1.2.2 *Sample Size*

A sufficient sample size is important to address any scientific question with empirical data. First, we have to realize that the effective sample size may often be much smaller than indicated by the total number of subjects in a study.[174] For example, when we study complications of a procedure that occur with an incidence of 0.1%, a study with 10,000 patients will contain only 10 events. The number 10 determines the effective sample size in such a study. In small samples, model uncertainty may be large, and we may not be able to derive reliable predictions from a model.

Second, a large sample size facilitates many aspects of prediction research. For example, large-scale international collaborations are increasingly set up to allow for the identification of gene-disease associations.[211] For multivariable prognostic modelling, a large sample size allows for selection of predictors with simple automatic procedures such as stepwise methods with $p<0.05$ and reliable testing of model assumptions. An example is the prediction of 30-day mortality after an acute myocardial infarction, where Lee et al. derived a prediction model with 40,830 patients of whom 2,850 died.[255] This example will be used throughout this book, with a thorough description in Chap. 22. In practice, we often have relatively small samples available. For example, a review of 31 prognostic models in traumatic brain injury showed that 22 were based on samples with less than 500 patients.[307] The main challenges are hence with the development of a good prediction model with a relatively small study sample.

Third, with small sample size we have to be prepared to make stronger modelling assumptions. For example, Altman illustrates the use of a parametric test (ANOVA) to compare 3 groups with 8, 9, and 5 patients in his seminal text "Practical statistics for medical research".[8] With larger samples, we would more readily switch to a non-parametric test such as a Kruskal–Wallis test. With small sample size, we may have to assume linearity of a continuous predictor (Chap. 9) and no interaction between predictors (Chap. 13). We will subsequently have limited power to test deviations from these model assumptions. It hence becomes more important what our starting point of the analysis is. From a Bayesian viewpoint, we could say that our prior information becomes more important, since the information contributed by our study is limited.

Fourth, we have to match our ambitions in research questions with the effective sample size that is available. When the sample size is very small, we should only ask relatively simple questions, while more complex questions can be addressed with larger sample sizes. A question such as: "What are the most important predictors in this prediction problem" is actually more complex than a question such as "What are the predictions of the outcome given this set of predictors" (Chap. 11). Table 1.1 lists questions on predictors (known or determined from the data?), functional form (known or determined from the data?), and regression coefficients (known or determined from the data?) and the consequence for the required sample size in a study.

Table 1.1 Stages of dev and required sample size

Predictors known?	Function form kno
–	–
+	–
+	+
+	+

1.3 Structure o

This book consists o applying prediction m while Part III focuse related to model modi cal in nature with a de some lessons learned lems with publicly ava

1.3.1 *Part I: Pre*

This book starts with ar cal practice and in med tistical model depends the analysis. A sophisti data collection procedu cohort studies for progn (Chap. 3). Various stati can be considered for a els for different types o models commonly suffe ting." This means that patterns.[174] A model ma ting of application is ve discussed with possible choosing between altern quality of predictions de