

# A Review of Statistical Approaches for the Analysis of Data in Rheumatology

Emmanuel Lesaffre and Jolanda Luime

## Introduction

Too often statistics is regarded as a set of rules, even recipes, to end up in a final *P*-value or at best a confidence interval. Good Statistical Practice (GSP) is much more than (even correctly) applying a bunch of statistical tests. In fact, it involves the whole research process from posing the appropriate research questions to writing up the results and drawing the appropriate conclusions. The order in which the statistical techniques are discussed in this chapter somewhat reflects how the statistical analysis of comparative studies is done. We start with descriptive statistics, look at methods to compare two or more treatments, and then discuss correlation and regression techniques to finally review methods for the analysis of follow-up studies. We also briefly discuss multivariate statistical approaches. In addition, we draw the attention to possible pitfalls of the discussed methods to provide guidance in analyzing data. While no clear-cut recipes for GSP can be expected, we hope that this chapter helps the reader in preparing a well-motivated statistical plan and analysis.

The appropriate choice of statistical analysis depends on many factors, such as the type of measurement (continuous, categorical, count, etc.), the research question (comparison of two groups, establishing relationship between one measurement and

---

E. Lesaffre, Dr. Sc. (✉)

Department of Biostatistics, Erasmus MC, Dr. Molewaterplein, 50-60,  
Rotterdam 3015 GE, The Netherlands

L-Biostat, KU Leuven, Leuven, Belgium

e-mail: [e.lesaffre@erasmusmc.nl](mailto:e.lesaffre@erasmusmc.nl)

J. Luime, PhD

Department of Rheumatology, Erasmus MC, P.O. Box 2040, Rotterdam 3000 CA,  
The Netherlands

e-mail: [j.luime@erasmusmc.nl](mailto:j.luime@erasmusmc.nl)

other measurements, etc.), the size of the study, the presence and amount of missing data, outliers in the study, etc. These aspects will be discussed in this chapter.

The statistical techniques are illustrated using the data from two rheumatoid arthritis studies conducted in Erasmus MC in Rotterdam, i.e., the RAPPORT and the tREACH study. Besides these two data sets, fictive data (inspired by the above two data sets) were generated to illustrate some statistical concepts. We used here exclusively the freely available R software [1]; however, software packages like SAS® [2], SPSS® [3], etc. could have also been used. For more elaborate texts on statistics (and some more technical details), many handbooks in the literature can be consulted, e.g., [4] and [5].

## Data Sets

### *RAPPORT Study*

The Rheumatoid Arthritis Patients rePort Onset Reactivation study (*RAPPORT study*) [6] was a longitudinal study that aimed to identify an increase in disease activity by self-reported questionnaires in the 3 months preceding the clinical assessment. In this study, 159 patients aged 18 years and older with rheumatoid arthritis (RA) or polyarthritis using disease-modifying antirheumatic drugs (DMARDs) for at least 3 months were recruited. Patient disease activities were evaluated using the Disease Activity Score of 28 joints (DAS28) every 3 months as part of their standard care by a rheumatologist at the clinic. The DAS28 is a composite index [7, 8], which varies between 0 and 10, built up from swollen joint count, tender joint count, a visual analog scale of the patient's assessment of general health, and erythrocyte sedimentation rate at the first hour. A higher score of DAS28 indicates a higher disease activity. Treatment was recommended to be intensified when  $\text{DAS28} > 3.2$  and may be tapered down at  $\text{DAS28} < 2.6$ .

In addition, the self-reported instruments consisting of Health Assessment Questionnaires (HAQ), Rheumatoid Arthritis Disease Activity Index (RADAI), a visual analog scale of the patient's global assessment of disease activity (VAS global), and a visual analog scale for fatigue (VAS fatigue), were measured using a web-based form producing patient-reported outcomes (PROs). The HAQ contains eight dimensions of daily functional activities such as dressing, rising, eating, walking, hygiene, reach, and grip, and is scored from 0 to 3 on a Likert scale with 3 corresponding to the worst condition [9]. Further, the RADAI measures the self-reported disease activity and is composed of five items, each varying between 0 and 10 (for items 4 and 5, see [10]): (1) global disease activity during the previous month, (2) disease activity in terms of swollen and tender joints throughout the day, (3) amount of arthritis pain throughout the day, (4) morning stiffness, and (5) self-assessed tender joints. The VAS global was used to estimate the patient's assessment for general health. Note that the VAS global is also a part of the DAS28. Finally, the VAS fatigue measures the severity of the patient's fatigue over the previous week by a similar VAS scale [11].

## *tREACH Study*

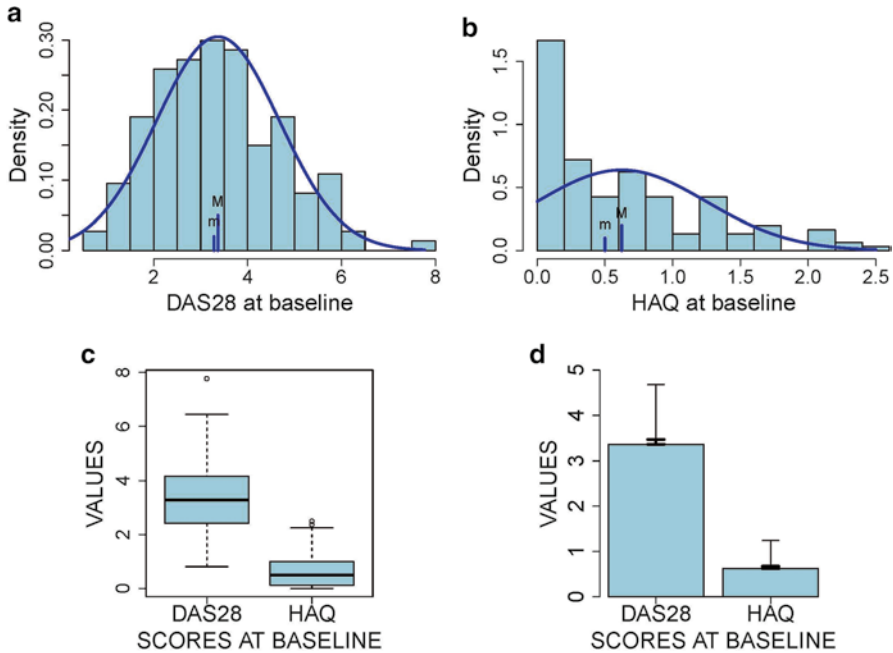
The tREACH is a trial within the Rotterdam Early Arthritis CoHort (tREACH) [12, 13]. It is a multicenter, stratified single-blinded trial conducted in eight rheumatology centers in the Netherlands. RA patients older than 18 years, with arthritis in  $\geq 1$  joint(s), and symptoms less than 1 year are included. Eligible patients were stratified into three strata (low, intermediate, and high) according to their likelihood of progressing to persistent arthritis (i.e., RA) based on the Visser prediction rule [14], a precursor of the new ACR/EULAR RA 2010 classification criteria [15]. For this chapter we use patients with a high risk who were randomized into one of the following initial treatment strategies: (1) triple DMARD therapy (MTX, sulphasalazine, and hydroxychloroquine with GCs intramuscular), (2) triple DMARD therapy with an oral GC tapering scheme, and (3) MTX with oral GCs as in strategy 2 [12].

## Describing the Collected Data

An essential first step in any empirical research is to describe the collected data with numerical values and/or graphical displays. The way this is done depends on the type of data. We consider here: categorical data (ordinal and nominal), counts, and continuous data. *Categorical* data, like adverse events in a drug trial, are typically summarized in tables with frequencies (and proportions or percentages) of each possible outcome as entries. A *bar chart* with these entries displayed as the heights of bars is a common graphical display for such data. When there is ordering in the values of the categorical variable, e.g., severity of the adverse event, one speaks of an *ordinal* variable. For a *nominal* variable, values are not ordered, e.g., for a particular type of adverse event. A special case of a categorical variable is a *binary* or *dichotomous* variable, where there are only two possible values. An example is gender and for this the nominal variable “1” could stand for men and “2” for women. In the RAPPORT study, there are 121 (76 %) men and 38 (24 %) women.

For variables with at least an ordinal character but with too many different values (e.g., DAS28), counts, and *continuous* variables (e.g., weight), the *histogram* provides a better way to graphically summarize the distribution of the data. Now the X-axis is split up into (often equally sized) intervals, and in each interval, the frequency (proportion/percentage) of values is represented as a bar (in another version the area of the bar represents frequencies/proportions/percentages). The histogram not only shows the spread of the collected data but can also spot *outlying values*, which are values that are located remotely from the bulk of the data. Figure 1a, b show the histograms of baseline DAS28 and HAQ values of the RAPPORT study, respectively. The histogram of DAS28 is (roughly) symmetric around its *mean*,

defined as  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$  where  $X_i$  represents here the DAS28 value for the  $i$ th patient and  $n = 159$  is the size of the RAPPORT study. The symbol  $\sum_{i=1}^n$  signifies that the sum



**Fig. 1** RAPPORT study: (a) histogram of DAS28 at baseline together with the best fitting normal distribution, (b) histogram of HAQ at baseline together with the best fitting normal distribution, (c) box plots of DAS28 and HAQ at baseline, and (d) error bar plots of DAS28 and HAQ at baseline. The histograms have the property that the total area of the bars is equal to one.  $M$  represents the mean and  $m$  represents the median. In the error bar plots, the longest bars have length equal to the standard deviation; the shortest bars represent the SEM

is taken of all 159 patients of the RAPPORT study. In contrast, HAQ has a right-skewed distribution (with a right tail). The *median* value corresponds to the value such that 50 % of the observations are left to it; it is also referred to as the 50%-ile and denoted as  $Q_{50}$ . While the median can always be interpreted as a central value, this is not necessarily the case for the mean value, see, e.g., Fig. 1b for HAQ. The spread of the collected data around a central value can be expressed in various ways. The *standard deviation* (denoted as  $s$  or  $SD$ ) is the square root of the *variance*  $s^2$ ,

which is equal to the average squared deviation of the data from their mean, i.e.,  $s^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ . The SD has the advantage over the variance that it is expressed in the original units of the data. For example, since blood pressure is measured in mmHg, the variance is expressed in mmHg<sup>2</sup>, while the SD is also expressed in mmHg.

When the SD is greater in treatment arm A than in arm B, we conclude that the variability of the data must be greater in A than in B. However, apart from this interpretation, it is not immediately obvious how the SD relates to the spread of the data. An alternative, easier-to-interpret, measure is the *interquartile range* (*IQR*), defined

as  $Q_{75} - Q_{25}$ , where  $Q_{25}$  is the 25 %-ile and  $Q_{75}$  is the 75 %-ile of the data. The IQR is therefore easily understood as the length of the central interval that contains 50 % of the data. The mean and median are called *summary statistics for location*, while SD and IQR summarize the *variability* of the data. The mean/median (SD/IQR) for DAS28 in Fig. 1a is 3.37/3.28 (SD=1.31/ IQR=3.28 – 2.42=0.86), while for HAQ (Fig. 1b) we have 0.62/0.50 (SD=0.63/IQR=1.00 – 0.125=0.875).

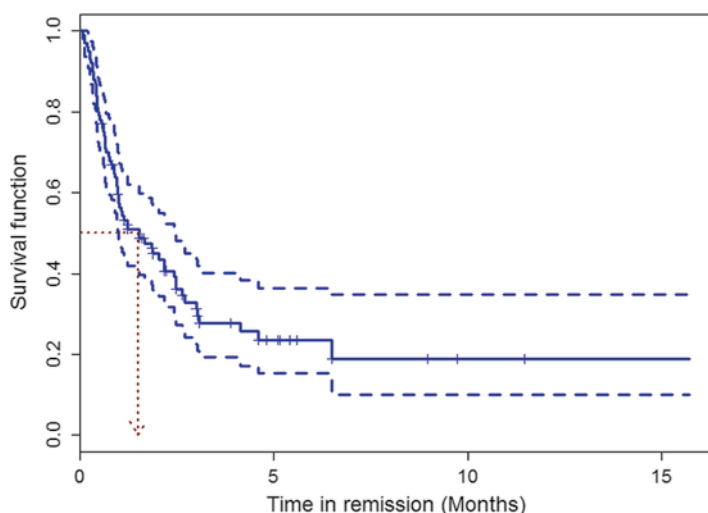
It is customary to summarize continuous data in medical publications with *mean* ± SD. While not always understood, this tradition stems from assuming that the interval [mean-SD, mean+SD] contains 68 % of the central data. However, this interpretation only holds when the histogram can be well approximated by the *Gauss curve or distribution*. The Gauss distribution, also called the “normal” distribution, is the most used distribution in mathematics and statistics, see, e.g., [4, 5]. It reflects the stochastic behavior of a random measure that is the result of the sum of many independent causative factors. Typical measurements that have a normal distribution in a general population are height and weight. For DAS28 the interval [3.37 – 1.31, 3.37 + 1.31] contains indeed 68 % of the central data. The 68 % CI for HAQ, equal to [0.62 – 0.63, 0.62 + 0.63], contains about 80 % of the data, but more importantly it contains negative values, which is clearly nonsense. In Fig. 1, we observe that the Gauss curve approximates well the histogram of DAS28, but not for HAQ.

The *error bar plot* is a popular way to graphically represent the characteristics of the data. In Fig. 1d we show this plot for DAS28 and HAQ. The height of the rectangle is equal to the mean, while the bar emanating from it has length equal to the SD. Hence, this plot graphically displays the interval [mean, mean+SD]. While popular, this plot cannot reveal a possibly skewed distribution of the data. An alternative graph is the *box (–whisker) plot* shown in Fig. 1c. The edges of the box represent  $Q_{25}$  (lower edge) and  $Q_{75}$  (upper edge); the horizontal line represents the median. The lines emanating from the box are called *whiskers*. The whiskers give a graphical impression of the skewness of the distribution. The dots indicate *outlying values*. The definition of the whiskers and outliers depend on the software (here, R).

*Time-to-event data* express the time until the event of interest occurs. This event includes, besides death, also nonterminal events such as remission (DAS < 1.6) in an RA study, a cardiac event in a cardiology study, caries in a dental study, etc. Another term for such data is *survival time*. Typically, survival times have a (right) skewed distribution, and hence the median (and IQR) is here preferred over the mean (and SD). However, most often the exact survival time is not known but is *censored*. A survival time is *right censored* when it is only known that the event hasn’t happened during the conduct of the study. *Left censoring* occurs when event happened before the patient entered the study. This may occur in retrospective studies but such patients are excluded in cohort studies where an association between a risk factor and the event is examined. *Interval censoring* is relatively common in clinical studies. A survival time is interval censored when it is only known that the event has occurred between two examinations. In this chapter we consider only right censoring. There are various reasons for (right) censoring. For instance, when patients are recruited rather late in the study, the probability is low that they will experience the event. Other reasons are: a patient leaves the study prior to experiencing the event because

he changed medication, because of adverse events, because the patient died, etc. It is important to mention that the time at which censoring occurs must not be correlated with the survival time. For instance, removing patients from the study immediately prior to experiencing an event will bias the results and the conclusions of every survival analysis applied to the time-to-event data.

Classical statistical descriptive and inference techniques are not appropriate for survival data. Indeed, for a right-censored survival time, the true survival is only known to be greater than its recorded value. Hence, the mean (median, SD, histogram, box plot, etc.) of the recorded (censored) survival times cannot provide a good estimate of the mean (median, etc.) of the true survival times. Indeed, dedicated techniques are needed in such a case, such as the *Kaplan–Meier curve*. This curve is a proper estimate of the distribution of the true survival times, called the *survivor function*. In Fig. 2, we show the Kaplan–Meier curve of a fictive RA study where RA patients were followed up from the first time they were in remission until their DAS increased above 1.6. This curve shows for each possible survival time (less than the maximum observed time) the estimated proportion of subjects in remission. The Kaplan–Meier curve provides also an estimate of the median survival time, which here is 1.5 months; see Fig. 2. However, the Kaplan–Meier curve cannot provide other descriptive statistics such as the mean survival and its SD. Note that, in the fictive RA study, we assumed right censoring, while in practice, interval censoring certainly would apply since DAS needs to be determined at examination times by the treating rheumatologist.



**Fig. 2** Fictive study: Kaplan–Meier curve that estimates for each time point the proportion of patients that are still in remission. The symbol “+” indicates when the “survival” time is right censored. The arrow points to the estimated median “survival” time. The dashed lines correspond to the 95 % CIs at each observed time of “death”

## Introduction to Statistical Inference

### *The Sample and the Population*

The goal of drug research is to establish the effect of an experimental medication for all possible eligible patients. These patients constitute the *population*. The population of subjects for whom the drug may apply is to some extent artificial since some of the eligible patients may not have been born yet. For this reason, but also due to practical (financial, time constraints, etc.) considerations, it is almost never possible to examine the whole population of interest, and one must confine to a limited set of subjects.

When the sample is taken in a random manner from the population, probability laws can tell us how the sample characteristics vary around the population characteristics. For instance, when a new DMARD reduces DAS28 after 3 months on the average with 0.5, then this average will fluctuate from study to study around its true mean  $\mu$  ( $=0.5$ ). The variability in the study mean is expressed by the *standard error of the mean* (SEM). It is in principle impossible to know SEM since studies are never repeated in exactly the same way. However, from probability laws, we know that it can be estimated from a single study using the formula  $SEM = \frac{s}{\sqrt{n}}$ . This

formula shows that when the patient population is homogeneous (small variance) and/or the study is large, there is little variation of the sample mean around the true mean and then taking the study mean for the population mean will not induce a great error. One can also guess the distance between the true and sample mean with the *confidence interval* (CI). Namely, the 95 % confidence interval given by  $[\bar{X} - 2 \times SEM, \bar{X} + 2 \times SEM]$  contains the true mean with 0.95 probability. Note that the coefficient “2” in the above expression is approximate and varies with the study size, as seen later. Thus, the smaller the 95 % CI, the more precise statement we can make about the true mean. For the RAPPORT study, the SEM of the mean DAS28 at baseline is equal to  $\frac{1.31}{\sqrt{147}} = 0.11$  (for some patients, DAS28 is missing), yielding a 95 % CI equal to [3.15, 3.58]. This implies that we are not sure about the true mean of DAS28, but we believe with 95 % certainty that it is greater than 3.15 and smaller than 3.58. For HAQ at baseline, we obtained an  $SEM = \frac{0.63}{\sqrt{153}} = 0.05$  and the 95 % CI now becomes [0.52, 0.72]. Bars have been added in Fig. 1 that represent the SEM.

Finally, note that the 95 % CI is most popular, but confidence intervals of any size can be determined. In fact, occasionally one reports the 90 % CI or the 99 % CI.

The above probability properties hold when the sample is taken from the population by random sampling (simple random sampling or a more sophisticated version) mechanism. This is often not possible but rather a *convenience sample* is taken, as with the RAPPORT study. This is a sample that is obtained by simply collecting the information from (consecutive) patients who are available to the investigator.

The problem with a convenience sample is that it is not obvious how the results can be extrapolated to a well-defined population. A similar problem occurs with randomized clinical trials (see chapter “[The randomized controlled trial: methodological perspectives](#)”).

## ***Basic Tools for Statistical Inference***

Statistical inference is the activity to draw conclusions from subjects examined in an experimental or observational study for use in future similar subjects. For example, in the RAPPORT study, we might be interested to know whether the change in average DAS28 (in a 12 months’ period) differs between men and women. The average difference (DAS28 at 12 months – DAS28 at baseline) for the 26 men for whom both measurements were recorded is equal to 0.10 (so in fact an increase in disease activity was noticed) and it is  $-0.052$  for the 81 women. The difference in averages is not equal to zero. But our interest lies in the difference of means between men and women for the populations from which the RAPPORT patients were taken, i.e., in the difference between  $\mu_{\text{male}}$  and  $\mu_{\text{female}}$ . The 95 % CI of  $\mu_{\text{female}} - \mu_{\text{male}}$ , computed from the patients with a recorded DAS28 value at both examinations, is equal to  $[-0.69, 0.38]$ . This interval expresses what we know about the true difference from the patients in the RAPPORT study. Since this interval includes zero, we cannot rule out a zero difference in the true means and we decide that there is no (strong) evidence of a different mean change in DAS28 after 12 months of treatment between men and women. Suppose now that we wish to know whether the mean age of women in the RAPPORT study is different from that of men. The mean age of the 121 women is 51.5 years, while for the 38 men, it is 58.4 years. Again we compute the 95 % CI of  $\mu_{\text{female}} - \mu_{\text{male}}$ , where  $\mu$  now represents the average age, and obtain  $[-11.60, -2.08]$  (in years). Now the interval excludes zero; hence, we conclude that there is (strong) evidence that on average women are younger than men in the RAPPORT population.

The confidence interval provides a direct way to draw inference from the study to the population. Yet, a more popular and indirect way of inference is based on the *P-value*. When comparing two (unknown true) means,  $\mu_1$  and  $\mu_2$ , one can distinguish two hypotheses:

$$H_0, \mu_1 = \mu_2 \text{ (or } \Delta = \mu_1 - \mu_2 = 0 \text{)} \quad \text{and} \quad H_a, \mu_1 \neq \mu_2 \text{ (or } \Delta = \mu_1 - \mu_2 \neq 0 \text{)}$$

The hypothesis of interest is given by  $H_a$ , called the *alternative hypothesis*. To test this hypothesis, one reasons indirectly and questions whether  $H_0$ , called the *null hypothesis*, can be rejected. This is done via the *P-value*. To establish the *P-value*, one computes the difference of the two observed means and evaluates the extremeness of this difference if  $\Delta = 0$  were true. The *P-value* is the result of a *statistical test* and expresses the probability that the observed difference (or more extreme) could



have been obtained under  $H_0$ . A  $P$ -value is sometimes referred to as a *surprise index*. When the  $P$ -value is small, doubt is raised about  $H_0$  and one is inclined to reject it. Classically a  $P$ -value less than 0.05 or less than 0.01 is considered a value too small to sustain the null hypothesis. When  $P < 0.05$ , one says that the result is *statistically significant at 0.05*; when  $P \geq 0.05$ , a *nonsignificant result* is obtained. The value of 0.05 is called the *significance level of the test* (in statistical handbooks denoted as  $\alpha = 0.05$ ). The significance level needs to be chosen prior to performing the computations. In this chapter we consider only  $\alpha = 0.05$ , which is the most popular choice but there is in principle nothing against choosing  $\alpha = 0.01$  or  $\alpha = 0.10$ , or any other value as long as the significance level is specified prior to performing the test. The average decrease in DAS28 in 1 year's time between men and women corresponds to  $P = 0.57$ , which is not smaller than 0.05, and hence we see no (strong) evidence against  $H_0$ . The conclusion is then that the two groups are not statistically significantly different at 0.05 (often denoted as *NS*). On the other hand, for the comparison of the average age between men and women, we find  $P = 0.0052$ . This result is now statistically significant at 0.05 (often indicated by \*) and we state that  $H_0$  is rejected at 0.05.

The statistical test used above is the *two-sample t-test*, also referred to as the *Student's t-test*. The test consists in computing a standardized difference of the two sample means  $\bar{X}_1$  and  $\bar{X}_2$ , i.e.,  $T = (\bar{X}_1 - \bar{X}_2) / \text{SE}(\bar{X}_1 - \bar{X}_2)$ , whereby  $\text{SE}(\bar{X}_1 - \bar{X}_2)$  is the standard error of the difference in means (similar to the SEM of a single mean). This standardized difference  $T$  is then compared to a reference distribution, here the *t-distribution* with  $(n_1 + n_2 - 2)$  *degrees of freedom* (*df*). This distribution reflects the natural variability of  $T$  under the null hypothesis that  $\Delta = 0$ . The degrees of freedom is a parameter that depends on the sample sizes of the groups and determines the particular *t-distribution*. Note that when  $\text{df} \geq 30$ , the *t-distribution* becomes close to the normal distribution. For the comparison of the change in DAS28 between men and women,  $\text{df} = 26 + 81 - 2 = 105$ . Under the null hypothesis one expects that  $T$  varies around zero, which translates into a statement that under  $H_0$  there is 95 % chance that  $T$  is located between two extreme values roughly equal to  $-2$  and  $2$  (which change with  $\text{df}$ ). Observed  $T$  values outside this central interval thus indicate that the null hypothesis may not be true and correspond to a  $P$ -value smaller than 0.05. For the DAS28 comparison, this interval is equal to  $[-1.983, 1.983]$ . We obtained  $T = -0.577$ , which belongs to the above central interval and therefore  $P > 0.05$ . For the comparison of the mean ages between men and women,  $\text{df} = 157$  and the central interval is now  $[-1.975, 1.975]$ . Since  $T = -2.836$  does not belong to this interval,  $P < 0.05$ .

The two-sample *t-test* is one of the many statistical tests that were developed over the last century to address the various research questions posed in empirical research. Much of this chapter deals with reviewing a variety of statistical tests. A list of popular statistical tests to compare two groups is given in Table 1 and will be further below discussed in section “Statistical tests to compare two groups.”

**Table 1** Overview of classical statistical tests to compare two groups

Type of measurement	Distributional assumptions	Large study	Small study
<i>Unpaired</i>			
Continuous	Normal in each group and = variance	Two sample <i>t</i> -test	Two sample <i>t</i> -test
	Normal in each group and ≠ variance	Welch test	Welch test
Continuous	Not normal and = variance	Two sample <i>t</i> -test	Wilcoxon rank-sum or Mann–Whitney test <sup>a</sup>
	Not normal and ≠ variance	Welch test	
Binary		Chi-square test	Chi-square test + correction
			Fisher’s exact test
<i>Paired</i>			
Continuous	Difference normally distributed	Paired <i>t</i> -test	Paired <i>t</i> -test
	Difference not normally distributed	Paired <i>t</i> -test	Wilcoxon signed-rank test <sup>a</sup>
Binary		McNemar test	McNemar test + correction
			Binomial test

<sup>a</sup>Can also be used for ordinal data

## ***One-Sided and Two-Sided Confidence Intervals and Tests***

The confidence intervals and *P*-values introduced in the previous section are *two sided*. For example, in section “The sample and the population,” we have seen that the 95 % CI of the mean DAS28 at baseline is equal to [3.15, 3.58]. This interval is bounded at both sides and contains with 0.95 probability the true value. Further, there is 0.025 probability that the true value is below 3.51 and 0.025 probability that the true value is greater than 3.58. We could, however, also give a *one-sided* interval like [3.15, infinity]. This interval expresses that there is 97.5 % probability that the true value is above 3.15. Most often, though, a 95 % two-sided interval is reported.

The *P*-values reported in the section “Basic tools for statistical inference” above are also two sided and therefore sometimes denoted as *2P*. When comparing two means, this means that the null hypothesis will be rejected when the standardized difference of means is either too large positively or too large negatively. Often in practice we must be able to reject the null hypothesis for large positive and large negative differences. Let's take the following example from drug research: A drug company is primarily interested to discover whether their drug is working better than the control drug. In other words, the prime interest lies in rejecting a difference in favor of the experimental drug. Suppose that in a large study, the standardized difference is equal to 1.69 (value obtained from standard normal table) in favor of the experimental drug. Since under the null hypothesis of equal treatment effects, 5 % of the studies show a better result than 1.68, the one-sided *P*-value is smaller

than 0.05. But the threshold for two-sided significance is 1.96, and hence at a two-sided level of 0.05, the result is not significant at 0.05. “One-sided” means that we look only in one direction, here in the direction of a better result for the experimental treatment. On the other hand, suppose that the standardized difference is equal to  $-3$ , then the two-sided  $P$ -value is smaller than 0.01 pointing toward a worse effect of the experimental treatment. However, the one-sided  $P$ -value in the direction of a beneficial effect of the experimental treatment is greater than 0.999. While there is no evidence for a significantly better result for the experimental treatment, there is also no evidence for a worse effect with the one-sided test because one looks away from worse experimental results. Therefore, regulatory agencies demand to use two-sided tests (except for non-inferiority tests, see chapter “[The randomized controlled trial: methodological perspectives](#)”).

### ***Type I Error, Type II Error, and the Power of a Test***

The fundamental problem in empirical research is that one is never sure about the truth. In fact, if the truth were known then empirical research is obsolete and statistical inference is not needed. Hence, it is upfront never clear whether the null or the alternative hypothesis is true so that every decision based on observed data is prone to two errors. The *Type I error* represents the error when one concludes that the alternative hypothesis is true (e.g., two treatments have a different effect), while the null hypothesis is in fact true (the true treatments have equal effect). But the researcher may also decide that there is no evidence for the alternative hypothesis, while in fact it represents the truth. In the latter case, a *Type II error* is committed and then the researcher fails to see that the two treatments really differ in efficacy. The Type I error is controlled by the construction of the statistical test. Namely, by choosing a significance level of 0.05, one automatically fixes the probability of the Type I error to 0.05, called *Type I error rate*. However, the probability of the Type II error is not fixed in advance and depends on, among other things, the study size. The probability of not committing the Type II error is known as the *power of the test* and is equal to the probability of finding a clinically relevant difference in the two groups, if it exists. Establishing the sample size to achieve a desirable power is a necessity in randomized controlled trials but is also desirable in explorative studies. Such a computation is, however, quite technical (see chapter “[The randomized controlled trial: methodological perspectives](#)”).

The above reasoning indicates that statistical inference is based on *repeated sampling* ideas. That is, the significance level of 0.05 means that the probability of a Type I error is fixed at 0.05. In other words, (even) if the null hypothesis is true then roughly five out of hundred (independent) statistical tests are significant at 0.05. The practical implication is that, when a large number of statistical tests are performed in a study, say that a few hundred of variables are compared between two groups with about 5 % of them statistically significant at 0.05, then, quite likely, the two groups are not different at all (null hypothesis is probably true). Similarly, the power

is also expressed in terms of repeated samplings. Namely, when the power is 0.80 for a clinically relevant different effect, say  $\Delta_a$ , then we expect in 100 similar studies at least 80 of them with a statistically significant result at 0.05 if the difference is indeed at least  $\Delta_a$ . Finally, the technical definition of the 95 % CI is that in 95 % of the studies set up in the same way as the current study, the true population value is included in the 95 % CI. But for the current study, the true value is inside or outside that interval. This approach of statistical inference, called the *frequentist approach*, is still most popular in clinical research.

In the frequentist approach, the null hypothesis of equality of group means, proportions, etc. can never be demonstrated. Admitted, such a hypothesis never holds in practice (except when two identical treatments are administered). A nonsignificant result must therefore be interpreted as the “absence of evidence against the null hypothesis” possibly due to a too small study size.

The *Bayesian approach* is an increasingly popular statistical approach for inference but based on quite different principles. In this approach, the role of the  $P$ -value is taken over by a probability that the hypothesis of interest is true after having done the experiment, called the *posterior probability*. This probability addresses, in contrast to the  $P$ -value, the research question directly. In section “The Bayesian approach” we will elaborate on this approach.

## ***Choice Between P-Value and Confidence Interval***

The analyses of the RAPPORT study in section “Basic tools for statistical inference” show that zero is inside/outside the 95 % CI of a difference in means when the result is not statistically/statistically significant at 0.05. This is true for most statistical tests. We have:

$P \geq 0.05$  ( $<0.05$ ) if and only if the 95 % CI of the difference does (not) include zero.

The 95 % CI is, however, more informative than the  $P$ -value since it also provides the uncertainty with which the true effect is estimated. With the  $P$ -value, inference is disconnected from the substantive problem and may easily lead to interpretational problems. For instance, there is a long-standing debate in the literature about whether a significant  $P$ -value weighs more in a large rather than in a small study [16]. Major clinical journals like the *NEJM*, the *Lancet*, etc. now require reporting confidence intervals. For instance, the NEJM guidelines for the authors stipulate: “Measures of uncertainty, such as confidence intervals, should be used consistently, including in figures that present aggregated results.” Nevertheless, the  $P$ -value is still here to stand for some time. However, it will probably not be the only basis for statistical inference in the future.

## *Use and Misuse of the P-Value*

The  $P$ -value remains the most used but also the most misused tool for statistical inference. For instance, the  $P$ -value is often misinterpreted as the probability that the posed hypothesis is correct. This, in fact, is the very probability which clearly interests the researcher most. However, it can only be obtained by the Bayesian approach, as will be seen in section “The Bayesian approach.”

Another quite frequent misuse of the  $P$ -value consists in ignoring the increased risk of committing a Type I error when repeatedly testing for significance. This is called the *multiple testing problem*. An example illustrates the problem. An experimental treatment is compared to a control treatment in two different studies, with a  $P$ -value of 0.03 in the first study and a  $P$ -value of 0.06 in the second study, both in favor of the experimental arm. With  $\alpha=0.05$ , there is in each study a risk of 5 % to claim that the two treatments are different while they are in fact equally effective. If better performance of the experimental treatment is concluded when at least one of the studies shows a significant result at 0.05, then the total risk under the null hypothesis of committing a Type I error is about 10 % and not 5 %—what we aimed at!

The *Bonferroni correction* provides an easy but somewhat crude way to deal with the multiple testing problem. For two tests, the Bonferroni correction consists in dividing the significance level by two, i.e.,  $\alpha=0.5/2=0.025$ . Significance in each test is then claimed only if  $P<0.025$ , reducing the overall risk back to approximately 5 %. In our example the treatments cannot be claimed different in efficacy based on Bonferroni’s correction. For  $k$  tests, Bonferroni correction consists in dividing the significance level by  $k$ , i.e.,  $\alpha/k$ . For  $k$  large, it will then become hard to claim any result significant at 0.05. Equivalent to Bonferroni’s correction is multiplying the  $P$ -value with the number of statistical tests, and check whether the product is lower than  $\alpha$  [17]. For example, with 10 tests,  $10 \times P$  must be smaller than 0.05 for a test to be called significant at 0.05. In chapter “[The randomized controlled trial: methodological perspectives](#)”, we will treat more refined ways to correct for multiple testing in controlled clinical trials.

There are several versions of the multiple testing problem. Examples are: two treatments compared in several studies (above example), two treatments compared at several time points or for several variables, more than two treatments compared, etc. In (medical) publications, many statistical tests are often needed to arrive at a sound (clinical) conclusion. Correction for multiple testing may not always be an issue, especially for the exploratory part of the study, as long as one is clear about the nature (exploratory) of the tests. A greater concern is *opportunistic testing*, i.e., searching as long as the tests confirm what you always wanted to prove. This is called *data dredging* and emerges especially with a lot of data but no available scientific theory. Finally, we note that statistical testing does not always make sense. For instance, a significance test that compares the baseline characteristics of treatments in a randomized controlled trial makes no sense since at the start, the treatment groups are by definition sampled from the same population.

## Statistical Tests to Compare Two Groups

### *Factors That Determine the Choice of the Statistical Test*

Table 1 contains common statistical tests to compare two groups of subjects. The choice of the appropriate test depends on many factors and here we consider four factors: (1) paired versus unpaired data, (2) continuous or binary data, (3) small versus large study, and (4) whether distributional assumptions are met or not. Statistical tests for counts are not included in the table since they are often analyzed (after transformation) as continuous data. If needed, the reader can check the statistical literature for more appropriate tests.

Examples of *paired* data are two measurements taken on the same subject at two time points or sometimes measurements recorded on siblings. This comes down to two groups of related data, where one group contains the first measurements and the other group the second measurements. With *unpaired* data, there is no (systematic) relationship between the measurements. Two groups of continuous data are most often compared via the difference in means or via whole distributions, depending whether some distributional assumptions are met or not. Two proportions are compared in different ways, depending on the type of study. With two observed proportions  $p_1$  and  $p_2$ , the *absolute risk reduction*  $AR$  is defined as  $p_1 - p_2$ . In epidemiological research, it is more customary to work with the *relative risk*  $RR = p_2/p_1$  or the *odds ratio*  $OR = \frac{p_2 / (1 - p_2)}{p_1 / (1 - p_1)}$ .

$$OR = \frac{p_2 / (1 - p_2)}{p_1 / (1 - p_1)}.$$

Another factor is the size of the study. However, we must admit that a general definition of a large study is lacking, since it depends on technical aspects of the statistical test. For instance, two groups of 1,000 subjects certainly qualify for a large study to compare two means, but perhaps not when two proportions of rare events are compared.

Furthermore, in applying certain tests, some distributional assumptions need to be met, like that the data should have a normal distribution or that variances should be equal.

That the choice of a statistical test depends on the above (and even other) conditions is purely technical and depends on probability laws developed under the above-specified conditions; see, e.g., [4, 5]. When the aforementioned conditions are fulfilled, the reported  $P$ -value and 95 % CI are correct. But these conditions rarely apply exactly in practice. For instance, data are never exactly normally distributed. Usually *simulation studies* are conducted to determine the operational characteristics of these tests under deviations from these conditions. This gives us a hint of when the reported  $P$ -value and 95 % CI are to be trusted in practice. We say that a statistical test is *robust* against an assumed condition when the reported  $P$ -value is still correct despite this assumption violated by the data; see the section before, and below “Common statistical tests for the comparison of two groups” below for examples.

In addition to the above, still other factors may play a role in choosing a particular test. For instance, if one is concerned about the impact of outliers on the conclusions

of a statistical analysis, a test may be needed that is more robust against such outlying values.

In the next section, we review the statistical tests shown in Table 1. This table can be used as guide when performing simple comparisons between two groups or as a tool to understand better the Materials and Methods part of a clinical paper.

## ***Common Statistical Tests for the Comparison of Two Groups***

### **Continuous Data**

The *t*-test introduced in the section “Basic tools for statistical inference” compares the means of, say, two treatments. This test is appropriate for unpaired data from two groups each having a normal distribution with equal variances. For unequal variances but normal distributions, the *t*-test for unequal variances, also called the *Welch test*, applies. However, the classical *t*-test also works well in this case when the group sizes are about the same, called the *balanced case*. This was discovered via computer simulation studies. The variance of DAS28 at baseline of men and women in the RAPPORT study is equal to 1.50 and 1.17, respectively. Hence, the Welch test seems at its place here, giving  $P=0.54$ , but this is basically the same to what is obtained from the classical *t*-test. Another condition for the unpaired *t*-test is normality in each group. Computer simulations have shown that the *t*-test is robust against non-normality in the balanced case. For extremely skewed distributions, it may be prudent, however, to check the outcome of the *t*-test with a *non-parametric* test. Such a test does not depend on the normality assumption. In fact, for a nonparametric test, the data are replaced by their ranks, and hence the *P*-value from the test becomes independent of the distribution of the data. A popular non-parametric test is the *Wilcoxon rank-sum test*, also called the *Mann–Whitney U test*. A small fictive example illustrates how the test works. Suppose that the DAS28 scores after one year of treatment for group A are 1.0, 1.7, 2.9, and 4.5 and for group B are 2.1, 3.1, 3.3, and 5.9. To compute the Wilcoxon statistic, these scores are ranked irrespective of their group assignment, but their group membership is secured. The ordered values are then 1.0, 1.7, 2.1, 2.9, 3.1, 3.3, 4.5, and 5.9 with the underlined scores pertaining to group B. In the next step, these ordered values are replaced by their ranks 1, 2, 3, 4, 5, 6, 7, and 8, and the ranks pertaining to A are added to give the Wilcoxon rank-sum test statistic  $W=1+3+4+7=15$ . The extremeness of the obtained *W* is established using probability laws with a *P*-value as result. Here  $P=0.484$  demonstrating that there is no evidence that the treatments differ in efficacy after one year. In addition to robustness of deviations from normality, a nonparametric test is less vulnerable to outlying values. A disadvantage of a nonparametric test is that the link with the original data is broken, providing basically only a *P*-value. Note that Wilcoxon rank-sum test can also be used for ordinal data.

Another way to deal with non-normal distributions is to transform the original data such that the transformed data have a normal histogram. The logarithmic function is a popular choice for right-skewed data but may not work when there are a lot of ties in the data. For the HAQ score at baseline, 38 patients have a zero score in the RAPPORT study. Before applying the log transform, we added 1 to the score but then the 38  $\log(\text{HAQ} + 1)$  scores are equal to zero and thus  $\log(\text{HAQ} + 1)$  cannot have a normal distribution. In fact, none of the classical transformations, including the square root, can turn the distribution of HAQ into a normal distribution. Further, in a comparative study, it often happens that a different transformation is needed in each of the groups. In that case, transforming the data is not an option. In addition, an interpretation problem may arise when results are based on transformed data. For instance, when the data are log transformed, the 95 % CI of the difference in the means on log scale translates into a 95 % CI of the ratio of *geometric means* on the original scale. But such a 95 % CI is more difficult to interpret as the geometric mean is not equal to the classical mean.

In the case of paired data, inference is based on the difference between the two related values. A statistical significant result is obtained when the mean difference is remote from zero, taking into account statistical fluctuations under  $H_0$ . The classical statistical test is now the *paired t-test*. This test requires that the difference of the two related values has a normal distribution. If we do not wish to assume this, one could apply the nonparametric *Wilcoxon signed-rank test*, which is now based on the ranks of the differences. This test is also appropriate for ordinal data.

Nonparametric statistical tests can be applied to all studies regardless of their size. For large studies, the *t-test* is also applicable even when the data grossly deviate from normality. This is a consequence of *The Central Limit Theorem*, a key result in statistics which allows working with the original data (of any distribution) for large studies. In practice “large” means in the balanced unpaired case, group sizes of about 20 or more depending on the deviation from normality, but large(r) sample sizes may be needed in the unbalanced case.

## Binary Data

When the outcome of interest is binary, the comparison of two groups involves contrasting two proportions. For unpaired data and a large sample size, the recommended test is the *chi-square test*. This test essentially evaluates a standardized version of the squared difference of the two proportions under the null hypothesis, which is now that the true proportions  $\pi_A$  and  $\pi_B$  are equal. Suppose the observed proportions under treatments A and B are given by  $p_A$  and  $p_B$ , respectively, then the chi-square test computes  $X^2 = (p_A - p_B)^2 / \text{SE}(p_A - p_B)^2$ , with  $\text{SE}(p_A - p_B)$  the standard error of the difference in proportions under  $H_0$ . When  $X^2$  is too large (compared to what is expected under the null hypothesis),  $H_0$  is rejected (at  $\alpha = 0.05$ ). For the actual calculation of the *P-value*, the *chi-square distribution with one degree of freedom* is used as reference distribution. Table 2 represents a  $2 \times 2$  contingency



**Table 2** RAPPORT study: observed and expected frequencies of patients split up according to gender who need a more intensive treatment at month 12

	Observed		Expected	
	DAS28 ≤ 3.2	DAS28 > 3.2	DAS28 ≤ 3.2	DAS28 > 3.2
Men	a = 17	b = 11	A = 14.25	B = 13.75
Women	c = 40	d = 44	C = 42.75	D = 41.25

table contrasting the frequencies of men and women in the RAPPORT study who require step-up treatment (DAS28>3.2). This table is a special case of an  $r \times c$  contingency table when there are  $r$  rows and  $c$  columns in the table. In Table 2 the lower case symbols stand for the observed frequencies, while the upper case symbols refer to the expected frequencies, i.e., those that one would expect on average to happen under the null hypothesis. Comparing the observed with the expected frequencies leads to an equivalent expression of  $X^2$  given by

$$X^2 = \frac{(a - A)^2}{A} + \frac{(b - B)^2}{B} + \frac{(c - C)^2}{C} + \frac{(d - D)^2}{D}.$$

The above expression shows that  $X^2$  will be large when the observed frequencies deviate a lot from the expected frequencies. For the data in Table 2, we obtained  $X^2=1.44$  which corresponds to a P-value of 0.23.

For a small study, the *chi-square test with continuity correction* can be used, but *Fisher’s Exact test* is recommended. Both tests give a more accurate P-value than the chi-square test for a small study. Now “small” is given by the *Cochrane conditions*, which stipulate that the chi-square test may be used when the expected frequencies all exceed 5 (satisfied in our example). The P-value for the Fisher’s Exact test is equal to 0.28.

Instead of applying the chi-square test, which only provides a P-value, one could also compute the 95 % CI of the absolute risk reduction  $AR=p_A-p_B$ , with  $p_A=b/(a+b)$  and  $p_B=d/(c+d)$ . When the 95 % CI of AR does not include 0, the two treatments are statistically significantly different at 0.05. For the relative risk  $RR = \frac{p_B}{p_A}$  and the odds ratio  $OR = \frac{p_B/(1-p_B)}{p_A/(1-p_A)}$ , the value of 1 must not be within the 95 % CI to claim a significant effect. Using the observed frequencies in Table 2, the odds ratio is easily seen to be equal to  $ad/bc$ . For the entries in Table 2, we obtained  $RR=1.28$  with 95 % CI=[0.88, 1.85] and  $OR=1.7$  with 95 % CI=[0.71, 4.06]. Both intervals do include 1 and hence there is no evidence for a difference in the true proportions between men and women.

For paired binary data, a similar reasoning applies but of course the tests must differ. An example of paired proportions is the proportion of patients that have in the RAPPORT study a DAS28 less than 3.2 or greater than 3.2 at baseline (first proportion) versus this proportion at 12 months (second proportion). For a large study, a *McNemar test* is appropriate, which is a variation of the classical chi-square test. For a small study, a corrected version is used or the *binomial test*.

## Survival Times

In section “Describing the collected data,” we have introduced survival data and mentioned that censoring complicated the analysis of such data. Only right censoring is considered here, which means that it is only known that the survival time is greater than the one recorded in the study. Figure 2 shows the Kaplan–Meier estimate (+95 % CI) of the survival function. The Kaplan–Meier curve is a nonparametric estimate, i.e., no assumption is made about the distribution of the true survival times. If one is willing to assume that the survival times have, say, a *Weibull* or a *lognormal distribution*, then estimates of the mean survival time, its SD, etc. can be derived. However, in survival analysis, there is no generally accepted distribution. Therefore, one is reluctant to base inference on a particular parametric assumption.

We will defer statistical inference with survival data to section “Cox regression,” where the Cox proportional hazards (PH) model is introduced. For now, we will limit ourselves by mentioning that the nonparametric tests, such as the Wilcoxon test, have been generalized to survival analysis, as well.

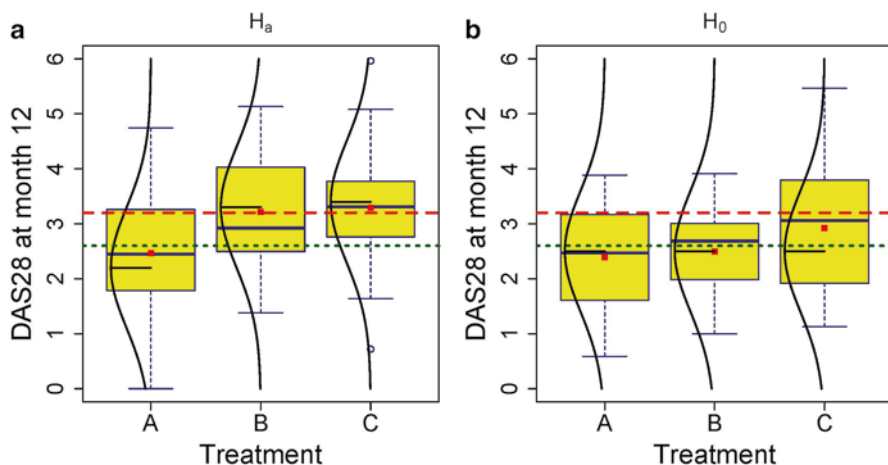
## Statistical Tests to Compare More Than Two Groups

Table 1 is limited to statistical tests for the comparison of two groups. In practice a variety of statistical tests are required to tackle the research questions that pop up in clinical research. In this section we review an extension of some of the techniques seen in section “Statistical tests to compare two groups” to compare more than two groups. We restrict ourselves here to the unpaired case. The paired case involves more complicated statistical techniques suitable for correlated data. Some of these techniques are discussed in section “Models for longitudinal studies.”

### *One-Way Comparisons with Continuous Measurements*

One possibility to compare  $k \geq 2$  groups is to contrast them two by two and perform for each pair a classical unpaired *t*-test. For  $k=5$  groups, this means 10 *t*-tests with each 5 % risk of committing a Type I error. A multiple testing problem arises if no correction (such as Bonferroni) is applied. A popular and better way to control the Type I error rate in this setting is to use an *analysis of variance (ANOVA) test*.

In an ANOVA test, the between-group variance is compared to the variance of the data within the groups. The standardized ratio of these two variances, called the *F-ratio*, should vary around 1 when the null hypothesis of equal means holds. When the alternative hypothesis is true, the *F-ratio* will often be greater than 1. To evaluate whether there is more variability of the group means than expected under  $H_0$ , one computes its extremeness using an *F-distribution* as reference distribution which now has two kinds of degrees of freedom depending on the number of groups and



**Fig. 3** Fictive study: one-way ANOVA with three treatment groups. (a) shows the case of three different true means, while in (b) the three groups have equal true means. The box plots are based on each 25 patients drawn from normal populations shown by the curved lines whereby the horizontal lines point to the true means. The observed means are indicated by squares. The *dashed horizontal line* indicates the threshold above which intensified treatment is needed, while below the *dotted horizontal line* indicates that treatment can be reduced

the group sizes. Two fictive studies illustrate the use of the ANOVA test below. This statistical approach is referred to as *one-way ANOVA*, because there is only a single factor involved in establishing the groups unlike the ANOVA tests reviewed below in section “Two and more way comparisons.”

In both panels of Fig. 3, the DAS28 measurements at month 12 are shown. In each of the two experimental treatments (A and B) and the control treatment (C), 25 patients have been included. All data are fictive and were randomly generated using a computer program. In Fig. 3a, it is seen that the true treatment means (indicated by the normal densities and their associated means) are unequal, i.e., 2.2, 3.3, and 3.4. The true standard deviation is for all groups equal to 1.1. An  $F$ -ratio equal to 4.20 with  $P=0.019$  is obtained. This  $F$ -ratio is judged too high to believe that the true means are equal. Because we have generated the data by ourselves, we know that this is the correct decision. In Fig. 3b, it is seen that the true treatment means are all equal to 2.5 with again  $SD=1.1$ . Now an  $F$ -ratio of 2.03 is obtained yielding a  $P$ -value of  $0.14 \geq 0.05$  and we cannot reject the  $H_0$ , which is again the correct decision.

The ANOVA test assumes normal distributions with equal variances in all groups, but a violation of these assumptions is not dramatic when the group sizes are roughly equal. When there is gross imbalance, one might need to choose for an alternative approach. There is, however, no commonly used test available that generalizes the Welch test. Another possibility is to use the *Kruskal–Wallis test*, which is a generalization of the Wilcoxon rank-sum test and is based on the same ranking principle. Applied to the same fictive data, we obtained the same qualitative conclusions, but different  $P$ -values of course, now equal to 0.029 and 0.24, respectively.

Transformation of the data to normality might sometimes help, but it is in general more difficult to find a transformation that is appropriate for all groups.

The ANOVA  $F$ -test only checks whether there is somewhere a difference between the treatment groups but does not give insight which groups are statistically significantly different. In the literature, pairwise  $t$ -tests are sometimes applied after a significant  $F$ -test, but this may again inflate the Type I error rate. The correct approach is to use so-called multiple comparison tests which penalize the  $P$ -value for multiple testing, i.e., the  $P$ -value is inflated instead of (equivalently) decreasing the significance level; see also above in section “Use and misuse of the  $P$ -value.” There are several types of multiple comparison tests, such as *Newman–Keuls*, *Tukey*, *Dunnnett*, etc. each with some optimality property. To illustrate their use, we take the first fictive example. The pairwise  $t$ -tests without correction for multiplicity result in  $P=0.02$  for treatments A and B, and  $P=0.011$ ,  $P=0.81$  for treatments A and C, B and C, respectively. With the Tukey multiple comparison test, we obtain (1)  $P=0.052$ , (2)  $P=0.028$ , and (3)  $P=0.98$ , respectively. Hence, by correcting for multiplicity, the first two treatments are not statistically significant anymore. For non-parametric tests, only the approximate Bonferroni correction can be applied or more advanced procedures, which are however not yet supported by common software.

### One-Way Comparisons with Categorical Measurements

When DAS28 is categorized into three classes with cutoff points of 2.6 and 3.2, the  $3 \times 3$  contingency Table 3 is obtained. The research question is now whether the probabilities of belonging to the three disease classes differ in the three treatment groups. When the Cochran conditions (section “Binary data”), above are fulfilled, we can apply a chi-square test with now 4 degrees of freedom. As for the  $2 \times 2$  contingency table, this is done by computing  $X^2$ , which is again a comparison of observed with expected frequencies. In general, for an  $r \times c$  contingency table, the degrees of freedom are  $(r - 1) \times (c - 1)$ . For Table 3,  $X^2=5.33$  with  $P=0.26$ . Compare this with the  $P$ -value equal to 0.019 obtained from a one-way ANOVA based on the continuous responses. This illustrates that discretizing continuous variables implies a loss of information and hence a decrease in the power of the study. We note that the chi-square test can also be used to test for an association between a row and a factor in an  $r \times c$  contingency table. For instance, suppose that in Table 3 the row factor is DAS28 categorized at baseline, then a test for difference in percentages is

**Table 3** Fictive study: contingency table of categorized DAS28 at month 12 using 2.4 and 3.2 as cut points (Fig. 3a) together with the row percentages

Treatment	DAS28 $\leq 2.6$	$2.6 < \text{DAS28} \leq 3.2$	DAS28 $> 3.2$
1	13 (52 %)	5 (20 %)	7 (28 %)
2	8 (32 %)	6 (24 %)	11 (44 %)
3	6 (24 %)	5 (20 %)	14 (56 %)

actually a test for association between DAS28 at baseline and at month 12. When the Cochran conditions are not satisfied, then *Exact tests*, which are generalizations of the Fisher's Exact test, are recommended.

In the case of a significant test result in an  $r \times c$  table, there is still the question how this significant result came about. For the chi-square test, a significant result can only be obtained when for one or more cells in the contingency table the observed frequency is remote from the corresponding expected frequency. There are advanced statistical approaches to look for the "significant" deviations in the contingency table, but they are beyond the scope of this chapter. Another, but ad hoc, approach is to collapse classes to create  $2 \times 2$  contingency tables from the  $r \times c$  table and to apply the Bonferroni correction afterward to guard ourselves against multiple testing. For instance, from Table 3 one can construct nine  $2 \times 2$  tables, e.g., left top table with entries 13, 5, 8, and 6 and left bottom table with entries 8, 6, 6, 5, etc. For each of these tables, one can evaluate the association, and the nine  $P$ -values are multiplied with nine. If  $9 \times P < 0.05$  for a particular subtable, then there is a significant association in that subtable. This was not the case here, however.

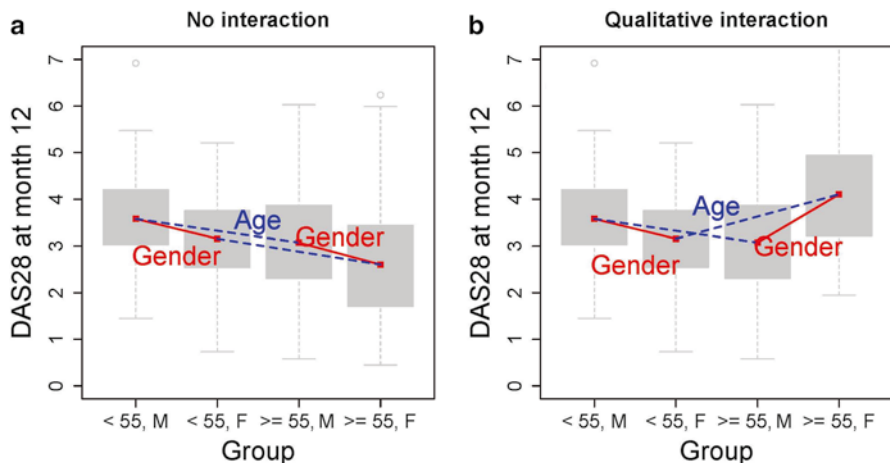
## Two- and More-Way Comparisons

In one-way ANOVA, the different values of one factor determine the groups. When multiple factors are involved, interest may lie in their joint effect on the response. For instance, suppose that RA patients are treated with either a control treatment C or an experimental treatment E (factor 1), but at the same time concomitant medication c or placebo p (factor 2) are administered. Suppose now that an RCT has been set up randomly allocating patients to the four possible combinations: (1) C and p, (2) E and p, (3) C and c, and (4) E and c. Suppose also that one is interested in the overall effect of the experimental treatment on, say, DAS28 but also in the overall effect of the concomitant treatment and additionally in their joint effect. The overall effect of E is called the *main effect* of E and similarly for the overall effect of c. Suppose that after 12 months of treating the patients with E, the average DAS28 is reduced by 1 unit whether or not concomitant medication is administered. In that case, one speaks of *no interaction* between the two factors. If also the concomitant medication reduces the average DAS28, say by 0.5, then in the absence of interaction, the joint effect of the experimental and concomitant treatment results in a decrease of the average DAS28 by  $1 + 0.5 = 1.5$ . A *statistical interaction* between the two factors is present when the joint administration does not result in a sum of the individual main effects. In our fictive example, the experimental treatment always reduces the average of DAS28. The interaction is therefore called *quantitative*. On the other hand, when the joint administration would raise DAS28 on average, then we are dealing with a *qualitative interaction*. For a quantitative interaction, adding the concomitant treatment to the experimental treatment does not change our conclusion about the experimental treatment, whereas for a qualitative interaction we must conclude that joint administration of the two treatments here has a negative impact on the patient.

In the above paragraph, we assumed that we knew the true treatment effects. In practice, they need to be estimated from the study at hand. In a second step, they are tested for equality. This is done by a *two-way ANOVA* analysis and it involves now three *F*-tests, one for each main effect and one for the interaction effect. Each time the null hypothesis corresponds to no effect. Under the assumption of normal distributions with equal variances, the null hypotheses are rejected when the corresponding *F*-values are judged too large under  $H_0$ . Note that we should test first the interaction. If significant, then one explores the main effect of one factor in each level of the other factor. Now follows a fictive example to better explain the practical procedure.

Inspired by the results of the RAPPORT study, we generated 200 DAS28 values (at month 12) from four groups split up according to age less than or more than 55 years (factor 1) and gender (factor 2). Figure 4a shows the generated DAS28 values under no interaction. The *F*-tests (*P*-values) are for (1) age (11.44 ( $P < 0.001$ )), (2) gender (8.15 ( $P = 0.0048$ )), and (3) interaction of age with gender (0.021 ( $P = 0.88$ )). Since there is no evidence for interaction, we can estimate the main effects immediately. They are equal to (95 % CI) for gender ((male–female):  $-0.45$  ( $[-0.75, -0.14]$ )) and for age ( $(\geq 55 - < 55)$ ):  $-0.53$  ( $[-.84, -0.22]$ )). The above 95 % (Tukey) CIs take into account the multiple testing problem. Figure 4b shows the qualitative interaction case. The *F*-value for the interaction is now 21.68 with  $P < 0.001$ . Now it does not make sense to interpret the main effects for gender and age. In fact, we need to estimate the effect of gender in each of the two age classes and the same holds for age.

Two-way ANOVA can be further generalized to involve more than two factors. In general, such data structures are called *factorial designs*. Factorial designs also



**Fig. 4** Fictive study: two-way ANOVA with two factors: age  $< 55$  or  $\geq 55$  and gender. (a) shows the case of no interaction between age and gender, while (b) shows the case of qualitative interaction. The *solid lines* represent the effect of gender in each of the two classes of age, while the *dashed lines* represent the effect of age in the two gender classes

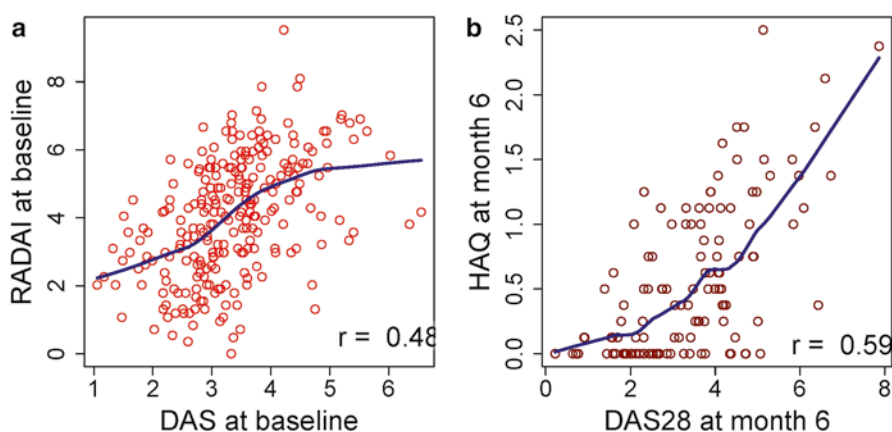
exist for categorical responses, but these will be treated in section “Regression models” where we look at regression models for binary outcomes.

## Measuring and Testing Associations

When two or more measurements are taken on the same subject, it may be of interest to see how much they are related. In this section we consider two situations. In the first case, we look at a general measure for association between two measurements of possibly different nature, so-called correlation measures. In the second case, interest lies in the association of two measurements of the same kind and to know whether they indeed measure the same characteristic. These are called *measures of agreement*. For both cases, continuous, ordinal, and binary data are considered here. We also explain the *Bland–Altman plot* which is a classical tool for continuous measurements often used in medical research to evaluate the dependence of agreement on the level of the measurement.

### Association

In the tREACH study, DAS and RADAI (patient-reported outcome of disease activity) are measured at several time points. One would expect that the two measures are positively related, i.e., we expect that when DAS is high, this will be also for RADAI. In Fig. 5a, we notice that when DAS28 is high (or low), then RADAI tends to be also high (or low). A popular measure to evaluate this association is the *Pearson correlation coefficient*  $r_p$ . The Pearson correlation  $r_p$  is zero, when the two



**Fig. 5** (a) Scatterplot of DAS and RADAI at baseline (tREACH study), (b) scatterplot of DAS and HAQ at month 6 (RAPPORT study). In addition a smooth line representing the relationship is added (using *R* function *lowess*)

measurements show no association. In that case the scatterplot exhibits a circular figure (as a pizza) or in general a figure centered on the horizontal line. For a positive correlation, the two measurements evolve in the same direction, while for a negative correlation the opposite is true. The Pearson correlation between DAS and RADAI is equal to 0.48, which is appreciable but not particularly high. In absolute value, the maximal Pearson correlation is 1. The *coefficient*  $r_p$  is an estimate of the “true” correlation  $\rho$  which would be obtained if we relate all possible DAS and RADAI values obtained from the population from which the sample was taken. A significance test can then determine whether the true correlation is zero or not, i.e., whether  $H_0: \rho=0$ . In addition a 95 % CI for  $\rho$  can be computed. For our example,  $P<0.001$ , indicating that most likely the true correlation is not equal to zero, which is confirmed by the 95 % CI equal to [0.38, 0.57] since it does not include zero.

The Pearson correlation measures the linear relationship between two measurements. If the relationship has a “banana” shape, then the Pearson correlation does not estimate the (nonlinear) association properly. Furthermore, the significance test to evaluate  $H_0: \rho=0$  assumes that both measurements have a normal distribution. The smooth line (see section “Regression models”) in the scatterplot shows an approximate straight line relationship. Thus, for the correct computation of the  $P$ -value associated with a Pearson correlation, both variables should have a normal distribution. In our example the distributions of DAS and RADAI appear to be normal. For non-normal distributions and/or a nonlinear relationship, the *Spearman rank correlation*  $r_s$  is preferred. To compute the Spearman correlation, the original data are first replaced by their ranks, and on these ranks the Pearson correlation is computed. Again, a  $P$ -value for  $\rho=0$  and a 95 % CI can be calculated. Typical for a nonparametric procedure, the Spearman correlation is robust against outlying values. The Spearman correlation between DAS28 and HAQ at month 6 in the RAPPORT study is  $r_s=0.59$  ( $P<0.001$ ). The scatterplot together with the smooth line in Fig. 5b suggests that there is a curvilinear relationship and a skewed distribution of HAQ. The Spearman correlation of DAS28 and HAQ at month 6 equals  $r_s=0.59$ , which is lower than the Pearson correlation of 0.64.

We note that it often does not really make sense to evaluate a correlation with a  $P$ -value. Indeed, often a zero correlation is not expected (e.g., we do expect a non-zero correlation between DAS20 at baseline and DAS28 at month 12), but we rather wish to know the size of the correlation using a 95 % CI.

The Spearman correlation can also be used for ordinal data. For two binary outcomes, several measures have been suggested such as the *tetrachoric correlation*. An alternative measure is the *cross-ratio*, which is in fact equal to the odds ratio but now the two binary variables are interchangeable without harming the interpretation of the association (symmetric case). For the odds ratio one variable is often considered to represent the “cause” and the other the “result” (asymmetric case).

It often happens in an explorative study that many correlations are tested without a clear hypothesis to evaluate. This clearly inflates the Type I error rate tremendously and may lead to wild speculations of possible relations.



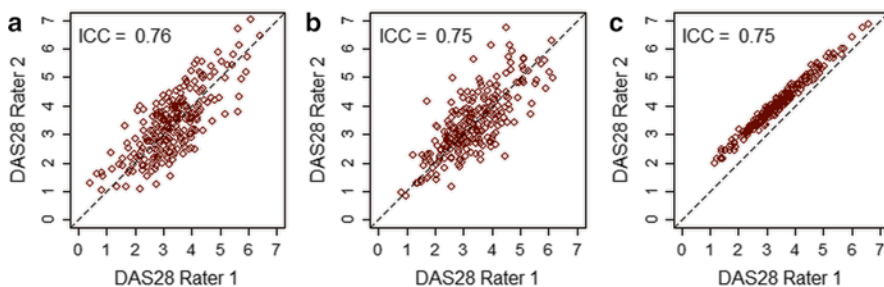
## Agreement

Suppose that we want to measure how reproducible clinicians can score DAS28, both between occasions as well as between clinicians. That is, we wish to know the *intra-rater variability* (between occasions) and the *inter-rater variability* (between clinicians). Both measures of variability can be estimated by the *intra-class correlation (ICC)*. We assume that there are either two clinicians who score DAS28 on RA patients, or that one clinician scores DAS28 twice on each patient. Let  $\sigma_B^2$  represent the variance of the average of the 2 scores across the patients and  $\sigma_W^2$  the variance of the scores within the same patient, then the population intra-class correlation is defined as

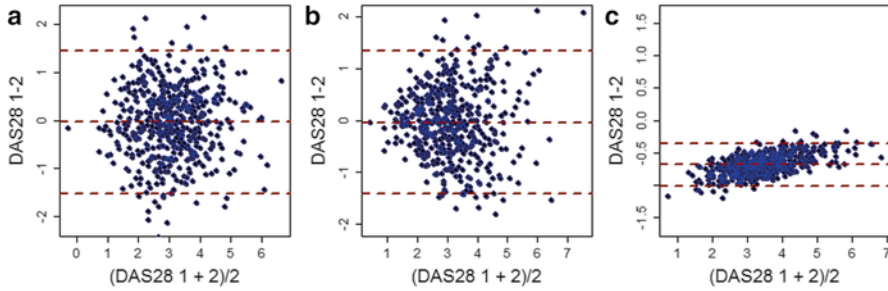
$$ICC = \frac{\sigma_B^2}{\sigma_B^2 + \sigma_W^2}.$$

The above formula shows that ICC will always be positive, and must therefore be different from the classical (Pearson) correlation. In fact, ICC measures in a scatterplot the closeness of the points to the bisecting line, while the classical correlation measures the closeness of the points to the best straight line. The sample estimate of ICC, i.e.,  $\hat{ICC}$ , is obtained by replacing in the above formula the true variances by their sample estimates, by  $s_B^2$  and  $s_W^2$ . In Fig. 6 three fictive cases of two observers scoring the same patients are shown together with the intra-class correlations: (a) the scores were generated around the bisecting line whereby the within-patient variance of the scores remains constant with increasing values of the scores, (b) the same situation except that now the within-patient variance of the scores increases with increasing values of the scores, (c) the scores between the raters are closely related but are not located on the bisecting line. The corresponding Pearson correlations are equal to 0.78, 0.78 and 0.99, respectively, which illustrates that the two measures are different in nature. Again one can test whether  $ICC=0$ , but we suggest to report the 95 % CI. As an example, the 95 % CI for ICC of Fig. 6a is equal to [0.71, 0.81].

Finally, to graphically represent whether the within-patient variability  $s_W^2$  depends on the actual value, the *Bland-Altman plot* is used. The Bland-Altman plot is a scatterplot of the difference of the two values with their average.



**Fig. 6** Fictive data: three examples of two raters scoring DAS28. Panel (a) corresponds to  $ICC=0.76$ , panel (b) corresponds to  $ICC=0.75$  and panel (c) corresponds to  $ICC=0.75$ . Further information is given in the text



**Fig. 7** Fictive data: three examples of a Bland–Altman plot corresponding to Fig. 6 (a), (b) and (c), respectively

In Fig. 7, the Bland–Altman plots corresponding to Fig. 6 are shown. In Fig. 6a  $s_w^2$  remains constant across the different DAS28 values, while for Fig. 6b  $s_w^2$  increases with DAS28 value. Clearly, Fig. 7a shows that the within-patient variability is constant, while for Fig. 7b the variability increases. Figure 7c shows that there is a problem with scoring.

Up to now, we have considered only the situation with two raters. The intra-class correlation, but also the agreement measures below can be extended to more than two observers. The actual expression of the agreement measure depends on whether the measure aims to estimate agreement between the selected observers in the study (study clinicians) or among all observers that belong to a particular population (all clinicians).

For binary, nominal or ordinal scores a popular measure of agreement is given by the *kappa coefficient*  $\kappa$  also called *Cohen's kappa*. For two binary scores, Cohen's kappa computes the relative degree of agreement, i.e., the agreement corrected for spontaneous agreement (also called agreement by chance). The theoretical formula for the true  $\kappa$  is given by

$$\kappa = \frac{\pi_o - \pi_e}{1 - \pi_e},$$

where  $\pi_o$  represents the (population) observed agreement and  $\pi_e$  the (population) agreement that happens by pure chance. In Table 4(a)  $\pi_o$  is estimated by  $\hat{\pi}_o = \frac{82 + 324}{498} = 0.82$  whereas  $\pi_e$  is estimated by  $\hat{\pi}_e = \frac{133}{498} \times \frac{124}{498} + \frac{368}{498} \times \frac{375}{498} = 0.62$ .

Then  $\hat{\kappa} = 0.52$  is the estimated excess agreement above the agreement obtained by pure chance. For Table 4(b) and (c) we obtained  $\hat{\kappa} = 0.60$  and  $\hat{\kappa} = 0.25$ , respectively. In addition one can compute a  $P$ -value for the null hypothesis. When kappa is zero, then the observed agreement is obtained by pure chance.

Agreement in ordinal data could be measured by *weighted kappas*. A greater weight is then assigned to cells that are further away from the diagonal in the table; popular are linear and quadratic weights. Note that there has been a lot of discussion in the statistical and epidemiological literature of the value of the kappa-statistic. For instance, it has been shown that it is difficult to compare kappa values obtained from different studies.

**Table 4** Fictive study: three examples upon which Cohen’s kappa is computed

	(a)			(b)			(c)		
	Rater 2			Rater 2			Rater 2		
Rater 1	82	51	133	97	33	130	21	98	119
	42	324	368	47	321	368	0	379	379
	124	375	498	134	364	498	21	477	498

The tables are obtained by binarizing the DAS28 scores with threshold 2.6, which represents the upper bound below which the disease activity is considered low

Regression Models

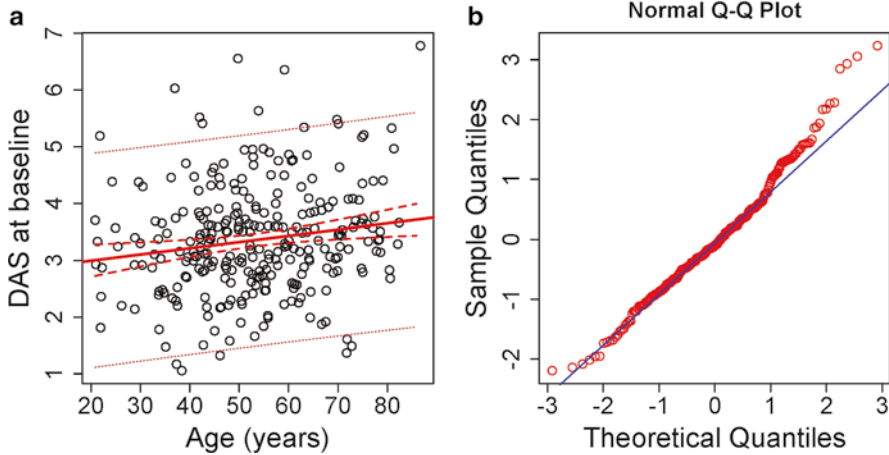
In this section we evaluate the strength of the relationship between two (or more) measurements. In addition, we now require also a mathematical expression that allows one measurement to “predict” from the other measurement(s). This entails an important class of statistical methods, called *regression methods*. First we treat linear regression models, where the response is continuous. Then binary and ordinal regression models are considered. We end with Cox regression, which is the most popular regression technique in survival analysis.

Linear Regression

Simple Linear Regression

In *simple linear regression* one variable, called the *response* or *outcome*, is predicted from another variable, called the *covariate*, *regressor* or *predictor*. In some textbooks the response is called the dependent variable and the covariate is referred to as the independent variable. However, this terminology may cause confusion in multiple linear regression introduced below and will therefore not be used here.

In Fig. 8 we show the regression line predicting DAS at month 12 in the tREACH study from the age (in years) of the patient. The straight line provides the “best” linear prediction of the response from the regressor, whereby “best” means that the squared deviations of the predicted response from the observed response are minimized with the regression line. The regression line is here given by the following formula:  $DAS = 1.11 + 0.0086 \times age$ . The coefficient 1.11 is called the (estimated) *intercept* and 0.0086 the (estimated) *slope*, they are also called the *regression coefficients*. The intercept represents the average DAS for age=0, while the slope represents the increase of the average DAS when age is increased by one year. Clearly, the intercept has no physical meaning here. When the slope is zero, the regression line is horizontal and hence the response and regressor are not related. With a positive slope the response increases on average when the regressor increases, while for a negative slope the opposite is true. Note that the regression coefficients depend on the scale of the response and the regressor. For instance, when age is replaced by



**Fig. 8** tREACH study: (a) simple linear regression, regressing DAS at month 12 on age at baseline. The *solid straight line* is the estimated regression line, the *dashed lines* express the 95 % confidence bounds for the predicted values, and the *dotted lines* express the 95 % confidence bounds for the individual observations; (b) Q–Q plot to check normality of the residuals

age/10 the slope must be multiplied with 10. This dependence on the scale of the regressor, but also of the response, makes it difficult to compare the magnitude of the regression coefficients.

The regression coefficients 1.11 and 0.0086 estimate the true regression coefficients which relate the two variables in the population. The classical assumption of simple linear regression is that the response deviates in a Gaussian way around a straight line with a variance that remains constant across the values of the regressor. The true relationship of the response with the regressor is not known but assumed to be linear. Hence for the observed response for the  $i$ th subject, denoted here as  $y_i$ , it is assumed that

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

with  $x_i$  the regressor value for the  $i$ th subject and  $\varepsilon_i$  the deviation of the response from the straight line, called the  $i$ th *residual*. It is assumed that this residual has a normal distribution. The true regression coefficients are estimated from the data. Here the estimates are:  $\hat{\beta}_0 = 1.11$ ,  $\hat{\beta}_1 = 0.0086$ . From these estimates the predicted response  $\widehat{\text{DAS}}_i = \hat{\beta}_0 + \hat{\beta}_1 \times \text{age}_i$  can be determined for each subject. The above statistical assumptions (linear relationship, Gaussian distribution around the regression line with constant variance) allow to: (1) derive the standard errors of the estimates, (2) test the null hypotheses that the true regression coefficients are zero, i.e.,  $\beta_0 = 0, \beta_1 = 0$  and (3) provide (95 %) confidence bounds for the predicted values and the responses at the different regressor values.

In Table 5, a classical regression output is shown, with the regression estimates, their standard error, the computed  $t$ -value (estimate/SE) and the corresponding  $P$ -value. For both regression coefficients the null hypothesis is rejected, i.e., there is

**Table 5** tREACH study: regression estimates for the regression model with response DAS at month 12 and age as regressor

Coefficient	Estimate	SE	<i>t</i> -value	<i>P</i>
Intercept	1.11	0.21	5.41	<0.001
Age (years)	0.0086	0.004	2.30	0.023

evidence that the true regression coefficients are not zero. However, we are only interested in verifying  $H_0: \beta_1=0$ . Since  $P<0.05$ , we believe that there is some relationship between DAS at month 12 and age. The strength of the relationship is classically expressed with the *coefficient of determination*, denoted as  $R^2$ . The coefficient of determination expresses the proportion of variability of the response that is explained by the regression model. The minimal value of  $R^2$  is 0 when the regressor has no predictive ability. The maximal value of 1 is obtained when all points lie on a (non-horizontal) straight line. For our example,  $R^2=0.021$  which is low and hence DAS is not well predicted from age. This example is an illustration that a significant relationship does not immediately result in good prediction. In addition, we can estimate the 95 % confidence boundaries for the predicted responses (dashed lines) and the future responses (dotted lines), they are both indicated in Fig. 8a. Finally, it can be verified that  $R^2$  is equal to the square of the Pearson correlation, i.e. equal to  $r_p^2$ .

Each regression analysis should be accompanied by diagnostic plots that verify the statistical assumptions. Such an exercise is too often neglected in practice. Linearity of the relationship can be graphically inspected by comparing the linear regression line with a *smooth fit*. This is a curvilinear plot that expresses the relationship between response and regressor nonparametrically, i.e., without any restrictions. The assumption of normality can be checked with a *Q-Q plot* that plots the obtained residuals, here  $\hat{r}_i = \widehat{DAS}_i - DAS_i$ , on the *Y*-axis and their expected value (under normality) on the *X*-axis. If normality applies, a straight line is (approximately) obtained. Another possibility is to apply a *normality test*, which formally tests whether the distribution of residuals is Gaussian. For the model predicting DAS28, the Q-Q plot in Fig. 8b shows some deviation from normality for the distribution of the residuals. Fortunately, since linear regression is rather robust against non-normality of the residuals, there is no immediate reason to look for another model.

Multiple Linear Regression

In *multiple linear regression*, several regressors are involved in a linear relationship with the response. The computational procedure to determine the regression coefficients is similar as for simple linear regression. Also, the assumptions upon which the statistical tests are based are the same as for simple linear regression. Yet, finding the appropriate model and interpreting the regression coefficients is now far more complex than with simple linear regression. To better explain, let us suppose that in the tREACH study we wish to predict DAS at month 12 from DAS at baseline, age, gender, and the treatment (1) triple therapy+prednisone oral (A), (2) triple therapy +prednisone injection (B), or (3) MTX+prednisone oral (C). In Table 6 we

**Table 6** tREACH study: regression model with response DAS at month 12 and DAS28 at baseline, age, gender, and treatment as regressors

Coefficient	Estimate	SE	<i>t</i> -value	<i>P</i>
Intercept	0.16	0.27	0.62	0.54
DAS baseline	0.13	0.051	2.45	0.015
Age (years)	0.010	0.003	2.86	0.0047
Gender (0= male)	0.50	0.11	4.58	<0.001
Treatment B	0.11	0.12	0.86	0.39
Treatment C	0.19	0.12	1.55	0.12

show the estimates of the multiple regression model. From that table we conclude that only age and gender significantly influence DAS at month 12.

The regression model estimated in Table 6 can be written as

DAS month 12 = 0.16 + 0.13 × DAS baseline + 0.010 × age + 0.50 × gender + 0.11 × treatment B + 0.19 × treatment C.

The estimated regression coefficients tell us that (on average) DAS at month 12 is higher for females and for greater values of DAS at baseline and age. Treatment, on the other hand, appears not to have any significant effect. Thus, the interpretation of regression coefficients appears to be the same as with simple linear regression. However, there is an important difference, namely, the value and interpretation of the regression coefficient depends on which other regressors are in the model. To better understand this, let us look at the following fitted regression models to DAS (at month 12) for the tREACH data:

- Model 1: DAS = 1.11 + 0.0086 (0.023) × age
- Model 2: DAS = 0.59 + 0.012 (0.0015) × age + 0.53 (<0.001) × gender
- Model 3: DAS = 0.24 + 0.010 (0.0054) × age + 0.52 (<0.001) × gender + 0.13 (0.011) × DAS baseline
- Model 4: DAS = 0.85 + 0.0070 (0.24) × age + 0.13 (0.76) × gender + 0.0074 (0.32) × age × gender
- Model 5: DAS = 0.93 − 0.0025 (0.91) × age + 0.54 (<0.001) × gender + 0.00013 (0.52) × age<sup>2</sup>

Each time, the *P*-value of the regression coefficient is given in parentheses. Model 1 is a simple linear regression model including only age. The model provides the *univariate effect* of age on the response, i.e., older age implies a higher DAS at month 12. The regression coefficient of age in Model 2 represents the effect of age when gender is kept constant, i.e., it represents the effect of age within males and females separately. It is said that the effect of age is *controlled for* gender and this significantly augments the effect of age here. This is called the *multivariate effect* of age when gender is included in the model. Note that the women are significantly younger in this study. Together with the fact that women have a higher DAS at month 12, it explains why this model shows a stronger effect of age. DAS at baseline value is included in Model 3, which has (as expected) a significant impact

on the DAS value at the end. In Model 4, the product of age with gender, also called the *interaction between age and gender*, is added to the model. Now none of the age regression coefficients is significant anymore. This is an illustration of *multicollinearity* in the regressors. Multicollinearity is a common phenomenon when highly correlated regressors are included in the model causing unstable regression computations. Model 4 is a sign that one must be quite careful in building up the model and interpreting the regression coefficients. Model 5 illustrates that with linear regression nonlinear relationships can also be expressed. There is, however, again a multicollinearity problem since age is positively and highly (closely) linearly correlated with  $\text{age}^2$ . The linear relationship between age and  $\text{age}^2$  can be removed by working with a centered age, namely, with  $\text{agec} = \text{age} - \text{mean}(\text{age})$ , and  $\text{agec}^2$ . For this model, the regression coefficients for gender and  $\text{agec}^2$  remain the same but the regression coefficient of agec changes drastically and is now statistically significant ( $P = 0.0014$ ).

The above shows that building an appropriate multiple linear regression model may be not that easy. Below is a list of challenges that one may face in the model-building process:

- When a large number of regressors is available, it is not immediately clear which regressors to include in the model. It is popular to select regressors in an automated manner, using, e.g., *stepwise selection procedures*. However, it is known that these procedures do not necessarily result in a meaningful model. In addition, since such a procedure involves many decisions to include or exclude a regressor, the reported  $P$ -values therefore suffer severely from the multiple testing problem.
- In multiple linear regression, there is much more freedom to deviate from the model assumptions. In order to achieve an appropriate model, we might have to transform the regressors, add double products, and/or transform the response. Such transformations might also be needed to improve the normality of the residuals. When the variance of the response is not constant, the model needs to be further expanded and another computational procedure is needed.
- To find out whether the constructed model is appropriate, i.e., satisfies the statistical assumptions, a battery of diagnostic plots is needed. One example of such a plot, the Q–Q plot of the residuals, was seen for simple linear regression. However, many other residual plots are needed to check the multiple linear regression model. In addition, diagnostic procedures should be used that can highlight *influential observations*, i.e., observations that have an unduly large effect on the estimates of the regression coefficients.

Constructing the appropriate multiple linear regression model may therefore need considerable statistical background. All these efforts do not, however, guarantee that we find the true model, if such a thing exists. We can only hope for a useful one.

## Logistic Regression

When the response is binary, linear regression is not appropriate anymore. Most popular is the *logistic regression model* whereby the probability of experiencing an event (scored as “1”) is expressed as a function of the covariates using the logistic function. More specifically, let  $\pi_i$  be the probability of experiencing the event and  $x_i$  the regressor for the  $i$ th individual, then the simple linear logistic model is given by

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_i,$$

whereby  $\log\left(\frac{\pi_i}{1-\pi_i}\right)$  is also denoted as  $\text{logit}(\pi_i)$ . An equivalent way of specifying the model is

$$\pi_i = \text{expit}(\beta_0 + \beta_1 x_i) = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)},$$

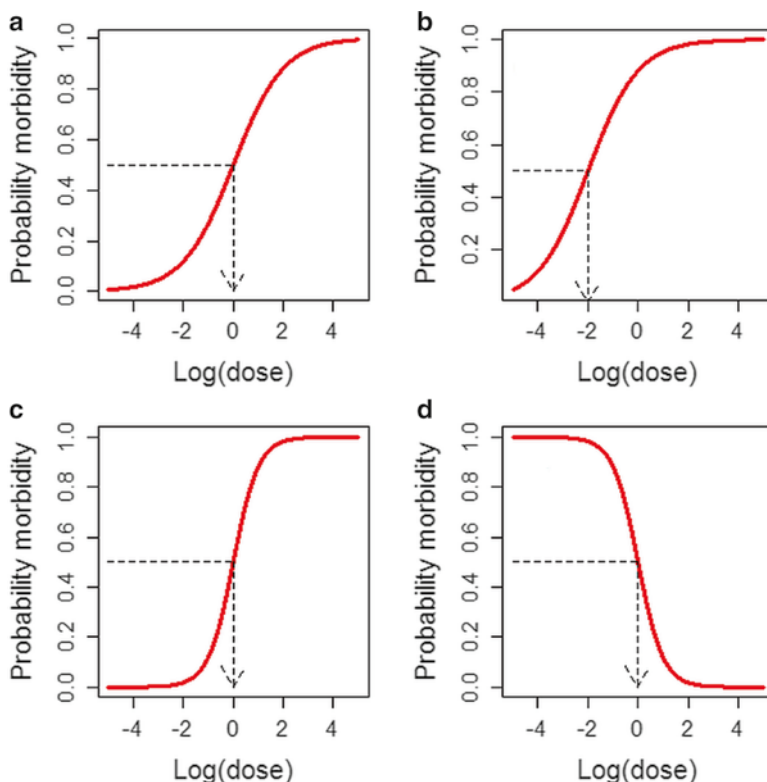
where  $\exp(a) = e^a$  and the function  $\exp(a)/(1 + \exp(a))$  is referred to in the literature as the *expit function*. The parameters  $\beta_0$  and  $\beta_1$  are again called the regression coefficients, with  $\beta_0$  playing the role of an intercept and  $\beta_1$  of a slope. In Fig. 9, four logistic models are shown in a fictive preclinical setting relating morbidity to the dose of an experimental drug. It is immediately seen that always  $0 < \pi_i < 1$ , whatever the values of  $\beta_0$  and  $\beta_1$  and of the regressor are. When  $\beta_1 > 0$  ( $\beta_1 < 0$ ), increasing (decreasing) the regressor will increase the probability of an event, while  $\beta_1 = 0$  implies that the regressor has no impact on the response.

For a binary regressor, it can be shown that  $\exp(\beta_1)$  expresses the odds ratio relating the binary response to the regressor. For a continuous regressor,  $\exp(\beta_1)$  expresses the odds ratio of the response with the regressor when the regressor increases by one unit. In practice  $\beta_0$  and  $\beta_1$  are estimated from the data, resulting in  $\hat{\beta}_0$  and  $\hat{\beta}_1$ . For each individual, the predicted response  $\hat{\pi}_i$  can be computed by plugging in the estimates  $\hat{\beta}_0$  and  $\hat{\beta}_1$  in the expression for  $\hat{\pi}_i$ . Based on these estimates, one can test whether the true regression coefficients  $\beta_0$  and  $\beta_1$  are equal to zero. As for linear regression, we will be only interested in the test  $\beta_1 = 0$ . Again 95 % CIs for  $\beta_0$  and  $\beta_1$  can be computed.

As for linear regression, more than one regressor can be included in the logistic regression model. The interpretation of the regression coefficients then depends on which other regressors are included in the model. For instance,  $\exp(\beta_1)$  then expresses the odds ratio of the regressor with the binary response but controlled for the other regressors.

The computational procedure to establish the estimates of the regression coefficients is iterative, i.e., the numerical algorithm needs several steps to end up in the estimates. This is in contrast to linear regression where analytical solutions are available for the regression coefficients. But, apart from the numerical procedure, the same challenges as in linear regression are to be dealt with in logistic regression.





**Fig. 9** Fictive preclinical study: four logistic models are shown relating  $\pi_i = \exp(\beta_0 + \beta_1 x_i) / (1 + \exp(\beta_0 + \beta_1 x_i))$  to  $x_i$  with: (a)  $\beta_0=0, \beta_1=1$ ; (b)  $\beta_0=2, \beta_1=1$ ; (c)  $\beta_0=0, \beta_1=2$ ; and (d)  $\beta_0=0, \beta_1=-1$ . The solid line represents the logistic curve for different values of the  $\log(\text{dose})$ . The dashed lines point to the  $\log(\text{dose})$  that corresponds to probability = 0.5

For instance, it is not immediately clear which regressors should be included in the model and whether they need to be transformed. Two measures of performance are popular: (1) an adapted  $R^2$ , called *Nagelkerke's  $R^2$* , such that its minimal value is 0 and maximal value is 1; and (2) a *concordance measure* (between 0.5 and 1), which measures the proportion of pairs of observations that have the same ordering in the observed (binary) responses as in the corresponding pair of predicted responses. For a non-predictive model, the concordance is equal to 0.5.

As an illustration, we explore in the tREACH study the relationship between a binary outcome bDAS (obtained from binarizing DAS at month 12 using threshold 2.4) and various regressors. Then, bDAS is 1 if there is low disease activity (<2.4), 0 otherwise. Three models express the probability of having low disease activity as a function of regressors. We obtain (within parentheses  $P$ -values):

- Model 1:  $\text{logit}(\text{pDAS}) = 2.53 - 1.23 (0.0078) \times \text{gender}$
- Model 2:  $\text{logit}(\text{pDAS}) = 4.02 - 1.39 (0.004) \times \text{gender} - 0.025 (0.054) \times \text{age}$
- Model 3:  $\text{logit}(\text{pDAS}) = 4.62 - 1.38 (0.004) \times \text{gender} - 0.023 (0.082) \times \text{age} - 0.22 (0.23) \times \text{DAS baseline}$

In Model 1, only gender is included. According to the fitted logistic model, males in the tREACH study show a lower disease activity. The odds ratio for gender is equal to 0.29, with 95 % CI=[0.11, 0.72]; hence, female patients have a lower probability to have a low disease activity at the end of treatment. One can verify that these estimates are the same as those obtained from a  $2 \times 2$  contingency table. In Model 2, age is added to the model. Now the regression coefficient of gender is controlled for age. The odds ratio is slightly decreased to 0.25 with 95 % CI=[0.098, 0.63]. In Model 3, the DAS value at baseline is added to the model, but surprisingly it appears to have no impact on the response. For the three models, Nagelkerke's  $R^2$  is equal to 0.058, 0.083, and 0.092, respectively, while the concordance is for the three models 0.611, 0.663, and 0.674, respectively. We see some increase in predictive performance when age is added to the model, but none of the models does a satisfactory job in predicting low disease activity at the end of the study.

The logistic regression model is one of the most popular models in epidemiology to search for risk factors for a variety of diseases. Its popularity has much to do with the property that the odds ratio obtained from a logistic model obtained from a case–control study is equal to the odds ratio obtained from a logistic model applied to a corresponding cohort study (see chapter “[Methodological issues relevant to observational studies, registries and administrative health databases in rheumatology](#)”).

Other models for binary outcomes in this class are the *probit* and *complementary log–log regression model*, for which the expit function is replaced by other S-shaped functions. An extended version of the logistic regression model has been suggested for an ordinal response, called the *ordinal logistic regression model*.

Finally, the logistic model belongs to a general and important class of statistical models, called *generalized linear models*. This class of models hosts many important models in statistics.

## Cox Regression

A regression model for a survival response needs to take care of (right) censored observations. By far the most popular regression model for survival data is *Cox regression model* proposed by Sir D.R. Cox in 1972 [18]. Cox's approach is based on a fundamental assumption on the *hazard function*, which we first introduce. We recall that “survival” does not need to be interpreted literally, but rather “death” means the occurrence of an event, such as drug survival or the occurrence of arthritis in ACPA-positive arthralgia over time or cardiovascular events.

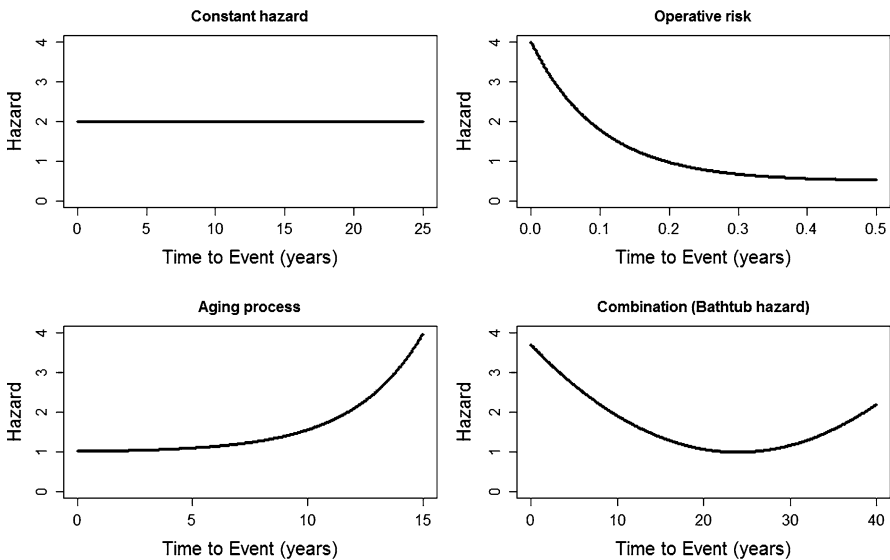
### The Hazard Function

The hazard function expresses the instantaneous risk for “dying” as a function of time that applies to a subject. More formally, the hazard function  $h(t)$  is defined as

$$h(t) = \frac{\text{Prob}(\text{death in } [t, t + \Delta t] | \text{alive at } t)}{\Delta t},$$

for small  $\Delta t$ . The formula reads as follows: given that a subject is alive at time  $t$ , the hazard at time  $t$  is the probability of dying in an interval of size  $\Delta t$  immediately after time  $t$ , divided by  $\Delta t$ .

The survival function and the hazard function provide complementary information, with the former describing the cumulative process of dying. Further, for each (theoretical) survival function, there is a corresponding hazard function. In fact, when we know the survival function, we also know the hazard function and vice versa. It is illustrative to look at some common hazard functions in Fig. 10. The constant hazard is the hazard caused by a variety of causes that may happen during any time in the life of an individual, such as a fall, a car accident, etc. The hazard caused by surgery is typically high at the time of surgery and then decreases with time. The aging process causes people to die when they get older; hence, the hazard function increases with age. Finally, the bathtub hazard function is seen when the different risks jointly apply to a population.



**Fig. 10** Some examples of theoretical hazard functions

The hazard function can be easily computed when a parametric assumption of the survival distribution is made, such as a Weibull distribution. Without such an assumption, it is difficult to obtain a reliable estimate of the hazard function from the data. This was recognized by Cox in 1972, who proposed a method that allows estimating the effect of risk factors on survival without needing to estimate the hazard function.

## The Proportional Hazards Assumption

The *proportional hazards assumption (PH assumption)* specifies that the impact of a regressor acts multiplicatively on the hazard function. In the case of a binary regressor, say gender, the PH assumption implies that the hazard function for males is proportional to that for females. When the hazard ratio is equal to 2, we have

$$\frac{h_{\text{Male}}(t)}{h_{\text{Female}}(t)} = 2.$$

This signifies that the instantaneous risk for men ( $h_{\text{Male}}(t)$ ) is twice the risk for women ( $h_{\text{Female}}(t)$ ). When the ratio is not constant, we say that the PH assumption is violated. The hazard ratio  $h_{\text{Male}}(t)/h_{\text{Female}}(t) = c$  is equivalent with  $h_{\text{Male}}(t) = h_0(t) \times \exp[\log(c)] = h_0(t) \times \exp[\beta_1 \times \text{gender}]$  with  $\beta_1 = \log(c)$  and gender = 0 for a female and 1 otherwise. In the above re-expression, the hazard function of the female patients plays the role of a *baseline hazard*. When there are  $p$  regressors in the model, the PH assumption generalizes to

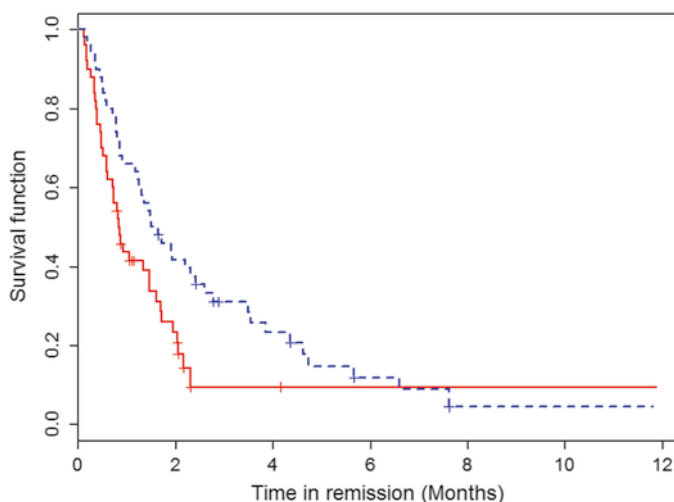
$$h_x(t) = h_0(t) \times \exp[\beta_1 x_1 + \beta_2 x_2 \cdots + \beta_p x_p].$$

Summarized, the PH assumption assumes that the regressors act multiplicative on the hazard function and their effect remains constant during the study. The regression coefficients represent, as for the other regression models, the strength of the regressors in the presence of the other regressors.

## Cox Regression

In 1972, Cox proposed a method to estimate the regression coefficients under the PH assumption. His approach does not require estimating the hazard function and became the most important survival regression method.

For an illustration of Cox regression, we take a fictive example that compares the time in remission between men and women, see Figure 11. The tREACH data cannot be used as an example here since patients are examined when visiting their rheumatologist at regular time intervals, which implies that any event of interest (but not fatal) is interval censored. Cox regression was, though, proposed for right-censored survival times.



**Fig. 11** Fictive study: Kaplan–Meier estimates of the survival functions for men (*dashed line*) and women (*solid line*)

As for logistic regression, an iterative procedure is needed to estimate the regression coefficients. In a Cox regression analysis, no intercept is estimated. This can be seen in the above expression of the general PH assumption. There is only one regressor  $x_1$  equal to gender (female = 1). The estimate of  $\beta_1$  is equal to 1.09, so that females go out of remission sooner than men. Now  $\exp(\beta_1) = 2.97$  is an estimate of the hazard ratio, which is the coefficient  $c$  in section “The proportional hazards assumption.” The 95 % CI for the hazard ratio, equal to [1.85, 4.77], does not include 1. We therefore conclude that females have a significantly higher risk to go out of remission than males ( $P < 0.001$ ). An adapted  $R^2$  and a concordance measure allow evaluating the predictive performance of the survival function. We obtained here  $R^2 = 0.176$  and 0.636 for concordance.

As for the other regression models, several regressors can be included in a Cox regression model. The inclusion of other regressors will change the value and the meaning of the original regression coefficients. All issues that popped up with linear and logistic regression, such as which regressors to include and in what scale, also apply to Cox regression.

We note that the PH assumption is an assumption that needs to be verified. A sign of nonproportional hazards are crossing survival functions (if based on enough subjects) but also formal diagnostic procedures are available. When the effect of the regressors is not constant over time, one might extend the model by including interaction terms with time. It is also possible to assume some smooth dependence of the hazard ratio with time. Estimating the regression coefficients is then considerably more complex. It could also be that regressors change during the conduct of the study. They are called *time-dependent regressors* and can be incorporated in a classical Cox regression analysis. Recently another approach, based on *joint modeling* of a survival and a longitudinal process, has been proposed and looks quite promising [19].

## Models for Longitudinal Studies

In a follow-up (FU) study, subjects are followed up in time. In the previous section, the time to an event was recorded in the FU study. Another example is when individuals are examined at several time points, which leads to a *longitudinal study*. Important examples of longitudinal studies are the randomized clinical trial (RCT) and the cohort study in epidemiology. To properly analyze longitudinal data, one needs to take into account the correlated nature of the repeated measures and one needs to address the fact that patients may miss examinations or drop out from the study. Many longitudinal studies in rheumatology are, however, analyzed inappropriately because of the unawareness of these two problems.

We first discuss the impact of missing data on the analysis of longitudinal studies, then give a brief review of some older, but still in use, statistical techniques possibly in combination with imputation techniques. We end with more modern techniques that incorporate flexibly the correlated nature of the data and allow for less restrictive missing data processes.

### *The Problem of Missing Data*

Missing data can affect all kinds of studies, but with longitudinal studies, we have more tools to address the problems that missing data cause. The amount and the reason why data are missing dictate what statistical technique to use. First, note that subjects may miss a visit *intermittently* and then return afterward to the study or they may *drop out* completely from the study. The most serious problem is the latter situation upon which we will focus here. Leaving the study may happen for a variety of reasons. A classical taxonomy introduced by Little and Rubin [20] still dominates the missing data terminology. Here we discuss this terminology in the context of regression models, where we assume that the response may be missing, but not the regressors. One distinguishes:

- *Missing completely at random (MCAR)*: A missing response occurs because of reasons completely unrelated to the response, i.e., by pure bad luck.
- *Missing at random (MAR)*: The missing data mechanism is related to observed responses. For example, when in an RCT patients are removed from the study by the clinical investigator because their DAS28 is too high, the dropout process depends on the latest value of DAS28 recorded in the study.
- *Missing not at random (MNAR)*: The missing data mechanism may not only be related to observed responses but also to unobserved responses. Take the previous example, but now assume that a visit to a rheumatology clinic outside the study reveals that the patient's DAS28 exceeds 5. The patient therefore decides to change medication and leaves unrecorded the study. Consequently, the dropout of the patient cannot be predicted within the study from the recorded past measurements.

Missing data may affect seriously the statistical analysis and the clinical conclusions. For instance, the descriptive statistics such as the mean, median, SD, etc. may be severely distorted with the MAR and MNAR missing data mechanisms, see, e.g., [21]. Most classical statistical techniques for the analysis of repeated measures are valid under MCAR (but their precision may be severely affected) but are likely to fail when the missing data processes are MAR or MNAR.

## *Classical Statistical Techniques*

In clinical research, it is still common practice to compare two treatments in a longitudinal study by significance tests at each visit. Simplicity is the only advantage of this approach. Indeed, the *repeated significance testing approach* is flawed with various problems: (1) it suffers from the multiple testing problem, (2) this approach turns a longitudinal study into several cross-sectional studies and therefore neglects the correlation among the responses, (3) the approach can only be applied when the examination times are (roughly) regular, and finally, (4) with this approach, it is difficult to imagine what the results imply for future patients because at each visit the comparison of the treatments is done on a different set of patients, i.e., on only those patients that are present at the respective visits. It is an example of an *available case approach*, whereby only the patients available at the examination can be compared. Finally, it is only valid under the MCAR assumption.

To address the multiple testing issue, one could apply an ANOVA approach. There are two classical ANOVA techniques to analyze repeated measurements: *repeated measurements ANOVA* (rANOVA) and *multivariate ANOVA* (MANOVA). Both approaches were popular among statisticians about 50 years ago. However, these approaches are not suitable for contemporary studies in clinical research since they require the data to be balanced, i.e., all subjects should be measured at the same time points and the data are not plagued by missing values. For rANOVA and MANOVA, a subject will be removed from the analysis if he/she has missed only one measurement. In addition, rANOVA assumes that the correlation among all repeated measures is the same irrespective of the time lag between the measurements (*compound symmetry*). For MANOVA, the correlation matrix must be general (*unstructured*) and this might require too many variance and correlation parameters to estimate. For instance, when there are 6 visits, 21 correlations and variances need to be determined. This causes two problems: (1) estimating too many parameters for the given data reduces the power of the analysis considerably; (2) when there are relatively few subjects and many measurements per subject, the model parameters may be not estimable ruling out MANOVA as an option. The two ANOVA approaches are examples of the *complete-case approach*. They are only valid on the MCAR missing data mechanism. Despite the abovementioned drawbacks, the two ANOVA methods are still frequently used in the clinical literature.

## Imputation Techniques

The above classical techniques may be combined with an approach that imputes reasonable values for the missing data. This may limit their efficiency loss in case of imbalance. A popular imputation approach in RCTs is the *last-observation-carried-forward (LOCF) approach*. This technique imputes for all the missing responses the last observed response value. Suppose that a patient dropped out at visit 3, then with the LOCF approach, the last recorded DAS28 value at visit 2 is repeatedly filled-in for all subsequent missing DAS28 values. There are, however, serious statistical as well as clinical problems with this approach. Indeed, it is now generally recognized that the LOCF procedure creates unrealistic profiles (both in terms of mean and variance). Further, the statistical properties of an analysis based on LOCF imputed data are unclear, see [21, 22].

An appropriate approach to impute missing data is the *multiple imputation (MI) approach*. The MI approach is based on a statistical model to impute the missing data stochastically. To reflect that filled-in data are subject to uncertainty, the imputation is done more than once (typically  $M=3$  to 5 times) yielding  $M$  imputed data sets. The imputed data sets are then combined in a second step for the statistical analysis of the data. Any statistical model can be combined with MI approach. The MI approach can also be applied to impute missing regressor values.

## More Recent Approaches to Analyze Longitudinal Data

We consider here two approaches: *mixed models* and *generalized estimating equation techniques*. We focus on continuous responses but mention also briefly the analysis of binary and ordinal responses.

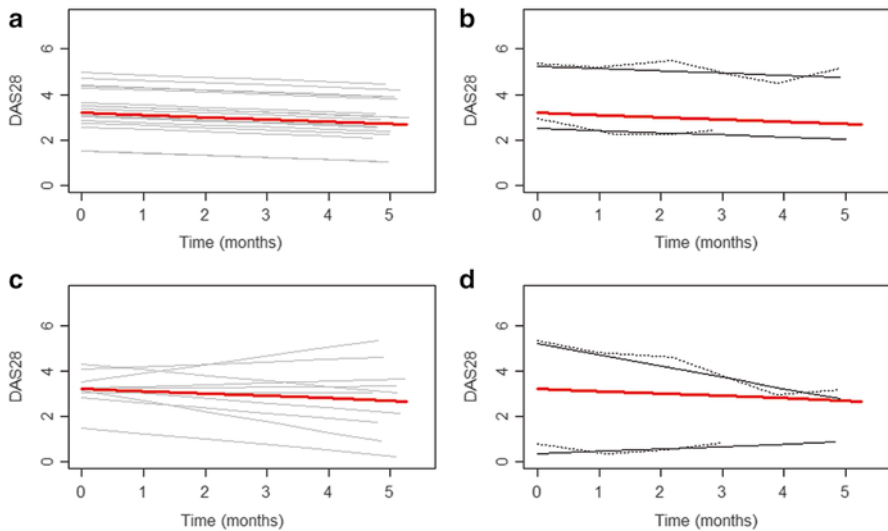
### Linear Mixed Models

A *linear mixed model* (LMM) assumes there exists an average profile for the population of patients from which the individual profiles deviate in a random manner by a subject-specific intercept, slope, quadratic term, etc. In Fig. 12, we give examples of LMMs whereby the evolution of the  $i$ th individual deviates from the overall downward linear trend in DAS28 by a subject-specific intercept  $b_{0i}$  (random intercept model) or additionally by a subject-specific slope  $b_{1i}$  (random intercept + slope model). The random intercept + slope model is given by

$$\text{DAS28}_{ij} = \beta_0 + \beta_1 \text{time}_{ij} + \dots + b_{0i} + b_{1i} \text{time}_{ij} + \varepsilon_{ij},$$

with  $\beta_0, \beta_1, \dots$  called *fixed effects*. The sub index  $i$  pertains to the patient number; the sub index  $j$  (here 1 to 5) pertains to the visits with  $j=1$  referring to the baseline visit and  $j=5$  to the 5th monthly visit. The dots indicate that additional fixed effects can be included in the model. The solid thick line in Fig. 12 represents  $\beta_0 + \beta_1 \text{time}_{ij}$ .

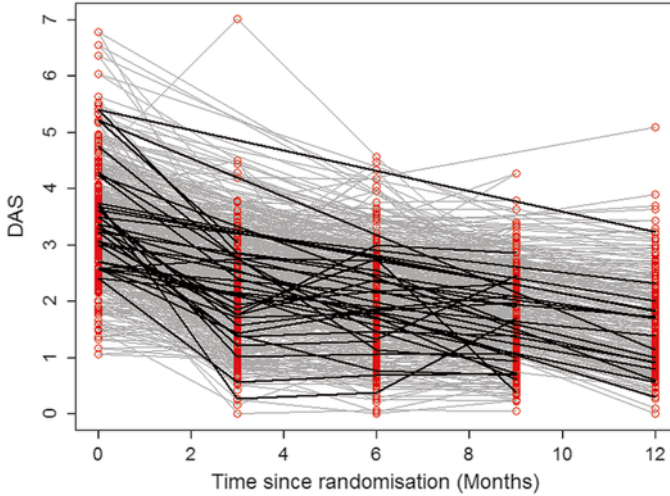




**Fig. 12** Examples of mixed models. (a): random intercept model showing a sample of subject-specific linear trends and (b) two specific trends in the random intercept model together with observed profiles. (c) random intercept+slope model showing a sample of subject-specific linear trends, and (d) two specific trends in the random intercept+slope model together with observed profiles. In the different plots the solid thick line corresponds to the population average evolution. The thin solid lines correspond to the individual linear evolutions. The dotted lines in panels (b) and (d) represent the actual observed profiles

Note that here  $\text{time}_{ij} = \text{time}_j$ , which means that the time intervals between visits to were taken the same for all subjects. On the other hand,  $b_{0i}$ ,  $b_{1i}$  represent the deviations of the subject-specific profiles from the population profile and are called *random effects*. Finally,  $\varepsilon_{ij}$  represents the measurement error, which is the fluctuation of the observed response around the subject-specific regression line ( $\beta_0 + \beta_1 \text{time}_{ij} + \dots + b_{0i} + b_{1i} \text{time}_{ij}$ ). The above model therefore reads as follows: the response (DAS28) at visit  $j$  of the  $i$ th patient is the sum of the overall trend seen in the population+the specific trend in patient  $i$ +the temporal fluctuation at visit  $j$ . As one can observe, the correlation among the repeated measurements is determined by the random intercept and slope, which ties together all observations from the same individual. For an LMM, no explicit imputation is involved, but there is still *implicit* imputation as can be seen in Fig. 12. In other words, for patients who drop out, it is assumed that their unobserved profile (after dropout) continues along their subject-specific profile.

The LMM allows for unequal time points. To estimate the model parameters ( $\beta_0$ ,  $\beta_1$ , ... and the variances of  $b_{0i}$ ,  $b_{1i}$ , and  $\varepsilon_{ij}$ ), distributional assumptions need to be made. Classically, it is assumed that  $b_{0i}$ ,  $b_{1i}$ , and  $\varepsilon_{ij}$  have normal distributions with a zero mean and variances to be estimated by the data. The random effects are allowed to be correlated but should be independent from measurement error. Based on these assumptions, all parameters can be estimated. Given that the model is correctly specified, the parameters are well estimated for an MCAR or MAR dropout process,



**Fig. 13** tREACH study: spaghetti plot of observed longitudinal DAS profiles; for a random sample, the profile is printed in black; all others are printed in gray

but in principle not for an MNAR dropout process. We refer to [23] for the motivation of this result and further technical details.

As an illustration, we analyzed the longitudinal DAS responses from the tREACH study. It was planned that the RA patients were examined at baseline, 3, 6, 9, and 12 months. However, some of the patients missed visits and/or were dropouts. Of the 281 patients who were randomized to the three treatments (91 patients to treatment A, 93 to B, and 97 to C), 264 patients were still in the study at 3 months, 255 patients at 6 months, 250 patients at 9 months, and 248 patients at 12 months. Hence, only a relatively few patients dropped out. From experience, we know that most often the dropout mechanism is at least MAR, which motivates the use of the linear mixed model.

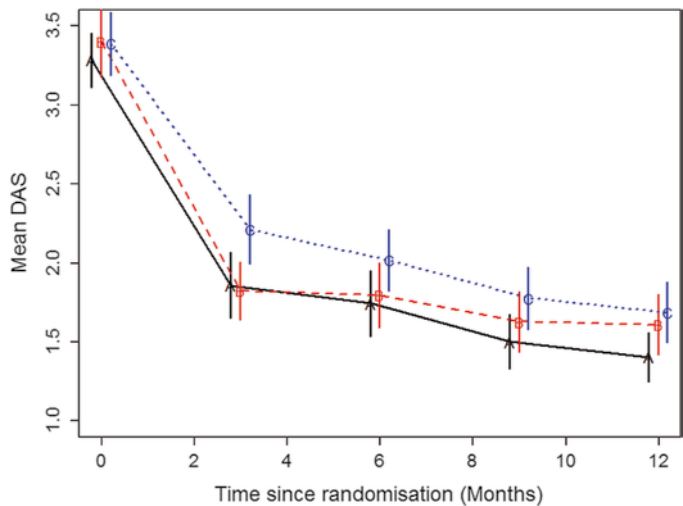
In Fig. 13 the individual profiles of all patients are plotted. We observe that overall there is a decrease in DAS but also that there is quite some variability. In Fig. 14, we show the mean  $\pm$  SEM plots based on the observed data. For an MAR dropout process, we have seen above that these plots may be misleading but here the amount of dropouts is limited and hence the descriptive measures probably give a good picture of the true values.

The following LMM was fit to the DAS responses using the *R* function *lmer*:

$$\text{DAS}_{ij} = \beta_0 + \beta_1 \text{time}_{ij} + \beta_2 x_{1i} + \beta_3 x_{2i} + \beta_4 x_{3i} + \beta_5 x_{4i} + b_{0i} + b_{1i} \text{time}_{ij} + \varepsilon_{ij},$$

with

- Fixed effects, the regression coefficients of time ( $\text{time}_{ij}$ ), gender ( $x_{1i}$ ), age at baseline ( $x_{2i}$ ), treatment ( $x_{3i}$ ), and duration of complaints ( $x_{4i}$ ).
- Random effects: random intercept ( $b_{0i}$ ) and slope ( $b_{1i}$ ).



**Fig. 14** tREACH study: mean±SEM plots of DAS split up into the three treatment groups. The solid line corresponds to treatment arm A, the dashed line to treatment arm B and the dotted line to treatment C

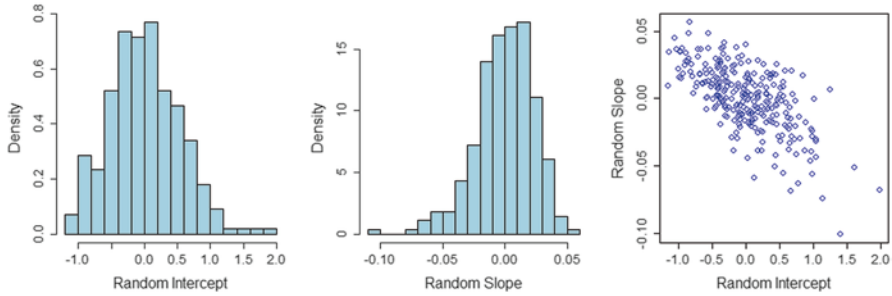
**Table 7** tREACH study: parameter estimates for the LMM with response: DAS and regressors: time since randomization, gender, age at baseline, treatment, and duration of complaints

Coefficient	Estimate	SE	t-value
Intercept	2.29	0.19	12.10
Time	−0.13	0.006	−22.93
Gender (0= male)	0.39	0.086	4.57
Age (years)	0.011	0.003	3.91
Treatment B	0.03	0.098	0.34
Treatment C	0.19	0.097	1.97
Complaints	0.0004	0.00044	0.92

The results of this analysis are shown in Table 7. Note that the lmer function does not provide *P*-values. The reason is that the degrees of freedom of the *t*-distribution are quite hard to determine in the LMM. Nevertheless, we can deduce from Table 7 that except for treatment and duration of complaints all regressors have a significant impact on the response (*t*-values much larger than 2 in absolute value).

Also, the random intercept and slope of each patient can be estimated. Recall that  $b_{0i}$  expresses the subject-specific deviation of the intercept for the  $i$ th subject, while  $b_{1i}$  expresses the subject-specific deviation of the slope. In Fig. 15, the estimates of the random effects are shown. We notice that the histograms of the random effects show some mild deviation from normality. The scatterplot shows that compared to the overall trend, patients who start relatively low may have their DAS value increase or be roughly stable over time, while those who start relatively high have a tendency to decrease considerably.

The model in Table 7 is a starting point. We can then explore which other regressors should be included, whether polynomial terms in time or age are needed or



**Fig. 15** tREACH study: histograms of the random intercept and slope and scatterplot of the random effects

double products, whether the random part should be made more complex by adding, say, a random quadratic term, etc. All of this can be done as in classical regression, and one can test which of the models is most appropriate.

We conclude that an LMM analysis provides a convenient way for analyzing contemporary follow-up studies, which are often hampered by many missing data. A condition is, however, that the LMM is (approximately) correctly specified and that the missing data process is at most MAR. However, one observes in simulations that the LMM often performs well in the case of MNAR especially if the repeated measurements are highly correlated. With regard to interpretation, a statistical analysis with a linear mixed model provides a treatment effect for the whole patient group if these patients were able to stay in the study until the end. This is different with the interpretation of a complete case analysis. Namely, a complete case analysis evaluates the treatment effect only for the patients still present at the end of the study and who never missed a visit. This is problematic since one cannot know in advance who will comply with the treatment during the whole of the study. On the other hand, with an LMM, none of the patients are excluded (provided they deliver at least one measurement), and all patients contribute to the estimated treatment effect.

## Generalized Linear Mixed Models

A popular model to analyze longitudinal binary responses is the *logistic random effects model*. It is a generalization of the logistic model seen in section “Logistic regression” to include random effects. As an example, suppose that we are interested in the probability of remission in the tREACH study at each clinical examination. Let then  $\pi_{ij}$  be the probability that  $\text{DAS} < 1.6$  at visit  $j$  and for patient  $i$ . A *logistic random intercept model* relating this probability to time and with the regressors of the previous section is given by the expression

$$\text{logit}(\pi_{ij}) = \beta_0 + \beta_1 \text{time}_{ij} + \beta_2 x_{1i} + \beta_3 x_{2i} + \beta_4 x_{3i} + \beta_5 x_{4i} + b_{0i}.$$

The random intercept  $b_{oi}$  is common to subject  $i$  and links all his repeated data. The fixed effects  $\beta_0, \beta_1, \dots, \beta_5$  have now a somewhat different interpretation than for a classical logistic regression model because of the inclusion of the random intercept into the model. Note that the above model does not have a measurement error. If needed, a random slope can be added to the model. The logistic random effects model has been further generalized to ordinal responses. We refer to [23, 24] for further technical details.

A special case of the logistic random intercept model is the *Rasch model*. This is a psychometric model for analyzing categorical data, such as answers to questions on a reading assessment or questionnaire responses. In [25], this model was used to determine whether the 14 questions (items) in the Completed Behçet's Disease Current Activity Forms form a hierarchical and unidimensional scale of disease activity. Specifically the authors used the model

$$\text{logit}(\pi_{ik}) = b_i - \beta_k,$$

with  $\pi_{ik}$  the probability that subject  $i$  will answer the item  $k$  correctly (or be able to do task  $k$ ),  $b_i$  playing here the role of the disease activity of that subject, and  $\beta_k$  the item activity parameter. Hence, in this model,  $b_i$  is the random intercept that expresses the personal ability of a subject to answer the item correctly, while  $\beta_k$  is a fixed effect expressing the overall difficulty of answer item  $k$  correctly. This model is then fitted to each of the items.

Other repeated responses, such as counts, can be also analyzed with mixed effects models. A general class of such models is given by the *generalized linear mixed model*, extending the generalized linear models mentioned in section "Logistic regression" to include random effects. For all these mixed effects models, computations to determine the parameter estimates are considerably more involved (involving integral calculations) but are still feasible. Inference is again robust under an MCAR and MAR missing data process provided the model is (approximately) correctly specified.

## Generalized Estimating Equations

The *generalized estimating equations (GEE)* approach is different in nature from the mixed model approach, where no complete model specification is required for the repeated measurements. The GEE approach can be applied to continuous and categorical outcomes. While for the mixed model approach, care should be taken that the mean structure and the correlation matrix should be correctly specified, with the GEE approach, only the mean structure needs to be specified correctly. For the correlation structure, just a rough guess is needed, called the *working correlation matrix*. While the GEE approach is a more robust approach to analyze longitudinal studies, it generally requires a larger sample size than the mixed model approach. Further, the basic version of GEE is only robust against an MCAR process. A *weighted GEE* or multiple imputation combined with GEE provides protection against an MAR process, at the expense of again a larger sample size.

## Frailty Models

A generalization of Cox regression that includes random effects is called the *frailty model*. This approach consists of a variety of survival techniques to analyze clustered survival times occurring, e.g., when a patient suffers from several RA flares over time or RA patients cluster in groups because they have been treated in different hospitals, etc.

## Approaches to Deal with MNAR Missingness

When the missing data or dropout process is of the MNAR type, in principle none of the approaches described above work. The problem of an MNAR process is that the probability of missing data/dropout depends on unobserved responses. Hence, there is no way to check what the specific missing data mechanism is. The only solution is to imagine different missing data processes and combine these with the primary analysis of the repeated measurements, i.e., to perform a *sensitivity analysis*. We refer to [23, 24, 26] for a further theoretical background and for practical guidelines.

## Multivariate Methods

Up to now, we have considered only one response at a time. There is a whole class of statistical methods that allows exploring many responses at the same time; these are called *multivariate methods*. Note that multiple regression is often referred to in the literature as multivariate regression; this is however a wrong term because this statistical approach only involves one response.

Examples of multivariate techniques are *principal component analysis*, *factor analysis*, *biplot graphs*, etc. These approaches have in common that they aim to discover the intrinsic dimensionality of the multivariate response. For instance, in [27] 272 consecutive Turkish patients with Behçet's disease (BD) were examined for target organ associations. The authors extracted four factors using a factor analysis of the variables: oral and genital ulcers, erythema nodosum, papulopustular skin lesions, uveitis, superficial and deep vein thrombosis, and joint, arterial, neurological, and gastrointestinal involvement. These four factors explained 69 % of the information in the original measurements. We refer to the statistical literature for further details on this rich class of models.

## The Bayesian Approach

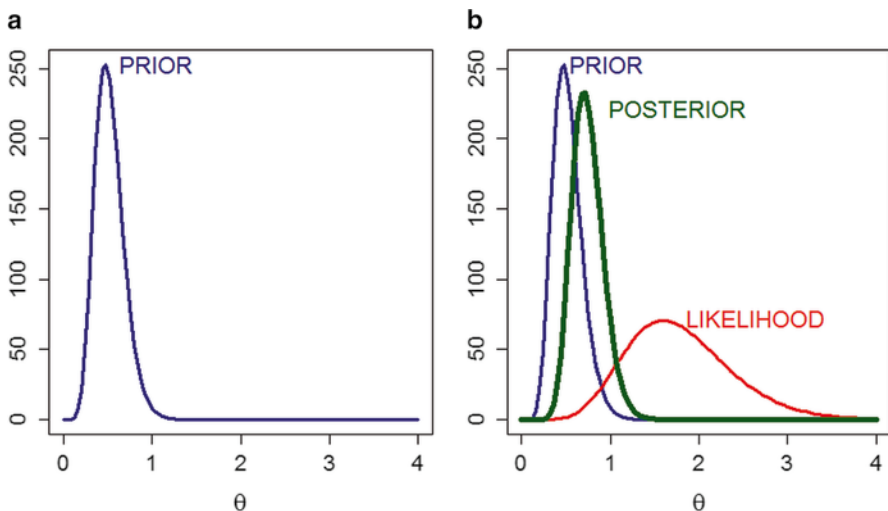
There is a growing interest in an alternative approach for statistical inference. The basis for this approach goes back 250 years with Bayes' theorem, which was published in 1763. Two years after the death of reverent Thomas Bayes, his friend

Richard Price published the document “An Essay toward a Problem in the Doctrine of Chances,” which is based on the writings of Bayes and which includes Bayes’ theorem. Bayes’ theorem expresses the uncertainty of the hypothesis of interest, after having collected experimental data and making use of what is known about this hypothesis. Formally, Bayes’ theorem is given by

$$p(\text{hypothesis}|\text{data}) = \frac{p(\text{data}|\text{hypothesis})p(\text{hypothesis})}{p(\text{data})}.$$

The theorem reads: the hypothesis is strongly supported by the data ( $p(\text{hypothesis}|\text{data})$  is high) when there is a relative strong prior belief in the hypothesis ( $p(\text{hypothesis})$  is high) and/or the observed data fit well with the hypothesis ( $p(\text{data}|\text{hypothesis})$  is high). The probability  $p(\text{hypothesis}|\text{data})$  is called the *posterior probability*,  $p(\text{data}|\text{hypothesis})$  is known as the *likelihood* of the data, and  $p(\text{hypothesis})$  is the *prior probability* of the hypothesis.

A similar result can be formulated when we wish to know what the true value of a parameter  $\theta$  is after having done an experiment. To explain this, suppose we wish to know the prevalence of RA in 2012 in Turkey. Browsing the Internet reveals that the RA prevalence around the globe varies from 0.2 to 1 % (excluding specific Indian tribes), but no value was found for Turkey. From these historical data, one could postulate that the prevalence for Turkey must be around 0.5 % but with uncertainty. This uncertainty can be expressed by a distribution, called the *prior distribution* shown in Fig. 16a. This distribution expresses that, with 95 % (prior) probability, we believe that the prevalence of RA lies between 0.25 and 0.89 %. Suppose now



**Fig. 16** Prevalence RA: (a) prior distribution of  $\theta$  and (b) prior likelihood and posterior distribution of  $\theta$ , with  $\theta$  = %RA in Turkey in 2012

that we have done a limited survey in 2012 examining 500 subjects in Turkey and found 8 subjects with RA. This gives an estimated prevalence of 1.6 % with a 95 % CI=[0.82 %, 3.12 %]. The likelihood function in Fig. 16b is based on the survey data and summarizes what we know of  $\theta$ . This function is maximal for 1.6 % (hence, best supported prevalence value by the data), but also other not too different values for  $\theta$  are relatively well supported (corresponding with a relatively high likelihood value). Bayes' theorem in this case is

$$p(\theta|\text{data}) = \frac{p(\text{data}|\theta)p(\theta)}{p(\text{data})}.$$

Hence, as before, the posterior probability is high for those values of  $\theta$  that are well supported by the data, and the prior as can be inferred from Fig. 16b. The posterior distribution thus combines prior information with the information from the survey to arrive at a more precise statement on  $\theta$ . The posterior uncertainty of the prevalence reduces to [0.44, 1.12 %]. The posterior distribution also delivers summary measures for  $\theta$  characterizing its most likely value using the posterior mean, median, or mode and its (posterior) standard error. All can be computed from the posterior distribution.

There are three major aspects that distinguish the Bayesian approach from the classical frequentist approach:

- The Bayesian approach allows to include prior information into the analysis of data.
- In the Bayesian approach, the parameters have a distribution, which arises from the fact that we are always uncertain about the true value of that parameter.
- In the Bayesian approach we do not look at other possible samples as is done when computing the  $P$ -value. One says that the Bayesian approach is only based on the currently observed data; in other words, in the Bayesian approach, one conditions on the observed data.

While the classical frequentist approach is still most popular among clinicians, one might favor the Bayesian approach for the following reasons. After having done the experiment, the researcher invariably wishes to know how well his hypothesis is supported. As seen above, this is not given by the  $P$ -value, which only provides evidence against the observed results given the null hypothesis. In fact,  $p(\text{hypothesis}|\text{data})$  is needed, but this can only be obtained from a Bayesian analysis. Further, the classical 95 % CI is interpreted as the interval that contains with 0.95 probability the true value. However, this is not the technical definition that applies in the frequentist approach (see section “Type I error, type II error and the power of a test”) but has in fact a Bayesian flavor. Hence, the Bayesian approach may offer philosophical and conceptual advantages.

Nowadays, the Bayesian approach definitely offers to analyze more complex problems than the classical approach. However, it has taken more than 200 years before it was considered as a tool for the practical statistician, since for a long time, the approach could only be applied to (simple) textbook examples. Indeed, one must



realize that in realistic examples the parameter  $\theta$  quickly becomes a high-dimensional vector complicating the computation of the denominator  $p(\text{data})$  in Bayes' theorem. Indeed,  $p(\text{data}) = \int p(\text{data})p(\theta)d\theta$  involves the evaluation of an integral which can be quite complicated for high dimensions and often impossible to compute with classical numerical techniques. In that case, the posterior distribution cannot be determined and no inference is available. In other words, if the integral in the denominator cannot be computed, then the Bayesian approach cannot be applied.

A breakthrough was achieved by Gelfand and Smith [28] who suggested using a sampling technique to replace the integral calculations. The development of these *Markov chain Monte Carlo sampling techniques*, together with the development of the corresponding (Win)BUGS software [29], led to the great popularity of the Bayesian approach nowadays. The reason is that the sampling approach (together with WinBUGS and other recently developed Bayesian software) allows analyzing in principle any complex problem. We refer to the literature, especially the statistical literature (e.g., [30]), to appreciate the strength of the Bayesian approach, since space restrictions prevent us to illustrate its elegance and power.

Despite its increasing popularity, the Bayesian approach is still criticized by many because it needs a prior distribution to make the computations happen. This prior distribution is inevitably (somewhat) subjective and therefore always a (small) subjective component creeps into a Bayesian analysis. The Bayesians argue that research is always somewhat subjective (and should be, otherwise it cannot be research). Secondly, they argue that (1) the prior distribution can be made so uninformative that it almost does not carry any prior information at all, (2) one can always vary the prior to evaluate its effect to see how much the posterior distribution is ruled by the prior information and how much by the data at hand, and (3) most often the information from the data dominates the prior information. What is important to realize is that the Bayesian approach offers a tool to combine prior knowledge with current data and hence mimics in this way how scientists organize their research and how humans in general go through life.

## Statistical Guidelines

Motivated by the need to improve the standards of the methodology in clinical research, several guidelines to improve clinical research have been published in the literature. The earliest, and perhaps most well known, are the *CONSORT guidelines*. On the website <http://www.consort-statement.org/>, we can read that “CONSORT, which stands for Consolidated Standards of Reporting Trials, encompasses various initiatives developed by the CONSORT Group to alleviate the problems arising from inadequate reporting of randomized controlled trials (RCTs).” The main product of CONSORT is the *CONSORT statement*, which is an evidence-based, minimum set of recommendations for reporting RCTs. It offers a standard way for authors to prepare reports of trial findings, facilitating their complete and transparent reporting and aiding their critical appraisal and interpretation. The

CONSORT statement comprises a 25-item checklist and a flow diagram, along with some brief descriptive text. The checklist items focus on reporting how the trial was designed, analyzed, and interpreted; the flow diagram displays the progress of all participants through the trial.

The CONSORT guidelines constitute the start of a series of guidelines in different kinds of studies, such as *PRISMA* (guidelines for systematic reviews and meta-analyses), *STROBE* (guidelines for observational studies in epidemiology), etc. More guidelines can be found on the website of the *COCHRANE collaboration* (<http://www.cochrane.org/>). As we can read from the website: “The Cochrane collaboration is an international network to help healthcare practitioners, policy-makers, patients, their advocates and carers, make well-informed decisions about health care.”

The above guidelines encompass more than just statistical guidelines; in fact, their purpose is to improve clinical research on the whole from the design to the reporting and interpretation stage. Various other guidelines can be found on the world wide web. And of course many statistical textbooks contain also guidelines; see, e.g., [4, 5].

## Conclusions

This chapter had the intention to give a brief overview of the statistical methodology used to analyze clinical studies, with examples from two rheumatologic studies. It was only possible to discuss the topics briefly, and we had to refer to the reader to the literature where for each topic a multitude of books has been written. In addition, the statistical discipline has seen an explosion in the last five decades and especially in the last two decades due to the enormous evolution in computing power. Therefore, many topics were not addressed at all or could only touched upon briefly, such as with the large class of multivariate statistical techniques, the exploratory Bayesian approaches, etc. The explosion in the development of new statistical approaches will not and cannot stop, since the medical society is collecting increasingly more data and more complex data. And statistics is by excellence the science that aims to make sense out of these data.

## References

1. R Development Core Team. R: a language and environment for statistical computing [computer software]. Vienna: R Foundation for Statistical Computing; 2010.
2. SAS® version 9.3 Cary, NC, USA, SAS Institute Inc. 2012.
3. IBM Corp. Released 2012. IBM SPSS statistics for windows, version 21.0. Armonk: IBM Corp.
4. Bland M. An introduction to medical statistics. 3rd ed. Oxford: Oxford University Press; 2002.
5. Petrie A, Sabin C. Medical statistics at a glance. 3rd ed. Chichester: Wiley; 2009.

6. Walter MJM, Mohd Din SH, Hazes JMW, Lesaffre E, Barendregt PJ, Luime JJ. Is tight controlled disease activity with online patient reported outcomes possible? *J Rheumatol*. 2014;41:640–7.
7. Van der Heijde D, Jacobs J. The original “DAS” and the “DAS28” are not interchangeable: comment on the articles by Prevoo et al. *Arthritis Rheum*. 1998;41:942–50.
8. Van der Heijde D, Van’t Hof M, Van Riel R, et al. Judging disease activity in clinical practice in rheumatoid arthritis: first step in the development of a disease activity score. *Ann Rheum Dis*. 1990;49:916–20.
9. Bruce B, Fries J. The health assessment questionnaire (HAQ). *Clin Exp Rheumatol*. 2005;23: S14–8.
10. Fransen J, Langenegger T, Michel B, Stucki G. Feasibility and validity of the RADAI, a self-administered rheumatoid arthritis disease activity index. *Br Soc Rheumatol*. 2000;39:321–7.
11. Wolfe F. Fatigue assessments in rheumatoid arthritis: comparative performance of visual analog scales and longer fatigue questionnaires in 7760 patients. *J Rheumatol*. 2004;31: 1896–902.
12. Claessen SJ, Hazes JM, Huisman MA, van Zeven D, Luime JJ, Weel AE. Use of risk stratification to target therapies in patients with recent onset arthritis; design of a prospective randomized multicenter controlled trial. *BMC Musculoskelet Disord*. 2009;18(10):71.
13. de Jong PH, Hazes JM, Barendregt PJ, et al. Induction therapy with a combination of DMARDs is better than methotrexate monotherapy: first results of the tREACH trial. *Ann Rheum Dis*. 2013;72(1):72–8.
14. Visser H, le Cessie S, Vos K, Breedveld FC, Hazes JM. How to diagnose rheumatoid arthritis early: a prediction model for persistent (erosive) arthritis. *Arthritis Rheum*. 2002;46(2):357–65.
15. Aletaha D, Neogi T, Silman AJ, et al. Rheumatoid arthritis classification criteria: an American College of Rheumatology/European League Against Rheumatism collaborative initiative. *Arthritis Rheum*. 2010;62(9):2569–681.
16. Royall R. Statistical evidence. A likelihood paradigm. London: Chapman and Hall; 1997.
17. Svejgaard A, Ryder LP. HLA and disease associations: detecting the strongest association. *Tissue Antigens*. 1994;43:18–27.
18. Cox DR. Regression models and life-tables (with discussion). *J R Stat Soc B*. 1972;34: 187–220.
19. Rizopoulos D. Joint models for longitudinal and time-to-event data: with applications in R. Boca Raton: Chapman and Hall/CRC; 2012.
20. Little RJA, Rubin DB. Statistical analysis with missing data. 2nd ed. New York: Wiley; 2002.
21. Lesaffre E. Longitudinal studies in rheumatology: some guidance for analysis. *Bull NYU Hosp Jt Dis*. 2012;70(2):65–72.
22. Panel on Handling Missing Data in Clinical Trials; National Research Council. The prevention and treatment of missing data in clinical trials. Washington, DC: The National Academic Press; 2010.
23. Verbeke G, Molenberghs G. Linear mixed models for longitudinal data. New York: Springer; 2000.
24. Molenberghs G, Verbeke G. Linear models for discrete longitudinal data. New York: Springer; 2005.
25. Lawton G, Bhakta BB, Chamberlain MA, Tennant A. The Behçet’s disease activity index. *Rheumatology*. 2004;43:73–8.
26. Molenberghs G, Kenward M. Missing data in clinical studies. West Sussex: Wiley; 2007.
27. Tunc R, Keyman E, Melikoglu M, Fresko I, Yazici H. Target organ associations in Turkish patients with Behçet’s disease: a cross sectional study by exploratory factor analysis. *J Rheumatol*. 2002;29(11):2393–6.
28. Gelfand AE, Smith AE. Sampling-based approaches to calculating marginal densities. *J Am Stat Assoc*. 1990;85:398–409.
29. Lunn DJ, Thomas A, Best N, Spiegelhalter D. WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics Comput*. 2000;10:325–37.
30. Lesaffre E, Lawson A. Bayesian biostatistics (statistics in practice). New York: Wiley; 2012.

Understanding Evidence-Based Rheumatology  
A Guide to Interpreting Criteria, Drugs, Trials,  
Registries, and Ethics

Yazici, H.; Yazici, Y.; Lesaffre, E. (Eds.)

2014, XII, 276 p. 32 illus., 20 illus. in color., Hardcover

ISBN: 978-3-319-08373-5