Smits LJ. Index event bias: problems with eliminating the paradox. J Stroke Cerebrovasc Dis. 2014;23(9):2464.
7. Smits LJ, van Kuijk SM, Leffers P, Peeters LL, Prins MH, Sep SJ. Index event bias-a numerical example. J Clin Epidemiol. 2013;66(2):192–6.
8. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol. 1996;49(12):1373–9.
9. Concato J, Peduzzi P, Holford TR, Feinstein AR. Importance of events per independent variable in proportional hazards analysis. I. Background, goals, and general strategy. J Clin Epidemiol. 1995;48(12):1495–501.
10. Peduzzi P, Concato J, Feinstein AR, Holford TR. Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates. J Clin Epidemiol. 1995;48(12):1503–10.
11. Austin PC, Steyerberg EW. The number of subjects per variable required in linear regression analyses. J Clin Epidemiol. 2015;68(6):627–36.
12. Harrell FE Jr. Regression modeling strategies. New York: Springer-Verlag; 2001.
13. Ogundimu EO, Altman DG, Collins GS. Adequate sample size for developing prediction models is not simply related to events per variable. J Clin Epidemiol. 2016;76:175–82.
14. Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. Stat Med. 2006;25(1):127–41.
15. Whitlock G, Lewington S, Sherliker P, Clarke R, Emberson J, Halsey J, et al. Body-mass index and cause-specific mortality in 900 000 adults: collaborative analyses of 57 prospective studies. Lancet. 2009;373(9669):1083–96. London
16. Zheng W, McLerran DF, Rolland B, Zhang X, Inoue M, Matsuo K, et al. Association between body-mass index and risk of death in more than 1 million Asians. N Engl J Med. 2011;364(8):719–29.
17. Berrington de Gonzalez A, Hartge P, Cerhan JR, Flint AJ, Hannan L, MacInnis RJ, et al. Body-mass index and mortality among 1.46 million white adults. N Engl J Med. 2010;363(23):2211–9.
18. Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ. 2009;338:b2393.
19. Rubin DB. Multiple Imputation for Nonresponse in Surveys. New York: John Wiley and Sons; 2004.
20. van Kuijk S, Viechtbauer W, Peeters L, Smits L. Bias in regression coefficient estimates when assumptions for handling missing data are violated: a simulation study. Epidemiol Biostat Public Health. 2016;13(1):1–8.
21. White IR, Carlin JB. Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. Stat Med. 2010;29(28):2920–31.