Chapter 3
Study Design for Prediction Models

Background In this chapter, we consider several issues in the design of studies for prediction research. These include the selection of subjects or patients for a cohort study, strengths and limitations of case series from a single center, from registries, or prospective trials. We further discuss issues in choosing predictors and outcome variables for prediction models. An important question is often how large a study needs to be for sufficient statistical power. Power considerations are given for studying effects of specific predictors, and for developing a prediction model that can provide reliable predictions. We use several case studies for illustration.

3.1 Study Design

Prognostic studies are inherently longitudinal in nature. They are most often performed in cohorts of patients, who are followed over time for an outcome to occur. The cohort is defined by the presence of one or more particular characteristics, e.g. having a certain disease, living in a certain place, having a certain age, or simply being born alive. For example, we may follow a cohort of patients with an acute myocardial infarction for long-term mortality according to ECG characteristics.[335]

Diagnostic studies are most often designed as a cross-sectional study, where predictive patient characteristics are related to an underlying diagnosis. The study group is defined by the presence of a particular symptom or sign that makes the subject suspected of having a particular (target) disease. Typically, the subjects undergo the index test and subsequently a reference test to establish the "true" presence or absence of the target disease, over a short time span. For example, we may aim to diagnose those with an acute myocardial infarction among patients presenting at an emergency department.[142]

3.2 Cohort Studies for Prognosis

Several types of cohort studies can be used for prognostic modelling. The most common type may be a single-center retrospective cohort study (Table 3.1). In this case, patients are identified from hospital records between certain dates, for example,

E.W. Steyerberg, *Clinical Prediction Models*,
DOI:10.1007/978-0-387-77244-8_3,

Table 3.1 Study designs for prognostic studies

Design	Characteristics	Strengths	Limitations
Retrospective	Often single-centre studies	Simple, low costs	Selection of patients Definitions and completeness of predictors Outcome assessment not by protocol
Prospective	Often multicentre RCT	Well-defined selection of patients Prospective recording of predictors Prospective assessment of outcome according to protocol	Poor generalizability because of stringent in- and exclusion criteria
Registry	Complete coverage of an area/participants covered by insurance	Simple, low costs Prospective recording of predictors	Outcome assessment not by protocol
Case–control	Efficient when outcome relatively rare	Simple, low costs	Selection of controls critical Definitions and completeness of predictors Outcome assessment not by protocol

those diagnosed between January 1, 1997, and December 31, 2003. These patients were followed over time for the outcome, but the investigator looks back in time (hence we may use the label "retrospective study"[463]).

3.2.1 Retrospective Designs

Strengths of a retrospective study design include its simplicity and feasibility. It is a design with relatively low costs, since patient records can often easily be searched, especially with modern hospital information systems or electronic patient records. A limitation is the correct identification of patients, which has to be done in retrospect. If some information is missing, or was incorrectly recorded, this may lead to a selection bias. Similarly, the recording of predictors has to have been reliable to be useful for prediction modelling. Finally, the outcome has to be reliable. This may be relatively straightforward for outcomes such as survival, where some deaths will be known from hospital records. But additional confirmation of vital status may often be required from nationwide statistical bureaus for a complete assessment of survival status. Other outcomes, e.g. related to functional status, may not be available at the time points that we wish to analyse. Finally, single centre studies may be limited by their sample size, which is a key problem in prediction research. Multicentre, collaborative studies can address this sample size issue. Moreover, the representativeness of the prediction model will then be better.

*3.2.2 Example: Predicting Early Mortality in Oesophageal Cancer

As an example, we consider outcome prediction in oesophageal cancer. A retrospective chart review was performed of 120 patients treated in a single institution between January 1, 1997, and December 31, 2003.[252] The patients had palliative treatment, which means therapy that relieves symptoms, but does not alter the course of the disease. A stent was placed in the oesophagus because of malignancy-related dysphagia (difficulty in swallowing). The authors studied 30-day mortality, which occurred in an unspecified number of patients (probably around 10%, n=12).[252]

3.2.3 Prospective Designs

In a prospective study, we can better check specific inclusion and exclusion criteria. The investigator is said to age with the study population (hence the label "prospective study"). We can use clear and consistent definitions of predictors, and assess

patient outcomes at pre-defined time points. Prospective cohort studies are therefore preferable to analyses in retrospective series.

Prospective cohort studies are sometimes solely set up for prediction modelling, but a more common design is that prediction research is done in data from randomized clinical trials (RCTs), or from prospective before–after trials. The strengths are in the well-defined selection of patients, the prospective recording of predictors, usually with quality checks, and the prospective assessment of outcome. Sample size is usually reasonably large. A limitation of data from (randomized) trials may be in the selection of patients. Often stringent inclusion and exclusion criteria are used, which may limit the generalizability of a model developed on such data. On the other hand, RCTs are often performed in multiple centres, sometimes from multiple countries or continents. Benefits of the multi-centre design include that consensus has to be reached on definition issues for predictors and outcome, and that generalizability of findings will be increased. This is in contrast to single centre studies, which only reflect predictive relationships from one specific setting.

A topic of debate is whether we should only use patients from an RCT who are randomized to a conventional treatment or placebo (the "control group"). If we combine randomized groups we assume that no specific subgroup effects are relevant for the prognostic model. This may generally be reasonable. Moreover, the prognostic effect of a treatment is usually small compared to prognostic effects of other predictors.

*3.2.4 Example: Predicting Long-Term Mortality in Oesophageal Cancer

In another study of outcome in oesophageal cancer, data from an RCT ("SIREC", n=209[197]) were combined with other prospectively collected data (n=396).[414] Long-term mortality was studied after palliative treatment with a stent or radiation ("brachytherapy").

3.2.5 Registry Data

Prognostic studies are often performed with registry data, for example cancer registries, or insurance databases. Data collection is prospective, but not primarily for prediction research. The level of detail may be a limitation for prognostic analyses. For example, the well-known US-based cancer registry (Surveillance, Epidemiology and End Results, SEER) contains information on cancer incidence, mortality, patient demographics, and tumour stage. It has been linked to the Medicare data base for information on comorbidity[233] and treatment (surgery,[80]

chemotherapy,[478] radiotherapy[471]). Socio-economic status (SES) is usually based on median income as available at an aggregated level.[24] SEER-Medicare does not contain detailed information on performance status, which is an important factor for medical decision-making and for survival of cancer patients. Also, staging may have some measurement bias.[118]

Another problem may occur when reimbursement depends on the severity that is scored for a patient. This may pose an upward bias on the recording of comorbidities in claims databases for example.

The outcomes for prognostic analyses usually suffer from the same limitations as retrospective studies, since usually no pre-defined assessments are made. Outcomes are therefore often limited to survival, although other adverse events can sometimes also be derived.[105,394] Strengths of prognostic studies with registry data include large sample sizes, and representativeness of patients (especially with population-based cancer registries). Such large databases may especially be useful for studying predictive relationships of a limited number of predictors with survival.

*3.2.6 Example: Surgical Mortality in Oesophageal Cancer

The SEER-Medicare database was used to analyze 30-day mortality in 1,327 patients undergoing surgery for oesophageal cancer between 1991 and 1996. Predictive patient characteristics included age, comorbidity (cardiac, pulmonary, renal, hepatic, and diabetes), preoperative therapy, and a relatively low hospital volume, which were combined in a simple prognostic score. Validation was done in another registry, and in a hospital series.[423]

3.2.7 Nested Case–Control Studies

A prospectively designed, nested case–control study is sometimes an efficient option for prediction research. A case–control design is especially attractive when the outcome is relatively rare, such as incident breast cancer.[131] For example, if 30-day mortality is 1%, it is efficient to determine detailed predictors in all patients who died, but for example 4% of the controls (1:4 case–control ratio). A random sample of controls is used as comparison for the cases. If the outcome is well defined, such as survival, selection bias cannot be a problem. Assessment of details of predictors is in retrospect, which is a limitation. If a prediction model is developed, the average outcome incidence has to be adjusted for final calculation of probabilities, while the regression coefficients can be based on the case–control study.[131]

*3.2.8 Example: Perioperative Mortality in Major Vascular Surgery

An interesting example is the analysis of perioperative mortality in patients undergoing major vascular surgery.[340] Predictors were determined in retrospect from a detailed chart review in all cases (patients who died), and in selected controls (patients who did survive surgery). Controls had surgery just before and just after the case. Hence a 1:2 ratio was achieved for cases against controls.

3.3 Studies for Diagnosis

3.3.1 Cross-Sectional Study Design and Multivariable Modelling

Ideally, a diagnostic study considers a well-defined cohort of patients suspected of a certain diagnosis, e.g. an acute myocardial infarction.[238] Such a diagnostic study then resembles a prognostic cohort study. The cohort is here defined by the suspicion of having (rather than actually having) a disease. The outcome is the underlying diagnosis. The study may therefore be labelled cross-sectional, since the predictor–outcome relationships are studied at a single point in time. Several characteristics may be predictive of the underlying diagnosis. For a model, we should start with considering simple characteristics such as demographics, and symptoms and signs obtained from patient history. Next, we may consider simple diagnostic tests, and finally invasive and/or costly tests.[295] The diagnosis (presence or absence of the target disease) should be established by a reference test or standard. This test used to be called "gold" standard, but no method is 24 carat gold. The result of the reference test is preferably interpreted without knowledge of the predictor and diagnostic test values. Such blinding prevents information bias (or incorporation, or "diagnostic review" bias).[296]

A common problem in diagnostic evaluations is the incomplete registration of all predictive characteristics. Not all patients may have undergone the entire diagnostic work-up, especially if they are considered as at low risk of the target disease. Similarly, outcome assessment may be incomplete, if a test is used as a gold standard which is selectively performed.[343] These problems are especially prominent in diagnostic analyses on data from routine practice.[313] Prospective studies are hence preferable, since these may use a pre-specified protocol for systematic diagnostic work-up and reference standard testing.

*3.3.2 Example: Diagnosing Renal Artery Stenosis

A cardiology database was retrospectively reviewed for patients who underwent coincident screening abdominal aorta angiography to detect occult renal artery stenosis. In a development set, stenosis was observed in 128 of 635 patients. This 20%

prevalence may be an overestimate if patients underwent angiography because of suspicion of stenosis.[347]

3.3.3 Case–Control Studies

Diagnostic studies sometimes select patients on the presence or absence of the target disease as determined by the reference test. Hence patients without a reference standard are not selected. In fact, a case–control study is performed, where cases are those with the target disease, and controls those without. This design has a number of limitations, especially related to the representativeness of the selected patients for all patients who are suspected of the diagnosis of interest. Selection bias is the most important limitation. Indeed, empirical evidence is now available on the bias that arises in diagnostic studies, especially by including non-consecutive patients in a case–control design, non-representative patients (severe cases compared to healthy controls), and when data are collected retrospectively.[259,361]

*3.3.4 Example: Diagnosing Acute Appendicitis

C-reactive protein (CRP) has been used for the diagnosis of acute appendicitis. Surgery and pathology results constituted the reference test for patients with a high CRP. Patients with a low CRP were not operated on and clinical follow-up determined whether they were classified as having acute appendicitis. As low-grade infections with low CRPs can resolve spontaneously, this verification strategy fails to identify all false-negative test results. In this way, the diagnostic performance of CRP will be overestimated.[259]

3.4 Predictors and Outcome

3.4.1 Strength of Predictors

For a well-performing prediction model, strong predictors have to be present. Strength is a function of the association of the predictor with the outcome, and the distribution of the predictor. For example, a dichotomous predictor with an odds ratio of 2.0 is more relevant for a prediction model than a dichotomous predictor with an odds ratio of 2.5, when the first predictor is distributed in a 50:50 ratio (50% prevalence of the predictor), and the second 1:99 (1% prevalence of the predictor). Similarly, continuous predictors have to cover a wide range to make them relevant for prediction.

When some characteristics are considered as key predictors, these have to be registered carefully, with clear definitions and preferably no missing values. This is

usually best possible in a prospective study, with a protocol and pre-specified data collection forms.

3.4.2 Categories of Predictors

Several categories of predictors have been suggested for prediction models.[174] These include

- Demographics (e.g. age, sex, race, socio-economic status)
- Type and severity of disease (e.g. principal diagnosis, presenting characteristics)
- History characteristics (e.g. previous disease episodes, risk factors)
- Comorbidity (concomitant diseases)
- Physical functional status (e.g. Karnofsky score, WHO performance score)
- Subjective health status and quality of life (psychological, cognitive, psychosocial functioning)

The relevance of these categories will depend on the specifics of the application. Publications tend to group predictors under general headings, see for example, the predictors in the GUSTO-I model (Chap. 22).[255] Of note, definitions of predictors may vary from study to study.[492] Socioeconomic status (SES) can be defined in many ways, considering a patient's working status, income, and/or education. Also, SES indicators are sometimes not determined at the individual level, but for example at census tract level ("ecological SES", e.g. in analyses of SEER-Medicare data[24,404]). Race/ethnicity can be defined in various ways, and sometimes be self-reported rather than determined by certain pre-defined rules. Comorbidity definitions and scoring systems are still under development.[91,126,201] Variation in definitions is a serious threat to the generalizability of prediction models.[16]

Another differentiation is to separate the patient's condition from his/her constitution. Condition may be reflected in type and severity of disease, history characteristics, comorbidity, physical and subjective health status. Constitution may especially be related to demographics such as age and gender. For example, the same type of trauma (reflected in patient condition) affects patients of different ages differently (constitution).

In the future, genetic characteristics will be used more widely in a prediction context. Inborn variants of the human genome, such as polymorphisms and mutations, may be considered as indicators of the patient's constitution. Other genetic characteristics, for example the genomic profile of a malignant tumour, may better be thought of as indicators of subtypes of tumours, reflecting condition.

3.4.3 Costs of Predictors

Predictors may require different costs, in monetary terms, but also in burden for a patient. In a prediction context, it is evident that information that is easy to obtain

should be considered before information that is more difficult to obtain. Hence, we should first consider characteristics such as demographics and patient history, followed by simple diagnostic tests, and finally invasive and/or costly tests. Expensive genetic tests should hence be considered for their incremental value over classical predictors rather than alone.[225] Such an incremental evaluation is well possible with predictive regression models, where a model is first considered without the test, and subsequently a model with the test added.[399]

3.4.4 Determinants of Prognosis

Prognosis can also be viewed in a triangle of interacting causes (Fig 3.1). Predictors may be separated as related to environment (e.g. socio-economic conditions, health care access and quality, climate), the host (e.g. demographic, behavioral, psychosocial, premorbid biologic factors), and disease (e.g. imaging, pathophysiologic, genomic, proteomic, metabolomic factors).[184]

3.4.5 Prognosis in Oncology

For prognosis in oncology, it has been proposed to separate factors related to the patient, the tumour and to treatment (Fig.3.2).[186] Examples of patients characteristics include demographics (age, sex, race/ethnicity, SES), comorbidity, functional status. Tumour characteristics include the extent of disease (e.g. reflected in TNM stage), pathology, and sometimes values of tumour markers in the blood. Treatment may commonly include (combinations of) surgery, chemotherapy, and radiotherapy.

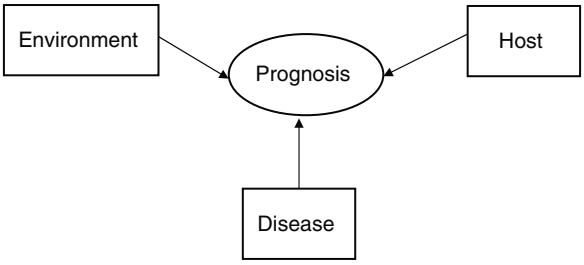

Fig. 3.1 Prognosis may be thought of as determined by predictors related to environment, host and disease[184]

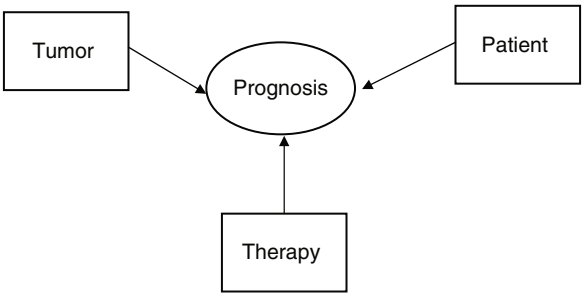

Fig. 3.2 Prognosis of a patient with cancer may be thought of as determined by predictors related to the tumour, the patient, and therapy[186]

3.5 Reliability of Predictors

3.5.1 Observer Variability

We generally prefer predictors that are well defined and reliably measured by any observer. In practice, observer variability is a problem for many measurements.[185,246] Disciplines include, for example pathologists, who may unreliably score tissue specimens for histology, cell counts, colouring of cells, and radiologists, who, for example, score *X*-rays, CT scans, MRI scans, and ultrasound measurements. This variability can appropriately be measured with kappa statistics.[248] The interobserver and intraobserver variability can be substantial, which will be reflected in low kappa values.

**3.5.2 Example: Histology in Barrett's Oesophagus*

Barrett's oesophagus is a pre-malignant condition. Surgery is sometimes performed in high-grade dysplasia, whereas other physicians defer treatment until adenocarcinoma is diagnosed. The agreement between readings of histology in Barrett's oesophagus for high-grade dysplasia or adenocarcinoma was only fair, with kappa values around 0.4.[314] The agreement between no dysplasia and low-grade dysplasia had been reported as even lower.[389] Because of observer variability, sometimes a central review process is organized, where an expert reviews all readings. This should be done independently and blinded for previous scores. Subsequently a rule has to be determined for the final score, for example that only the expert score is used, or that an additional reader is required in case of disagreement. Also, consensus procedures can be set up with experts only, for example with scoring by two experts, and involvement of a third if these disagree.[230] Some use the unreliability of classical pathology as an argument for using modern biomarkers.[247]

3.5.3 Biological Variability

Apart from observer variability, some measurements are prone to biological variability. A well-known example is blood pressure, where a single measurement is quite unreliable.[318] Usually at least two measurements are made, and preferably more, with some spread in time. Again, definitions have to be clear (e.g. position of patient at the measurement, time of day).

3.5.4 Regression Dilution Bias

The effect of unreliable scoring by observers, or biological variability, generally is a dilution of associations of predictors with the outcome. This has been labelled "regression dilution bias", and methods have been proposed to correct for this bias.[257] A solution is to repeat unreliable measurements, either by the same observer (e.g. use the mean of three blood pressure measurements), or different observers (e.g. double reading of mammograms by radiologists). Practical constraints may limit such procedures.

*3.5.5 Example: Simulation Study on Reliability of a Binary Predictor

Suppose we have a binary predictor that we measure with noise. Suppose two observers make independent judgments of the predictor. Their judgments agree with the true predictor status with sensitivity of 80% (observer scores 1 if true = 1) and specificity of 80% (observer scores 0 if true = 0, Table 3.2). If both observers score the predictor independently and without correlation, the observers agree with each other with a kappa of only 0.36 (Table 3.3).

The true predictor status predicts outcome well, with an odds ratio of 4. The observed predictor status has a diluted predictive effect, with odds ratio 2.25. Similarly, the discriminative ability is diluted (*c* statistic decreases from 0.67 to 0.60, Table 3.4).

Table 3.2 Sensitivity and specificity for observers in determining the true predictor status (sensitivity = specificity = 80%)

		True predictor status	
		0 *N*(col%)	1 *N*(col %)
Observer	0	750 (80%)	187 (20%)
	1	187 (20%)	750 (80%)

Table 3.3 Agreement between observer 1 and observer 2 (kappa = 0.36)

		Observer 2	
		0	1
Observer 1	0	637	300
	1	300	637

Table 3.4 Association with outcome for the true predictor status and observed predictor status (by observer 1 or 2, Table 3.3)

		Outcome			
		0 *N* (row%)	1 *N* (row%)	Odds ratio	*c* statistic
True predictor status	0	625 (67%)	312 (33%)	4.0	0.67
	1	312 (33%)	625 (67%)		
Observer	0	562 (60%)	375 (40%)	2.25	0.60
	1	375 (40%)	562 (60%)		

3.5.6 Choice of Predictors

In aetiologic research we may often aim for the best assessment of an exposure variable. We will be concerned about various information biases that may occur. In the context of a prediction model we can be much more pragmatic. If we aim to develop a model that is applicable in daily practice, we should use definitions and scorings that are in line with daily practice. For example, if medical decisions on surgery are made considering local pathology reports, without expert review, the local pathology report should be considered for a prediction model applicable to the local setting. As illustrated, such less reliable assessments will affect the performance of a predictive model, since predictive relationships are disturbed. If misclassification is at random, a dilution of the relationship occurs (Table 3.4). On the other hand, if measurements are more reliable in clinical practice than in a research setting, e.g. repeated assessments of blood pressure, we might argue that a correction has to be made in the prediction model. In practice, prediction models tend to include predictors that are quite readily available, not too costly to obtain, and can be measured with reasonable precision.

3.6 Outcome

3.6.1 Types of Outcome

The outcome of a prediction model should be relevant, either from an applied medical perspective or from a research perspective. From a medical perspective, “hard” end points are generally preferred. Especially mortality is often used as an end point in prognostic research. Mortality risks are relevant for many acute and chronic

conditions, and for many treatments, such as surgery. In some diseases, mortality may not be a relevant outcome. Other outcomes include non-fatal events (e.g. disease recurrence), patient centred outcomes (e.g. scores on quality of life questionnaires), or wider indicators of burden of disease (e.g. absence from work, Table 3.5, based on Hemingway[184]).

3.6.2 Survival End points

When cause-specific mortality is considered, a reliable assessment of the cause of death is required. If cause of death is not known, relative survival can be calculated.[166,167] This is especially popular in cancer research. Mortality in the patients with a certain cancer is compared with the background mortality from the general population. The difference can be thought of as mortality due to the cancer.

The pros and cons of relative survival estimates are open to debate. Some have proposed to also study conditional survival for patients already surviving for some years after diagnosis. These measures may sometimes be more meaningful for clinical management and prognosis than 5-year relative survival from time of diagnosis.[139,214] Others have proposed that median survival times are better indicators of survival than 5-year relative survival rates, especially when survival times are short.[319]

**3.6.3 Example: Relative Survival in Cancer Registries*

Five-year relative survival was studied for patients enrolled in the SEER registry in the period 1990–1999.[139] The 5-year relative survival rate for persons diagnosed with cancer was 63%, with substantial variation by cancer site and stage at diagnosis.

Table 3.5 Examples of prognostic outcomes[184]

Prognostic outcome	Example	Characteristics
Fatal events	All-cause, or cause-specific	Hard end point, relevant in many diseases, but sometimes too infrequent for reliable statistical modeling
Non-fatal events	Recurrence of tumor, cardiovascular events (e.g. myocardial infarction, revascularization)	Somewhat softer end point, reflecting decision-making by physicians, increases power for analysis
Patient centered	Symptoms, functional status, health-related quality of life, utilities	Subjective end point, focused on the patients themselves; often used as secondary end point
Wider burden	Absence from work because of sickness	Especially of interest from an economical point of view

Five-year relative survival increased with time since diagnosis. For example, for patients diagnosed with cancers of the prostate, female breast, corpus uteri, and urinary bladder, the relative survival rate at 8 years after diagnosis was over 75%.

Similar analyses were performed with registry data from the Eindhoven region, where it was found that patients with colorectal, melanoma skin, or stage I breast cancer could be considered cured after 5–15 years, whereas for other tumours survival remained poorer than the general population.[214]

3.6.4 Composite End Points

Sometimes composite end points are defined, which combine mortality with non-fatal events. Composite end points are especially popular in cardiovascular research (see also Chap. 23). For example, the Framingham models have been used to predict incident cardiovascular disease in the general population. A popular Framingham model (the Wilson model) defines cardiovascular events as fatal or non-fatal myocardial infarction, sudden death, or angina pectoris (stable or unstable).[487] Composite end points have the advantage of increasing the effective sample size and hence the power for statistical analyses.

*3.6.5 Example: Mortality and Composite End Points in Cardiology

A prediction model was developed in 949 patients with decompensated heart failure. The outcome was 60-day mortality or the composite end point of death or rehospitalization at 60 days. The discriminatory power of the model was substantial for mortality (*c* statistic 0.77) but less for the composite end point (*c* statistic 0.69).[121] These findings are in line with prediction of acute coronary syndromes, where predictive performance was better for mortality than for a composite end point of mortality or myocardial (re)infarction.[43] The case study in Chap. 23 also considers a composite end point.

3.6.6 Choice of Prognostic Outcome

The choice of a prognostic outcome should be guided by the prediction problem, but the outcome should be measured as reliable as possible. Prediction models may be developed with pragmatic definitions of predictors, since this may resemble the future use of a model. But the outcome should be determined with similar rigour as in an aetiologic study or randomized clinical trial. In the future, decisions are to be based on the predictions from the model. Predictions hence need to be based on robust statistical associations with an accurately determined outcome.

If there is a choice between binary and continuous outcomes, the latter are preferred from a statistical perspective, since they provide more power in the analysis. Also, ordered outcomes provide more power than binary outcomes. In practice, binary outcomes are however very popular, making logistic regression and Cox regression the most common techniques for prediction models in medicine.

3.6.7 Diagnostic End Points

The outcome in diagnostic research naturally is the underlying disease, which needs to be defined according to a reference standard.[48,49,238,296] The reference standard can sometimes be anatomical, e.g. findings at surgery. Other definitions may include blood or spinal fluid cultures (e.g. in infectious diseases), results of high-quality diagnostic tests such as angiography (e.g. in coronary diseases), and histological findings (e.g. in oncology). Methods are still under development on how to deal with the absence of an acceptable reference standard. In such situations the results of the diagnostic test can, for example, be related to relevant other clinical characteristics and future clinical events.[360]

The relevance of the underlying diagnosis may be high when treatment and prognosis depends directly on the diagnosis. This is for example the case with testing for genetic defects such as trisomy 21 (Down syndrome). However, often a diagnosis covers a spectrum of more and less severe disease, and longer-term outcome assessment would be desirable. This is especially relevant in the evaluation of newer imaging technology, which may detect disease that remained previously unnoticed.[34,266]

**3.6.8 Example: PET Scans in Oesophageal Cancer*

In oesophageal cancer, positron-emission tomography (PET) scans provide additional information on extent of disease compared to CT scanning alone.[316,495] However, the clinical relevance of the additionally detected metastases can only be determined in a comparative study, preferably a randomized controlled trial. Diagnosing more metastases is not sufficient to make PET/CT clinically useful.[462]

3.7 Phases of Biomarker Development

Pepe has proposed a phased approach to developing predictive biomarkers, in particular for early detection of cancer[332] (Table 3.6). These phases are also relevant to the development of future prediction models, which may add novel biomarkers to traditional clinical characteristics. The development process begins with small studies focused on classification performance and ends with large studies of impact on

Table 3.6 Phases of development of a biomarker for cancer screening[332]

Phase	Objective	Study design
1. Preclinical exploratory	Promising directions identified	Case–control (convenient samples)
2. Clinical assay and validation	Determine if a clinical assay detects established disease	Case–control (population based)
3. Retrospective longitudinal	Determine if the biomarker detects disease before it becomes clinical. Define a "screen positive" rule	Nested case–control in a population cohort
4. Prospective screening	Extent and characteristics of disease detected by the test; false referral rate	Cross-sectional population cohort
5. Cancer control	Impact of screening on reducing the burden of disease on the population	Randomized trial

populations. The aim is to select promising markers early while recognizing that early studies do not answer the ultimate questions that need to be addressed.

As an example, Pepe considers the development of a biomarker for cancer screening. Phase 1 is exploratory and may consider gene expression arrays or protein mass spectrometry that yields high dimensional data for biomarker discovery. Reproducibility between laboratories is an aspect to consider before moving on to phase 2, where a promising biomarker is compared between population-based cases with cancer and population-based controls without cancer. Phase 3 is a more thorough evaluation in a case–control study to determine if the marker can detect subclinical disease. In phase 4, the marker may be applied prospectively as a screening test in a population. Finally, the overall impact of screening is addressed in phase 5 by measuring effects on clinically relevant outcomes such as mortality.

The study design implications are also shown in Table 3.6. In the exploratory phase 1 it may be acceptable to use "convenient samples", which will likely lead to spectrum bias in the assessment of the biomarker. In phase 2, population-based samples are desired for a simple case–control design. In phase 3, we require samples taken from cancer patients before their disease became clinically apparent. A nested case–control study design can be efficient for data from a cohort study. For phase 4, a prospective cohort study is required to determine the characteristics and treatability of early detected disease. Finally, an RCT is desired for unbiased assessment of the impact of screening.

3.8 Statistical Power

An important issue is how large a study needs to be for sufficient statistical power to address the primary research question. Power considerations are given for studying effects of a specific predictor, and for developing a prediction model that can provide reliable predictions.

3.8.1 Statistical Power to Identify Predictor Effects

We may primarily be interested in the effect of a specific predictor on a diagnostic or prognostic outcome. We may then aim to test the effect of this predictor for statistical significance. This leads to similar sample size considerations as for testing of treatment effects, e.g. in the context of an RCT. Sample size calculations are straightforward for such univariate testing. The required sample size is determined by choices for the acceptable Type I and Type II error. The Type I error is usually set at 5% for statistical significance. The Type II error determines the power, and may, e.g. be set at 20% for 80% power. Other considerations are the variability of the effect estimate. For binary predictors of a binary outcome, the prevalence of the predictor and the incidence of the outcome are important. Finally, the magnitude of the effect determines the required sample size, with larger sample size required to detect smaller effects.

*3.8.2 Examples of Statistical Power Calculations

Sample size calculations can be performed for most types of regression models with standard software. For illustration, we consider the statistical power for a binary predictor of a binary outcome (Fig. 3.3). We find that the required sample size increases steeply with a very low or very high incidence of the outcome. With

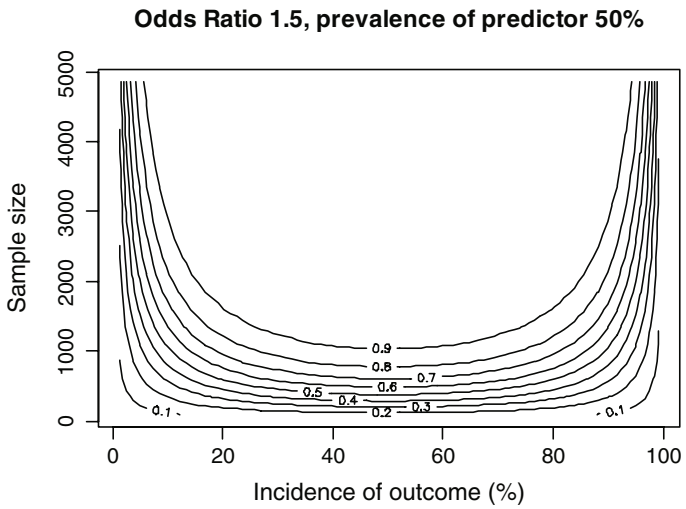

Fig. 3.3 Power corresponding to sample sizes for incidence of the outcome ranging from 0 to 100%. A binary predictor was considered with 50% prevalence with odds ratio 1.5

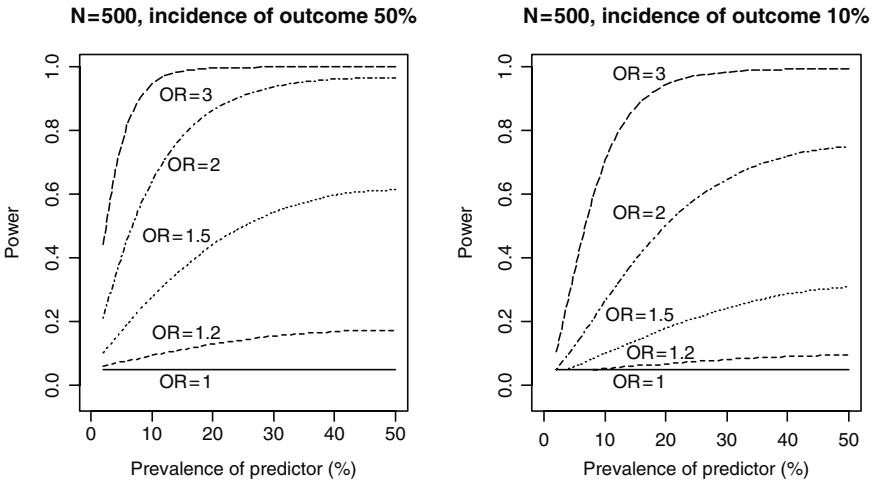

Fig. 3.4 Power in relation to prevalence of a binary predictor, for odds ratios from 1 to 3 in samples with 500 subjects. Incidences of the outcome were 50% (*left panel*) and 10% (*right panel*)

an odds ratio of 1.5, 80% power requires approximately 2,000 subjects at a 10% incidence, 1,000 subjects at 20% incidence, and 800 subjects at 50% incidence.

Next, we illustrate that statistical power is related to the prevalence of a binary predictor (Fig. 3.4). We consider odds ratios from 1 to 3, as may often be encountered in medical prediction research. In a sample size of 500 subjects, 250 with and 250 without the outcome, 80% power is reached with prevalences of 16% and 5.5% for odds ratios of 2 and 3, respectively. Odds ratios of 1.2 and 1.5 require sample sizes of 3,800 and 800 at 50% prevalence, respectively. With 10% incidence of the outcome, power is substantially lower (Fig. 3.4, right panel). An odds ratio of 3 now requires 18% instead of 5.5% prevalence of the predictor for 80% power. Without an effect (OR=1), statistical significance is by definition expected in 5%.

3.8.3 Statistical Power for Reliable Predictions

Instead of focusing on predictors, we can consider the reliability of predictions that are provided by a prediction model. Some rules of thumb have been proposed, supported by simulation studies. The sample size requirements are commonly formulated as events per variable (EVP). The minimum EVP for obtaining good predictions may be 10.[175,326,327] Clinical prediction models that are constructed with EVP less than 10 are overfitted, and may perform poorer than a simpler model which considers fewer predictors, such that the EVP is at least 10 (see further illustration in Chap. 24). EPV values for reliable selection of predictors from a larger set of candidate predictors may be as large as 50 (events per candidate predictor, see Chap. 11). For

pre-specified models, shrinkage may not be required with EPV of at least 20 (Chap. 13).[410] Validation studies may need to include at least 100 events (details in Chap. 19).[465]

Other EPV values may apply for specific circumstances. Regression analyses can technically well be performed with lower EVP values. Adjusted analyses of an exposure variable may be performed with EPV less than 10 when we only aim to correct for confounding.[473]

3.9 Concluding Remarks

Prognostic studies are ideally designed as prospective cohort studies, where the selection of patients and definition of predictors is pre-specified. Data from randomized clinical trials may often be useful, although representativeness of the included patients for the target population should be considered as a limitation. Data may also be used from retrospective designs, registries, and case–control studies, each with their strengths and limitations. Diagnostic studies are usually cross-sectional in design, and should prospectively select all patients who are suspected of a disease of interest. In practice, designs are still frequent where patients are selected by a reference test which is not performed in all patients.

Predictors should be defined pragmatically, and cover the relevant domains for prediction of outcome in a disease. The outcome of a prediction model should be measured with high accuracy. Hard end points such as mortality are often preferred but statistical power considerations may motivate the use of composite and other end points.

Questions

3.1 Cohort studies
One could argue that both diagnostic and prognostic studies are examples of cohort studies.

(a) What is the difference between diagnostic and prognostic outcomes in such cohorts?
(b) What is the implication for the statistical analysis?

3.2 Prospective vs. retrospective designs (Sect. 3.2)
Prospective study designs are generally noted as preferable to retrospective designs. What are the pros and cons of prospective vs. retrospective designs?

3.3 Accuracy of predictors and outcome (Sect. 3.5 and 3.6)

(a) Why do we need to be more careful with reliable assessment of outcome than reliable assessment of predictors?
(b) What is the effect of imprecise measurement of a predictor?

3.4 Composite end points (Sect. 3.6.4)
Composite end points are often motivated by the wish to increase statistical power for analysis. What is the price that we pay for this increase in term of assumptions on predictive relationships? See a recent JCE paper for a detailed discussion.[140]

3.5 Statistical power (Figs. 3.3 and 3.4)

(a) What is the required total sample size for 50% power at 10%, 30%, or 50% incidence of the outcome in Fig. 3.3?
(b) What is the similarity between Fig. 3.3 and 3.4 with respect to the ranges of the incidence of the outcome or prevalence and associated statistical power?