Part I
Prediction Models in Medicine

Chapter 2
Applications of Prediction Models

Background In this chapter, we consider several areas of application of prediction models in public health, clinical practice, and medical research. We use several small case studies for illustration.

2.1 Applications: Medical Practice and Research

Broadly speaking, prediction models are valuable for medical practice and for research purposes (Table 2.1). In public health, prediction models may help to target preventive interventions to subjects at relatively high risk of having or developing a disease. In clinical practice, prediction models may inform patients and their treating physicians on the probability of a diagnosis or a prognostic outcome. Prognostic estimates may for example be useful for planning of remaining life-time in terminal disease; or give hope for recovery if a good prognosis is expected after an acute event such as a stroke. Classification of a patient according to his/her risk may also be useful for communication among physicians. A key condition for this type of application of a prediction model is that predictions are reliable. This means that when a 10% risk is predicted, on average 10% of patients with these characteristics should have the outcome ("calibration", Chap. 4 and 15).

In the diagnostic work-up, predictions can be useful to estimate the probability that a disease is present. When the probability is relative high, treatment is indicated; if the probability is low, no treatment is indicated and further diagnostic testing may be considered necessary. In therapeutic decision-making, treatment should only be given to those who benefit from the treatment. Prognostic predictions may support the weighing of harms vs. individual benefits. If risks of a poor outcome are relatively low, the maximum benefit will also be relatively low. Any harm, such as a side effect of treatment, may then readily outweigh any benefits. The claim of prediction models is that better decisions can be made with a model than without.

In research, prediction models may assist in the design and analysis of randomized trials. Models are also useful to control for confounding variables in observational research, either in traditional regression analysis or with modern approaches such

E.W. Steyerberg, *Clinical Prediction Models*,
DOI:10.1007/978-0-387-77244-8_2, © Springer Science+Business Media, LLC 2009

Table 2.1 Some areas of application of clinical prediction models

Application area	Example in this chapter
Public health	
Targeting of preventive interventions	
Incidence of disease	Models for (hereditary) breast cancer
Clinical practice	
Diagnostic work-up	
Test ordering	Probability of renal artery stenosis
Starting treatment	Probability of deep venous thrombosis
Therapeutic decision-making	
Surgical decision making	Replacement of risky heart valves
Intensity of treatment	More intensive chemotherapy in cancer patients
Delaying treatment	Spontaneous pregnancy chances
Research	
Inclusion in an RCT	Traumatic brain injury
Covariate adjustment in an RCT	Primary analysis of GUSTO-III
Confounder adjustment with a propensity score	Statin effects on mortality
Case-mix adjustment	Provider profiling

as “propensity scores”. Several areas of application are discussed in the next sections.

2.2 Prediction Models for Public Health

2.2.1 *Targeting of Preventive Interventions*

Various models have been developed to predict the future occurrence of disease in asymptomatic subjects in the population. Well-known examples include the Framingham risk functions for cardiovascular disease.[487] The Framingham risk functions underpin several of the current policies for preventive interventions. For example, statin therapy is only considered for those with relatively high risk of cardiovascular disease. Similarly, prediction models have been developed for breast cancer, where more intensive screening or chemoprophylaxis can be considered for those at elevated risk.[130,131]

*2.2.2 *Example: Incidence of Breast Cancer*

In 1989, Gail et al. presented a by now famous risk prediction model for developing breast cancer.[131] The model was based on case–control data from the Breast Cancer Detection Demonstration Project (BCDDP). The BCDDP recruited 280,000 women from 1973 to 1980 who were monitored for 5 years. From this cohort, 2,852 white women developed breast cancer and 3,146 controls were selected, all with complete risk factor information. The model includes age at menarche, age at first live birth, number of previous biopsies, and number of first-degree relatives with breast cancer. Individualized breast cancer probabilities were calculated from information on relative risks and the baseline hazard rate in the general population. The calculations accounted for competing risks (the risk of dying from other causes).

The predictions were validated later on other data sets from various populations, with generally favorable conclusions.[83,94] Practical application of the original model involved cumbersome calculations and interpolations. Hence, more easily applicable graphs were created to estimate the absolute risk of breast cancer for individual patients for intervals of 10, 20, and 30 years.[33] The absolute risk estimates have been used to design intervention studies, to counsel patients regarding their risks of disease, and to inform clinical decisions, such as whether or not to take tamoxifen to prevent breast cancer.[132]

Other models for breast cancer risk include the Claus model, which is useful to assess risk for familial breast cancer.[74] This is breast cancer that runs in families but is not associated with a known hereditary breast cancer susceptibility gene. Unlike the Gail model, the Claus model requires the exact ages at breast cancer diagnosis of first or second-degree relatives as an input.

Some breast cancers are caused by a mutation in a breast cancer susceptibility gene (BRCA), referred to as hereditary breast cancer. A suspicious family history for hereditary breast cancer includes many cases of breast and ovarian cancers, or family members with breast cancers under age 50. Simple tables have been published to determine the risk of a BRCA mutation, based on specific features of personal and family history.[127] Another model considers the family history in more detail (BRCAPRO[323]). It explicitly uses the genetic relationship in families, and is therefore labeled a Mendelian model. Calculations are based on Bayes' theorem. BRCAPRO was shown to perform at least as good as experienced genetic counselors.[116]

Friedenson provides an interesting overview of risk models in breast cancer and their clinical implications (Table 2.2).[128] Various measures are possible to reduce breast cancer risk, including behavior (e.g. exercise, weight control, alcohol intake) and medical interventions (e.g. tamoxifen use).

2.3 Prediction Models for Clinical Practice

2.3.1 Decision Support on Test Ordering

Prediction models may be useful to estimate the probability of an underlying disease, such that we can decide on further testing. When a diagnosis is very unlikely, no further testing is indicated, while more tests may be indicated when the diagnosis is not yet sufficiently certain for decision-making on therapy. Further testing usually involves one or more imperfect tests (sensitivity below 100%, specificity below 100%). Ideally, a gold standard test is available (sensitivity=100%, specificity=100%). A gold standard test is the diagnostic test that is regarded as definitive

Table 2.2 Risk factors in four prediction models for breast cancer: two for breast cancer incidence, two for presence of mutation in BRCA1 or BRCA2 genes[128]

Risk factor	Gailmodel	Clausmodel	Myriad tables	BRCAPRO model
Woman's personal information				
Age	+	+	+	+
Race/ethnicity	+			
Ashkenazi Jewish			+	+
Breast biopsy	+			
Atypical hyperplasia	+			
Hormonal factors				
Age at menarche	+			
Age at first live birth	+			
Age at menopause	+			
Family history				
1st degree relatives with breast cancer	+	+	Age <50/≥50	Age for all affected
2nd degree relatives with breast cancer		+	Age <50/≥50	Age for all affected
1st or 2nd degree with ovarian cancer			+	Age for all affected
Bilateral breast cancer				+
Male breast cancer				+
Outcome predicted	Incident breast cancer			BRCA 1/2 mutation

in determining whether a subject has the disease. The gold standard test may not be suitable to apply in all subjects suspected of the disease because it is burdensome (e.g. invasive), or costly.

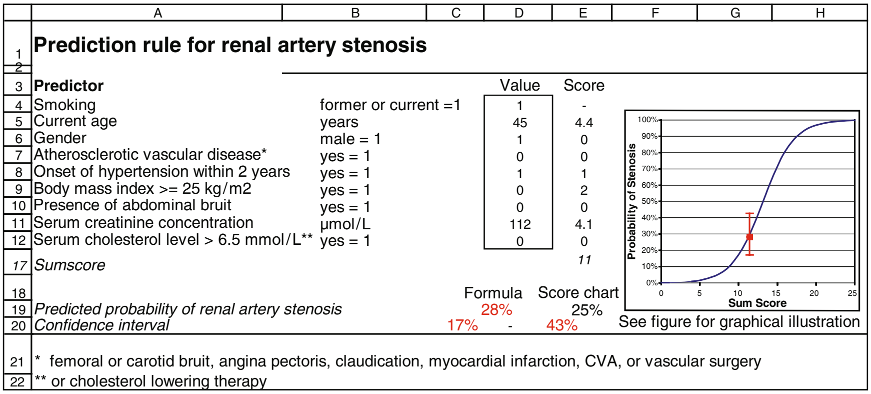
2.3.2 *Example: Predicting Renal Artery Stenosis

Renal artery stenosis is a rare cause of hypertension. The gold standard for diagnosing renal artery stenosis, renal angiography, is invasive and costly. Krijnen et al. aimed to develop a prediction rule for renal artery stenosis from clinical characteristics. The rule might be used to select patients for renal angiography.[243] Logistic regression analysis was performed with data from 477 hypertensive patients who underwent renal angiography. A simplified prediction rule was derived from the regression model for use in clinical practice. Age, sex, atherosclerotic vascular disease, recent onset of hypertension, smoking history, body mass index, presence of an abdominal bruit, serum creatinin concentration, and serum cholesterol level were selected as predictors. The diagnostic accuracy of the regression model was similar to that of renal scintigraphy, which had a sensitivity of 72% and a specificity of 90%. The conclusion was that this clinical prediction rule can help to select patients for renal angiography in an efficient manner by reducing the number of angiographic procedures without the risk for missing many renal artery stenoses. The modelling steps summarized here will be described in more detail in Part II.

An interactive Excel program is available to calculate diagnostic predictions for individual patients. Figure 2.1 shows the example of a 45-year-old male with recent onset of hypertension. He smokes, has no signs of atherosclerotic vascular disease, a BMI<25, no abdominal bruit is heart, serum creatinin is 112 μmol/L, and serum cholesterol is not elevated. According to a score chart (see Chap. 18), the sum score was 11, corresponding to a probability of stenosis of 25%. According to exact logistic regression calculations, the probability was 28% [95% confidence interval 17–43%].

2.3.3 Starting Treatment: the Treatment Threshold

Decision analysis is a method to formally weigh pros and cons of decisions. For starting treatment after diagnostic work-up, a key concept is the treatment threshold. This threshold is defined as the probability where the expected benefit of treatment is equal to the expected benefit of avoiding treatment. If the probability of the diagnosis is lower than the threshold, no treatment is the preferred decision, and if the probability of the diagnosis is above the threshold, treatment is the preferred decision.[325] The threshold is determined by the relative weight of false-negative vs. false-positive decisions. If a false-positive decision is much less important than a false-negative decision, the threshold is low. For example, a 1:100 ratio leads to a 1% threshold. On the other hand, if false-positive decisions confer serious risks, the threshold should be higher. Further details on the threshold concept are beyond the scope of this book, but the issue returns when we discuss the performance of prediction models with decision curves[469] (Chap. 16).



	A	B	C	D	E	F	G	H
1	**Prediction rule for renal artery stenosis**							
2								
3	**Predictor**			Value	Score			
4	Smoking	former or current =1		1	-			
5	Current age	years		45	4.4			
6	Gender	male = 1		1	0			
7	Atherosclerotic vascular disease*	yes = 1		0	0			
8	Onset of hypertension within 2 years	yes = 1		1	1			
9	Body mass index >= 25 kg/m2	yes = 1		0	2			
10	Presence of abdominal bruit	yes = 1		0	0			
11	Serum creatinine concentration	μmol/L		112	4.1			
12	Serum cholesterol level > 6.5 mmol/L**	yes = 1		0	0			
17	*Sumscore*				*11*			
18				Formula	Score chart			
19	*Predicted probability of renal artery stenosis*			28%	25%			
20	*Confidence interval*		17%	-	43%	See figure for graphical illustration		
21	* femoral or carotid bruit, angina pectoris, claudication, myocardial infarction, CVA, or vascular surgery							
22	** or cholesterol lowering therapy							

Fig. 2.1 Prediction rule for renal artery stenosis as implemented in an Excel spreadsheet

Note that a single treatment threshold applies only when all diagnostic work-up is completed, including all available tests for the disease. If more tests can still be done, a more complex decision analysis needs to be performed to determine the optimal choices on tests and treatments. We then have two thresholds: a low threshold between no treatment and further testing; and a higher threshold between further testing and treatment. This concept is illustrated with the diagnosis of deep venous thrombosis using ultrasound.

*2.3.4 Example: Probability of Deep Venous Thrombosis

A systematic review of 54 studies indicated that individual clinical features are of limited value in diagnosing deep venous thrombosis (DVT). Characteristics such as previous DVT, malignant disease, recent immobilization, and recent surgery only modestly increased the probability of DVT.[144] A clinical prediction rule developed by Wells et al. combines nine signs, symptoms and risk factors to categorize patients as having low, moderate or high probability of DVT.[482] This rule stratifies a patient's probability of DVT much better than individual findings.[144]

Patients who are found to be at low pretest probability ("score ≤ 1") can have DVT safely excluded (1) on the basis of a single negative ultrasound result, or (2) a negative plasma D-dimer test. Patients who are at increased pretest probability ("score > 1") require both a negative ultrasound result, and a negative D-dimer test to exclude DVT.[481] A possible diagnostic algorithm is shown in Fig. 2.2.[369]

2.3.5 Intensity of Treatment

Prognostic estimates are also important to guide decision-making once a diagnosis is made. Decisions include, for example, more or less intensive treatment approaches. The framework for decision-making based on prognosis is very similar to that based on diagnostic probabilities as discussed before.

A treatment should only be given to a patient if a substantial gain is expected, which exceeds any risks and side effects (Fig. 2.3). Glasziou and Irwig illustrate this approach with a case study in anticoagulants and risk of atrial fibrillation.[138] Anticoagulants are very effective in reducing the risk of stroke in patients with nonrheumatic atrial fibrillation. However, using these drugs increases the risk of serious bleedings. Hence, the risk of stroke has to outweigh the bleeding risk before treatment is considered.

The specific calculation of the net benefit of a treatment requires various steps:[138]

(1) Estimate benefit and harm: randomized controlled trials (RCTs) may often provide the most reliable source for relative risk estimates for both benefits and harms of treatment.

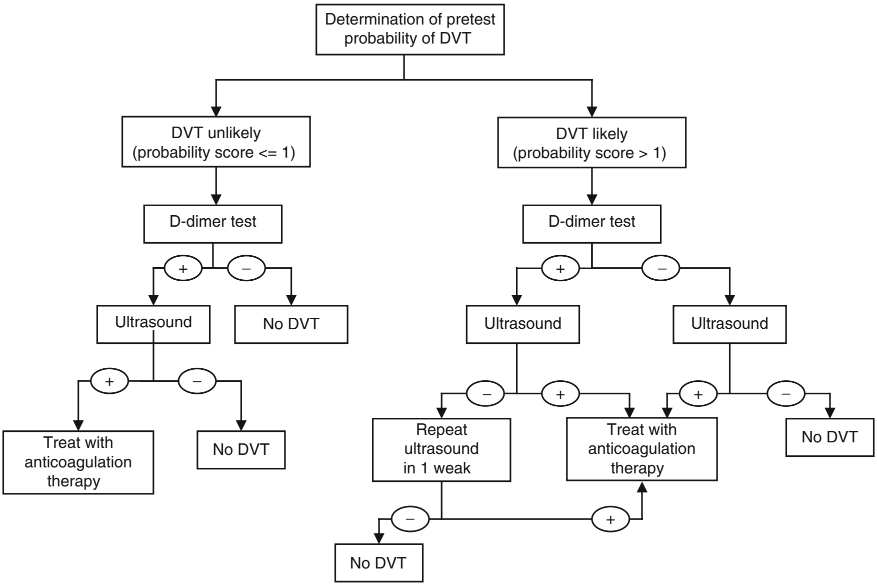

Determination of pretest probability of DVT
DVT unlikely (probability score <= 1)
DVT likely (probability score > 1)
D-dimer test
D-dimer test
+
−
+
−
Ultrasound
No DVT
Ultrasound
Ultrasound
+
−
−
+
+
−
Treat with anticoagulation therapy
No DVT
Repeat ultrasound in 1 weak
Treat with anticoagulation therapy
No DVT
−
+
No DVT

Fig. 2.2 A possible diagnostic algorithm for patients suspected of DVT with D-dimer testing and ultrasound imaging[369]

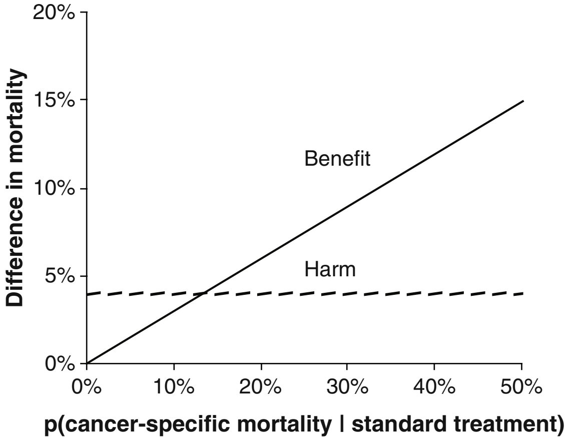


Fig. 2.3 Graphical illustration of weighing benefit and harm of treatment. Benefit of treatment (reduction in absolute risk) increases with cancer-specific mortality (relative risk set to 0.7). Harm of treatment (excess absolute risk, e.g. due to toxicity of treatment) is assumed to be constant at 4%. Net benefit occurs only when the cancer-specific mortality given standard treatment is above the threshold of 11%[451]

(2) Check assumptions of relative benefit and absolute harm: subgroup effects of treatment may exist, both for benefit and harm, which invalidate the simple decision-analytic model in Fig. 2.3.
(3) Weigh up benefit and harm: If the assumptions of relative risk reduction and constant harm are fulfilled the predicted benefit needs to be weighed up against the potential harm. This results in a graph as Fig. 2.3, with actual numbers on the *Y*-axis.
(4) Predict patient's risk: To identify patients who should expect benefit to be greater than harm, we need to predict each patient's risk. Prognostic models are important for this step.

*2.3.6 Example: Defining a Poor Prognosis Subgroup in Cancer

As an example we consider high-dose chemotherapy (HD-CT) as first line treatment to improve survival of patients with non-seminomatous testicular cancer.[451] Several non-randomized trials reported a higher survival for patients treated with HD-CT as first line treatment (including etoposide, ifosfamide, cisplatin) with autologous stem cell support, compared to standard-dose (SD) chemotherapy (including bleomycin, etoposide, cisplatin). However, HD-CT is related to a higher toxicity, both during treatment (e.g. granulocytopenia, anaemia, nausea/vomiting, diarrhoea), shortly after treatment (e.g. pulmonary toxicity), and long after treatment (e.g. leukemia, cardiovascular disease). HD-CT should therefore only be given to patients with a relatively poor prognosis.

We can specify the threshold for such a poor prognosis group by weighing expected benefit against harms. Benefit of HD-CT treatment is the reduction in absolute risk of cancer mortality. Benefit increases linearly with risk of cancer mortality, if we assume that patients with the highest risk have most to gain. Harm is the increase in absolute risk of treatment mortality (e.g. related to toxicity) due to treatment. The level of harm is the same for all patients, assuming that the toxicity of treatment is independent of prognosis. Patients are candidates for more aggressive treatment when their risk of cancer mortality is above the threshold, i.e. when benefit is higher than harm (Fig. 2.3).

2.3.7 Cost-Effectiveness of Treatment

Cost-effectiveness of treatment also directly depends on prognosis. Treatments may not be cost-effective if the gain is small (for patients at low risk), and the costs high (e.g. for all patients the same drug costs are made). For example, statin therapy should only be given to those at increased cardiovascular risk.[157] And more aggressive thrombolysis should only be used in those patients with an acute myocardial infarction (AMI) who are at increased risk of 30-day mortality.[63] Many other examples can be found, where the relative benefit of treatment is assumed to be constant across various risk groups, and the absolute benefit hence increases with higher risk.

Another approach is to search for differential treatment effects among subgroups of patients. The assumption of a fixed relative benefit is then relaxed. Some patients respond well to a certain treatment and others do not. Patient characteristics such as age, or the specific type of disease, may interact with treatment response. Effects of drugs are affected by the drug metabolism, which is, e.g. mediated by cytochrome P450 enzymes and drug transporters (P-glycoprotein).[103] Research in the field of pharmacogenomics aims to further understand the relation between an individual patient's genetic make-up (genotype) and the response to drug treatment, such that response can better be predicted.[45] Cost-effectiveness will vary depending on the likelihood of response to treatment.

2.3.8 Delaying Treatment

In medical practice, prediction models may provide information to patients and their relatives, such that they have realistic expectations of the course of disease. A conservative approach can sometimes be taken, which means that the natural history of the disease is followed. For example, many men may opt for a watchful waiting strategy if a probably unimportant ("indolent") prostate cancer is detected.[227,424] Or women may be reassured on their pregnancy chances if they have relatively favourable characteristics.

**2.3.9 Example: Spontaneous Pregnancy Chances*

Several models have been published for the prediction of spontaneous pregnancy among subfertile couples.[76,111,393] A "synthesis model" was developed for predicting spontaneous conception leading to live birth within 1 year after start of follow-up based on data from three previous studies.[205] This synthesis models hence had a broader empirical basis than the original models. The predictors included readily available characteristics such as the duration of subfertility, women's age, primary or secondary infertility, percentage of motile sperm, and whether the couple was referred by a general practitioner or by a gynaecologist (referral status). The chance of spontaneous pregnancy within 1 year can easily be calculated. First a prognostic index score is calculated. The score corresponds to a probability, which can be read from a graph (Fig. 2.4).

For example, a couple with a 35-year-old woman (7 points), 2-year duration of infertility (3 points), but with one child already (secondary infertility, 0 points), normal sperm motility (0 points), and directly coming to the gynecologist (secondary care couple, 0 points), has a total score of 10 points. This corresponds to a chance of becoming pregnant of 42%.

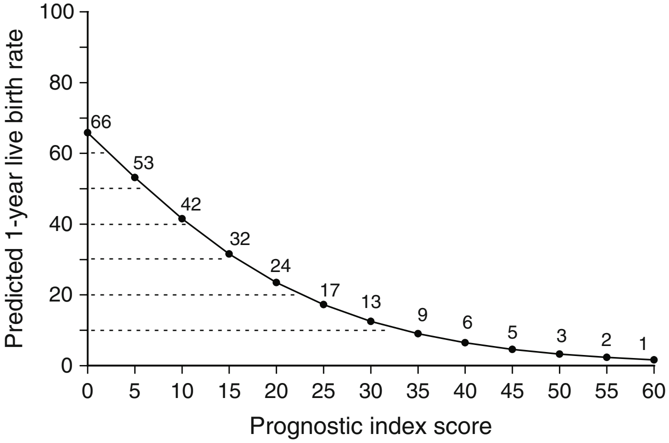


Fig. 2.4 Score chart to estimate the chance of spontaneous pregnancy within 1 year after intake resulting in live birth. *Upper part*: calculating the score; *lower part*: predicting 1-year pregnancy rate.[205] Procedure: circle the subfertility score for each of the variables, transfer to rightmost column and add to get the prognostic index score. Insert the score in the figure below to read off the chance of spontaneous pregnancy within 1 year resulting in live birth

							Subfertility Score
Woman's age (years)	21–25	26–31	32–35	36–37	38–39	40–41	
Score	**0**	**3**	**7**	**10**	**13**	**15**	……..
Duration of subfertility (yrs)	1	2	3–4	5–6	7–8		
Score	**0**	**3**	**7**	**12**	**18**		……..
Type of subfertility	Secondary			Primary			
Score	**0**			**8**			
Motility (%)	≥60	40–59	20–39	0–19			
Score	**0**	**2**	**4**	**6**			……..
Referral status	Secondary care			Tertiary care			
Score	**0**			**4**			……..
				Prognostic index score (Sum)			……..

Most couples who have tried for more than 1 year to become pregnant demand immediate treatment.[205] In their judgment, further waiting is senseless because they consider themselves as infertile. Moreover, the psychological pressure caused by feelings of uncertainty and frustration may increase a desire for immediate action. In addition, most couples overestimate the success of assisted reproduction, such as in vitro fertilization, and underestimate the related risks. The estimations of spontaneous pregnancy leading to live birth can be a tool in advising these couples in the following manner. If the chances are low, e.g. below 20%, there is no point in further waiting, and advising the couple to quickly undergo treatment is realistic. In contrast, if the chances are favourable, e.g. above 40%, the couple should be strongly encouraged to wait for another year, because there is a substantial chance of success.

2.3.10 Surgical Decision-Making

In surgery, it is typical that short-term risks are taken to reduce long-term risks. Short-term risks include both morbidity and mortality. The surgery aims to reduced long-term risks that would occur in the natural history. Acute situations include surgery for trauma, and for acute conditions such as a ruptured aneurysm (a widened artery). Elective surgery is done for many conditions, and even for such planned and well-prepared surgery, the short-term risk and burden are never zero. In oncology, increased surgical risks typically lead to the choice for less risky treatments, e.g. chemotherapy or radiation, or palliative treatments. For example, in many cancers, older patients and those with comorbidity do less often undergo surgery.[6,169,207]

Many prognostic models have been developed to estimate short-term risks of surgery, e.g. 30-day mortality. These models vary in complexity and accuracy. Also, long-term risks have been modeled explicitly for various diseases, although it is often hard to find a suitable group of patients for the natural course of a disease without surgical intervention. As an example, we consider a surgical decision problem on replacement of risky heart valves (Fig. 2.5). Prognostic models were used to estimate surgical mortality, individualized risk of the specific valve, and individual survival.[37,415,449]

**2.3.11 Example: Replacement of Risky Heart Valves*

Björk–Shiley convexo–concave (BScc) mechanical heart valves were withdrawn from the market in 1986 after reports of mechanical failure (outlet strut fracture). Worldwide, approximately 86,000 BScc valves had been implanted by then. Fracture of the outlet strut occurs suddenly and is often lethal.[448] Therefore, prophylactic replacement by another, safer valve, may be considered to avert the risk of fracture. Decision analysis is a useful technique to weigh the long-term loss of life expectancy due to fracture against the short-term surgical mortality risk (Fig. 2.5). The long-term loss of life expectancy due to fracture depends on three aspects:

1. The annual risk of fracture, given that a patient is alive
2. The fatality of a fracture
3. The annual risk of death (survival).

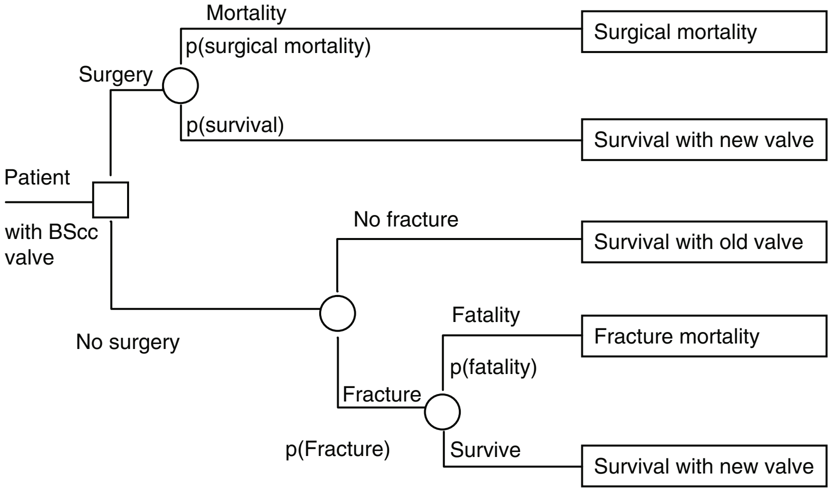


Fig. 2.5 Schematic representation of surgical decision-making on short-term vs. long-term risk in replacement of a risky BScc heart valve. *Square* indicates a decision, *circle* a chance node. Predictions ('p') are needed for four probabilities: surgical mortality, long-term survival, fracture, and fatality of fracture

This long-term loss of life expectancy has to be weighed against the risk of surgical mortality. If the patient survives surgery, the fracture risk is assumed to be reduced to zero. Predictive regression models were developed for each aspect, based on the follow-up experience from 2,263 patients with BScc valves implanted between 1979 and 1985 in The Netherlands.[223,415] We considered 50 fractures that had occurred during follow-up and 883 patients who died (excluding fractures).

The risk of fracture is the key consideration in this decision problem. But the low number of fractures makes predictive modelling challenging, and various variants of models have been proposed. A relatively detailed model included four traditional predictors (age, position (aortic/mitral), type (70° opening angle valves had higher risks than 60° valves), size (larger valves had higher risks)), and two production characteristics.[415] The fatality of a fracture depended on the age of the patient, and the position (higher fatality in aortic position). Survival was related to age, gender, position of the valve, and also to time since implantation. This meant that patients of a given age (e.g. 50 years), had higher risks when the implantation of the valve was longer ago (e.g. implantation at age 35 vs 45 years). Finally, surgical risk was modelled in relation to age and position of the valve. This was a relatively rough approach, since many more predictors are relevant, and a later prediction model was much more detailed.[454]

The results of this decision analysis depended strongly on age: replacement was only indicated for younger patients, who have lower surgical risks, and a higher long-term impact of fracture because of longer survival (Table 2.3). Also, the posi-

Table 2.3 Patient characteristics used in the decision analysis of replacement of risky heart valves[415]

Characteristic	Surgical risk	Survival	Fracture	Fatality fracture
Patient related				
Age (years)	+	+	+	+
Sex (male/female)		+		
Time since implantation (years)		+		
Valve related				
Position (aortic/mitral)	+	+	+	+
Opening angle (60°/70°),			+	
Size (<29 mm or >=29 mm)			+	
Production characteristics			+	
Type of prediction model	Logistic regression	Poisson regression	Poisson regression	Logistic regression

tion of the valve affects all four aspects (surgical risk, survival, fracture, fatality). Before, results were presented as age-thresholds for eight subgroups of valves: by position (aortic/mitral), by type (70°/60°), and by size (large/small).[449] The more recent analysis was so detailed that individualized calculations were necessary, which were performed for all patients who were alive in The Netherlands in 1998. The recommendations from this decision analysis were rather well followed in clinical practice.[455]

2.4 Prediction Models for Medical Research

In medical research, prediction models may serve several purposes. In experimental studies, such as a randomized controlled trial (RCT), predictive baseline characteristics may assist in the inclusion and stratification of patients, and improve the statistical analysis. In observational studies, adequate controlling for confounding factors is essential.

2.4.1 Inclusion and Stratification in an RCT

In a RCT, prognostic estimates may be used for selection of subjects for the study. Traditionally, a set of inclusion and exclusion criteria is applied to define the subjects for the RCT. Some criteria aim to create a more homogeneous group according to expected outcome. Traditionally, all inclusion criteria have to be fulfilled, and none of the exclusion criteria. Alternatively, some prognostic criteria can be combined in a prediction model, with selection based on individualized predictions. This leads to a more refined selection.

Stratification is often advised in RCTs for the main prognostic factors.[18,338,496] In this way, balance is obtained between arms of a trial with respect to baseline prognosis. This may facilitate simple, direct comparisons of treatment results, especially for smaller RCTs, where some imbalance may readily occur. Prediction models may refine stratification of patients, especially when many prognostic factors are known.

*2.4.2 Example: Selection for TBI Trials

As an example, we consider the selection of patients for RCTs in traumatic brain injury (TBI). Patients above 65 years of age and those with non-reacting pupils are often excluded because of a high likelihood of a poor outcome. Indeed we find a higher than 50% mortality at 6-month follow-up in patients fulfilling either criterion (Table 2.4). Hence, we can simply select only those less than 65 years with at least one reacting pupil (Table 2.5, part A). However, we can use a prognostic model for more efficient selection that inclusion based on separate criteria. A simple logistic regression model with "age" and "pupils" can be used to calculate the probability of mortality in a more detailed way. If we aim to exclude those with a predicted risk over 50%, this leads to an age limit of 30 years for those without any pupil reaction, and an age limit of 76 years for those with any pupil reaction (Table 2.5, part B). So, patients under 30 years of age can always be included, and patients between 65 and 75 years can be included if they have at least one reacting pupil (Table 2.5).

Table 2.4 Analysis of outcome in 7,143 patients with severe moderate traumatic brain injury according to reactive pupils and age dichotomized at age 65 years[276]

	>= 1 Reactive pupil		Non-reactive pupils	
	<65	>=65 years	<65	>=65 years
6-month mortality	926/5101 (18%)	159/284 (56%)	849/1644 (52%)	97/114 (85%)

Table 2.5 Selection of patients with two criteria (age and reactive pupils) in a traditional way (A) and according to a prognostic model (probability of 6-month mortality < 50%, B)

		A: Traditional selection		B: Prognostic selection		
		<65	>=65 years	<30	30–75	>=76 years
Pupillary reactivity	No reactivity	Exclude	Exclude	Include	Exclude	Exclude
	>=1 pupil	Include	Exclude	Include	Include	Exclude

2.4.3 Covariate Adjustment in an RCT

Even more important is the role of prognostic baseline characteristics in the analysis of an RCT. The strength of randomization is that comparability is created between treated groups both with respect to observed *and unobserved* baseline characteristics (Fig. 2.6). No systematic confounding can hence occur in RCTs. But some observed baseline characteristics may be strongly predictive of outcome. Adjustment for such covariates has several advantages:[133,182,188,190,339,348]

1. To reduce any distortion in the estimate of treatment effect that occurred by random imbalance between groups
2. To improve the precision of the estimated treatment effect
3. To increase the statistical power for detection of a treatment effect

Remarkably, covariate adjustment works differently for linear regression models and generalized linear models (e.g. logistic, Cox regression, Table 2.6).

1. For randomized clinical trials the randomization guarantees that the bias is zero a priori, both for observed and unobserved baseline characteristics. However, random imbalances may occur, generating questions such as: What would have been the treatment effect had the two groups been perfectly balanced? We may think of this distortion as a bias a posteriori, since it affects interpretation similarly as in observational epidemiological studies.

Regression analysis is an obvious technique to correct for such random imbalances. When no imbalances have occurred for predictors considered in a regression model, the adjusted and unadjusted estimates of the treatment effect would be expected to be the same. This is indeed the case in linear regression analysis. Remarkably, in generalized linear models such as logistic regression, the adjusted and unadjusted estimates of a treatment effect are not the same, even when predictors are

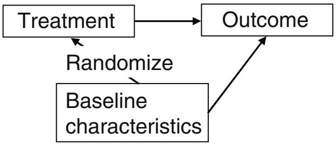


Fig. 2.6 Schematic representation of the role of baseline characteristics in an RCT. By randomization, there is no systematic link between baseline characteristics and treatment. Baseline characteristics are still important, since they are prognostic for the outcome

Table 2.6 Comparison of adjustment for predictors in linear and generalized linear models (e.g. logistic regression) in estimation and testing of treatment effects, when predictors are completely balanced

Method	Effect estimate	Standard error	Power
Linear model	Identical	Decreases	Increases
Generalized linear model	Further from zero	Increases	Increases

completely balanced[133] (see Questions 2.3 and 22.2). Adjusted effects are expected to be further from zero (neutral value, OR further from 1). This phenomenon is referred to as a “stratification effect”, and does not occur with linear regression.[403]

2. With linear regression, adjustment for important predictors leads to an improvement in precision of the estimated treatment effect, since part of the variance in the outcome is explained by the predictors. Contrary, in generalized linear models such as logistic regression, the standard error of the treatment effect always increases with adjustment.[348]
3. In linear regression, adjusted analyses provide more power to the analysis of treatment effect, since the standard error of the treatment effect is smaller. For a generalized linear model such as logistic regression, the effect of adjustment on power is not so straightforward. It has however been proven that the expected value of the treatment effect estimate increases more than the standard error. Hence, the power for detection of a treatment effect is larger in an adjusted logistic regression analysis compared to an unadjusted analysis.[348]

2.4.4 Gain in Power by Covariate Adjustment

The gain in power by covariate adjustment depends on the correlation between the baseline covariates (predictors) and the outcome. For continuous outcomes, this correlation can be indicated by Pearson’s correlation coefficient (r). Pocock et al. showed that in the continuous outcome situation, the sample size can be reduced with $1 - r^2$, to achieve the same statistical power with a covariate adjusted analysis as an unadjusted analysis.[339] A very strong predictor may have r=0.7 (r^2 50%), e.g. a baseline covariate of a repeated measure such as blood pressure, or a questionnaire score. The required number of patients is then roughly halved. The saving is less than 10% for r=0.3 (r^2 9%).[339]

Similar results have been obtained in empirical evaluations with dichotomous outcomes, where Nagelkerke’s R^2 [309] was used to express the correlation between predictor and outcome.[188,190,403] The reduction in sample size was slightly less than

Table 2.7 Illustration of reduction in sample size with adjustment for baseline covariates with dichotomous outcomes

Application area	Correlation baseline–outcome	Reduction in sample size
Acute MI: 30-day mortality[403]		
Age adjustment	R^2 13%	12%
17 predictor adjustment	R^2 25%	19%
Traumatic brain injury: 6-month mortality[189]		
3 predictor model	R^2 30%	25%
7 predictor model	R^2 40%	30%

1 – R^2 in simulations for mortality among acute MI patients[403] and among TBI patients[189] (Table 2.7).

*2.4.5 Example: Analysis of the GUSTO-III Trial

The GUSTO-III trial considered patients with an acute myocardial infarction.[4] The outcome was 30-day mortality. The protocol pre-specified a prognostic model for the primary analysis of the treatment effect. This model combined age, systolic blood pressure, Killip class, heart rate, infarct location, and age-by-Killip-class interaction. These predictors were previously found to comprise 90% of the predictive information of a more complex model for 30-day mortality in the GUSTO-I trial.[255] A review of RCTs published in the major medical journals after the year 2000 shows that covariate adjustment is used in approximately 50% of the cases.[339]

2.4.6 Prediction Models and Observational Studies

Confounding is the major concern in epidemiological analyses of observational studies. When treatments are compared, groups are often quite different because of a lack of randomization. Subjects with specific characteristics are more likely to have received a certain treatment than other subjects ("indication bias", Fig. 2.7). If these characteristics also affect the outcome, a direct comparison of treatments is biased, and may merely reflect the lack of initial comparability ("confounding"). Instead of treatment, many other factors can be investigated for their causal effects. Often, randomization is not possible, and observational studies are the only possible design. Dealing with confounding is an essential step in such analyses.

Fig. 2.7 Schematic representation of confounding in an observational study. Baseline characteristics act as confounders since they are related to the treatment and to the outcome

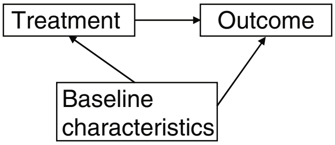
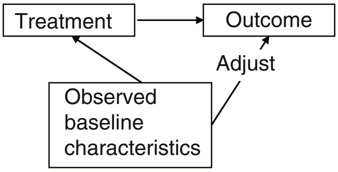


Fig. 2.8 Schematic representation of adjustment for baseline characteristics in an observational study. By adjustment, we aim to correct for the systematic link between observed baseline characteristics and outcome, hence answering the question what the treatment effect would be if observed baseline characteristics were similar between treatment groups



Regression analysis is a commonly used method to control for imbalances between treatment groups, e.g. with logistic or Cox regression.[235] Many baseline characteristics can be simultaneously adjusted for (Fig. 2.8). Similarly, regression analysis can be used to control for confounders in aetiologic research.

2.4.7 Propensity Scores

A problem arises when the outcome is relatively rare. Constructing a regression model with many predictors is then problematic. This may lead to biased and inefficient estimates of the difference between groups in the adjusted analysis.[66] An alternative in the setting of rare outcomes is to use a propensity score.[55] The propensity score defines the probability that a subject receives a particular treatment ("Tx") given a set of confounders: p(Tx | confounders). For calculation of the propensity score, the confounders are usually used in a logistic regression model to predict the treatment, without including the outcome.[60,359] The propensity score is subsequently used in a second stage as a summary confounder (Fig. 2.9). Approaches in this second stage are matching on propensity score, stratification of propensity score (usually by quantile), and inclusion of the propensity score with treatment in a regression model for the outcome.[89]

Empirical comparisons provided no indication of superiority of propensity score methods over conventional regression analysis for confounder adjustment.[381,429] Simulation studies however suggest a benefit of propensity scores in the situation of few outcomes relatively to the number of confounding variables.[66]

**2.4.8 Example: Statin Treatment Effects*

Seeger et al. investigated the effect of statins on the occurrence of acute myocardial infarction (AMI).[378] They studied members of a Community Health Plan with a recorded LDL>130 mg dl^{-1} at any time between 1994 and 1998. Members who initiated therapy with a statin were matched using propensity scores to members who did not initiate statin therapy. The propensity score predicted the probability of sta-

Fig. 2.9 Schematic representation of propensity score adjustment for baseline characteristics in an observational study. The propensity score estimates the probability of receiving treatment. By subsequent adjustment for the propensity score, we mimic an RCT, since we removed the systematic link between baseline characteristics and treatment. We can however only include observed baseline characteristics, and have no control over unobserved characteristics

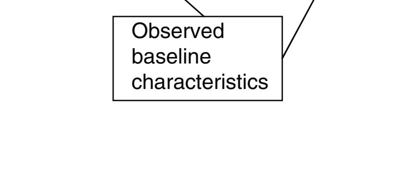

Treatment
Outcome
Propensity score adjustment
Observed baseline characteristics

Table 2.8 The effect of statins on the occurrence of acute myocardial infarction[378]

	Confounders	*N* with AMI	HR [95% CI]
Unadjusted	–	325 vs. 124	2.1 [1.5–3.0]
Propensity score adjusted	52 main effects, 6 quadratic terms	77 vs. 114	0.69 [0.52–0.93]

tin initiation. Scores were estimated using a logistic regression model that included 52 variables and 6 quadratic terms (Table 2.8). Statin initiators were matched to a noninitiator within a 0.01 caliper of propensity. Initiators for whom no suitable noninitiator could be found were excluded, leaving 2,901 matched initiators out of 4,144 initiators (70%). The 4,144 statin initiators had a higher prevalence of established coronary heart disease risk factors than did unmatched noninitiators. The follow-up of these unmatched cohorts identified 325 AMIs in the statin initiator group and 124 in the noninitiator group (hazard ratio 2.1, 95% confidence interval 1.5–3.0). The propensity score-matched cohorts (2 × *n*=2,901) were very similar with respect to 51 of the 52 baseline characteristics. There were 77 cases of AMI in statin initiators compared with 114 in matched non-initiators (hazard ratio 0.69, 95% confidence interval 0.52–0.93). The authors hence conclude that statin use in the members of this Community Health Plan was beneficial on the occurrence of AMI, but warn that predictors that are not part of the model may remain unbalanced between propensity score matched cohorts, leading to residual confounding.

2.4.9 Provider Profiling

Another area of application of prediction models is in the comparison of outcomes from different hospitals (or other providers of care, "provider profiling").[47] The quality of health care providers is being compared by their outcomes, which are considered as performance indicators. Simple comparisons between providers may obviously be biased by differences in case-mix; for example, academic centers may see more severe patients, which accounts for poorer outcome on average. Prediction models are useful for case-mix adjustment in such comparisons.

**2.4.10 Example: Ranking Cardiac Outcome*

New York State was among the first to publicly release rankings of outcome of coronary artery bypass surgery by surgeon and hospital. Such cardiac surgery report cards have been criticized because of their methodology.[136] Adequate risk adjustment is nowadays better possible with sophisticated prediction models. An example is a model published by Krumholz et al., who present a prediction model for 30-day mortality rates among patients with AMI.[245] The model used information from administrative claims and aimed to support profiling of hospital performance. They analyzed 140,120 cases discharged from 4,664 hospitals in 1998. They compared the model from claims data with a model using medical record data and found high agreement. They also found adequate stability over time (data from years 1995 to 2001). The final model included 27 variables and had an area under the receiver operating characteristic curve of 0.71. The authors conclude that this administrative claims-based model is as adequate for profiling hospitals as a medical record model. Chapter 21 provides a more in-depth discussion of this research area.

2.5 Concluding Remarks

We have discussed several areas of potential application of prediction models, including public health (targeting of preventive interventions), clinical practice (diagnostic work-up, therapeutic decision making), and research (design and analysis of RCTs, confounder adjustment in observational studies). More types of application can probably be thought of. Obtaining predictions from a model has to be separated from obtaining insights in the disease mechanisms and patho-physiological processes. Such insights are related to the estimated effects of predictors in a model. Often, prediction models serve the latter purpose too, but the primary aim considered in this book is outcome prediction.

Questions

2.1 Examples of applications of prediction models

(a) What is a recent application of a prediction model that you encountered? Search PubMed [http://www.ncbi.nlm.nih.gov/sites/entrez] if nothing comes to mind.

(b) How could you use a prediction model in your own research or in your clinical practice?

2.2 Cost-effectiveness

How could prediction models contribute to targeting of treatment and to increasing cost-effectiveness of medical care?

2.3 Covariate adjustment in an RCT

What are the purposes of covariate adjustment in an RCT? Explain and distinguish between logistic and linear regression.

2.4 Propensity score

(a) What is the definition of a propensity score?

(b) Explain the difference between adjustment for confounders through regression analysis and through a propensity score.

(c) When is propensity score specifically appropriate? See papers by Braiman and by Cepeda.[55,66]