Table 23. *Mental Health Composite Scores in SF-12: An Example for Wilcoxon Signed-Rank Test*

Subject	Before Surgery Score: X_1	After Surgery Score: X_2	Sign of $X_1 - X_2$	$\|X_1 - X_2\|$	Rank of $\|X_1 - X_2\|$
1	80	65	+	15	9
2	60	72	−	12	8
3	55	62	−	7	2
4	70	66	+	4	1
5	85	95	−	10	5.5
6	83	92	−	9	4
7	66	74	−	8	3
8	52	92	−	40	10
9	73	62	+	11	7
10	80	90	−	10	5.5

general anesthesia or regional anesthesia during total knee replacement.

Let $o_{1,1}$, $o_{1,2}$, $o_{2,1}$, and $o_{2,2}$ denote the observed frequency of each combination of anesthesia type and incidence of nausea in cells $c_{1,1}$, $c_{1,2}$ $c_{2,1}$, and $c_{2,2}$, and $e_{1,1}$, $e_{1,2}$, $e_{2,1}$, and $e_{2,2}$ denote the corresponding expected frequency if there were no association, where the expected frequency:

$$e_{i,j} = \frac{\textit{row } i \textit{ total} \times \textit{column } j \textit{ total}}{\textit{total } (N)}, \quad i,j = 1,2$$

One way to assess the association between the row and the column variables is by measuring the difference between the observed and expected frequencies with a χ^2 test. The χ^2 test statistic is defined by the following:

$$\hat{\chi}^2 = \sum_{(i,j)} \frac{(o_{i,j} - e_{i,j})^2}{e_{i,j}}$$

The test can be applied to any row-by-column, or $m \times n$, contingency table, $m,n > 1$. Under the null hypothesis of no association between the row and the column variables, the test statistic follows a χ^2 distribution with $(m - 1)(n - 1)$ degrees of freedom.

Example 6: The expected frequencies in Table 24 are as follows:

$$e_{1,1} = \frac{90 \times 100}{180} = 50, \; e_{1,2} = \frac{90 \times 80}{180} = 40,$$

$$e_{2,1} = \frac{90 \times 100}{180} = 50, \; e_{2,2} = \frac{90 \times 80}{180} = 40$$

Hence,

$$\hat{\chi}^2 = \frac{(70-50)^2}{50} + \frac{(20-40)^2}{40} + \frac{(30-50)^2}{50} + \frac{(60-40)^2}{40} = 36$$

and the degree of freedom is $(2 - 1) \times (2 - 1) = 1$. The P value obtained with statistical software is less than .05, implying that patients who received general anesthesia during total knee replacement are more likely to have nausea than patients who received regional anesthesia.

Table 24. *Association Between Anesthesia Type and Incidence of Nausea*

Anesthesia Type	Incidence of Nausea Yes	No	Row Total
General	$O_{1,1} = 70\ (c_{1,1})$	$O_{1,2} = 20\ (c_{1,2})$	90
Regional	$O_{2,1} = 30\ (c_{2,1})$	$O_{2,2} = 60\ (c_{2,2})$	90
Column total	100	80	Total N = 180

Fisher Exact Test

When the expected frequency in any cell of a contingency table is less than 5, the χ^2 test becomes inaccurate and loses its power because it relies on large samples. For example, in a study comparing mortality rates between patients undergoing unilateral or bilateral knee replacement, the incidence of death is very low, resulting in highly unbalanced data allocations among the cells of the table. In such a case, the Fisher exact test is an alternative to the χ^2 test. Because, in general, the computation of the Fisher exact test is not feasible by hand, we avoid the detailed formula here.

McNemar test

If the χ^2 (or Fisher exact) test could be considered the independent 2-sample t test for categorical variables, the McNemar test is the counterpart of the

Table 25. *Case-Control Study of Cancer*

	Non-cancer patient		
	Smoker	Non-smoker	Row Total
Cancer patient			
Smoker	a	b	a + b
Non-smoker	c	d	c + d
Column total	a + c	b + d	Total N

paired *t* test for comparing dependent categorical variables. For example, the investigators studied the association between smoking and lung cancer in a case-control study where *N* cancer patients (cases) were matched with *N* non-cancer patients (controls) in Table 25 based on age, gender, location, and other related variables. In this case the χ^2 test and the Fisher exact test are not appropriate because they assume that the samples are independent.

The McNemar test is a modification of the χ^2 test, taking into account the correlation between the matched samples. Because the concordance cells where both case and control are smokers (a) or non-smokers (d) do not provide information about the association between cancer and smoking, the McNemar test only contains the frequencies in the disconcordance cells (b and c) and is defined as:

$$\hat{\chi}^2 = \frac{(|b-c| - 0.5)^2}{b+c}$$

The test statistic follows a χ^2 distribution with 1 degree of freedom under the null hypothesis of no association between cancer and smoking.

MULTIPLE-SAMPLE PARAMETRIC TESTS

These tests are used when comparing data sets among 3 or more groups. The ANOVA is used for normally distributed samples/data, whereas the Kruskal-Wallis test and GEE are appropriate for samples with normal or non-normal distributions.

One-Way ANOVA

A 1-way ANOVA is an alternative to the independent 2-sample *t* test for testing the equality of 3 or more means by use of variances.

The assumptions of ANOVA include the following:

- The samples are drawn from populations following normal distributions.
- The samples are independent.
- The populations have equal variances.

The null hypothesis of ANOVA is that all population means are equal, and the alternative hypothesis is that at least one population mean is different.

The basis of ANOVA is to partition the total variation into "between-group variation" and "within-group variation" and compare the two. These and other terms related to ANOVA are defined below.

Grand mean is the average of all sample values.

Between-group variation is the sum of squared differences between each group mean and the grand mean. The between-group variance is the between-group variation divided by its degrees of freedom. If there are *g* groups, the degrees of freedom is then equal to $g - 1$.

Within-group variation is the sum of squared differences between each sample and its group mean. The within-group variance is the within-group variation divided by its degrees of freedom. If there are *g* groups and *n* samples within each group, the degrees of freedom is then equal to $g(n - 1)$ or $N - 1$, where *N* is the total sample size.

Total variation is the sum of between-group variation and within-group variation.

The ANOVA is used to compare the ratio (*F* test statistic) of between-group variance to within-group variance. If the between-group variance is much larger

Table 26. *One-Way ANOVA*

Source	Sum of Squares (Variation)	Degrees of Freedom	Mean Square (Variance)	*F* Statistic
Between group	SSB	$g - 1$	$\text{MSB} = \frac{SSB}{g-1}$	$\frac{MSB}{MSW}$
Within group	SSW	$g\,(n - 1)$	$\text{MSW} = \frac{SSW}{g(n-1)}$	
Total	SST = SSB + SSW	$N - 1$		

Abbreviations: SSB, sum of squares between groups; SSW, sum of squares within groups; SST, total sum of squares; MSB, mean squares between groups; MSW, mean squares within groups; g, number of groups; n, number of samples within each group.

Table 27. *Two-Way ANOVA*

Source	Sum of Squares	Degrees of Freedom	Mean Square	*F* Statistic
Main effect 1	SS1	$g_1 - 1$	MS1 = SS1/*df*	MS1/MSW
Main effect 2	SS2	$g_2 - 1$	MS2 = SS2/*df*	MS2/MSW
Interaction effect	SS12	$(g_1 - 1)(g_2 - 1)$	MS12 = SS12/*df*	MS12/MSW
Within	SSW	$g_1 g_2 (n - 1)$	MSW = SSW/*df*	
Total	SST = SS1 + SS2 + SS12 + SSW	$g_1 g_2 n - 1$		

Abbreviations: SS1, sum of squares for Main Effect 1; SS2, sum of squares for Main Effect 2; SS12, sum of squares for interaction between Main Effect 1 and Main Effect 2; SSW, sum of squares within groups; SST, total sum of squares; MS1, mean squares for Main Effect 1; MS2, mean squares for Main Effect 2; MS12, mean squares for interaction between Main Effect 1 and Main Effect 2; MSW, mean squares within groups; g, number of groups; n, number of samples within each group.

than the within-group variance, then we conclude that the means are different. This is summarized in an ANOVA table (Table 26).

Two-Way ANOVA

In contrast to 1-way ANOVA, which tests the equality of population means in one variable, 2-way ANOVA is extended to assess the difference among population means in 2 independent variables or factors.

The 2-way ANOVA has the same assumptions as the 1-way ANOVA.

The null hypotheses in a 2-way ANOVA include:

- Main effect: The population means of each factor are equal.
- Interaction effect: There is no interaction between the 2 factors.

Similar to 1-way ANOVA, 2-way ANOVA partitions the total variation into 2 main effects or between-group variations, within-group variation, and interaction effects between the 2 factors. There is an *F* test for testing each main effect and the interaction effect. A similar table is created for 2-way ANOVA (Table 27).

MULTIPLE-SAMPLE NONPARAMETRIC TESTS

Kruskal-Wallis Test

The Kruskal-Wallis test is a generalization of the Mann-Whitney *U* test for testing the equality of 3 or more population medians and is a nonparametric alternative to 1-way ANOVA. Like other nonparametric tests, the Kruskal-Wallis test is based on the ranks of data and does not assume normality.

Assume there are g independent groups with n_i observations in the i group, $i = 1, 2, \ldots, n$. To calculate the Kruskal-Wallis test statistic, rank all data from the g groups with the smallest value obtaining a rank of 1. Ties are assigned average ranks. The test statistic is given by the following:

$$K = (n-1)\frac{\sum_{i=1}^{g} n_i(\bar{r}_i - \bar{r})^2}{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(r_{ij} - \bar{r})^2}$$

where r_{ij} is the rank among all data of observation j in group i, $\bar{r}_i$ is the mean rank of all observations in group i, and $\bar{r}$ is the mean rank of all observations across all groups.

The test statistic K follows a χ^2 distribution under the null hypothesis with $g - 1$ degrees of freedom. The P value can be obtained from the χ^2 distribution table.

Correlation Analysis

In addition to the previous statistical tests, we next briefly discuss correlation analysis. A variety of correlation coefficients are available and used to assess the relation between 2 or more random variables. We introduce 2 commonly used correlation coefficients, the Pearson correlation and Spearman rank correlation coefficients.

Pearson Correlation: Pearson correlation, also called Pearson product-moment correlation, was developed by Karl Pearson. It is applied to continuous variables and assumes a linear relation between 2 normally distributed variables. Pearson correlation lies in [–1, 1], with 1 (–1) indicating a perfect positive (negative) linear relationship. For a pair of independent variables, the Pearson correlation is 0.

Spearman Rank Correlation: Spearman rank correlation is a nonparametric correlation. When 2 variables are not normally distributed or do not have a linear relation, Spearman rank correlation is an alternative to Pearson correlation. Like those nonparametric tests we introduced earlier, Spearman rank correlation is also calculated based on ranks and therefore is not affected by the distribution of data.

ADVANCED STATISTICAL TESTS

In addition to those commonly used statistical tests, there are advanced statistical methods available for more complicated data settings. A couple of examples are presented here.

Regression Analysis

Regression analysis is a method for assessing the relation between a dependent variable and one or more independent variables. The most commonly used regression analysis is linear regression, which assumes a linear relation between the dependent and independent variables.

Repeated-Measures ANOVA

As with any ANOVA, repeated-measures ANOVA tests the equality of multiple means. However, repeated-measures ANOVA is used when the same group of random samples is measured under the same condition at multiple time points or under different conditions. It assumes data to be normally distributed and can be considered an extension of the paired *t* test to a sample with more than 2 repeated measures.

The GEE Method

The GEE method is for modeling clustered data and longitudinal data. When data are clustered dependent, the GEE allows for fitting the parameters of a generalized linear model without explicitly defining the correlation structure.

CONCLUSIONS

Statistical tests prove that observed differences are not due to random chance, providing scientific rigor to clinical and other experimental findings. Examples in this section show that specific tests have been developed to analyze most types of data sets that are of interest to the academic clinician-scientist. As outlined in this section, the appropriate test for a given data set is simple to determine based on 3 basic aspects of the data set(s): dimension (whether 2 or more groups are being compared), distribution (whether data are normally or non-normally distributed), and dependency (whether variables are dependent or independent). In the context of clinically relevant study design and interpretation of results, statistical tests establish nonrandom correlations that rigorously support efficacy, safety, or other outcomes of therapeutic interventions or other factors that are of interest to the clinician-scientist investigator.

SUGGESTED READING

Agresti A. *Categorical data analysis.* Hoboken, NJ: Wiley, 2002.

Rumsey D. *Statistical II for dummies.* Hoboken, NJ: Wiley, 2009.

Norman GR, Streiner DL. *PDQ statistics.* Ed 3. Shelton, CT: People's Medical Publishing House, 2003.

Pagano M, Gauvreau K. *Principles of biostatistics.* Ed 2. Pacific Grove, CA: Duxbury Press, 2000.

Kahn HA, Sempos CT. *Statistical methods in epidemiology.* Oxford: Oxford University Press, 1989.

Petrie A. Statistics in orthopaedic papers. *J Bone Joint Surg Br* 2006;88:1121-1136.

Hanley JA, Negassa A, Edwardes MD, Forrester JE. Statistical analysis of correlated data using generalized estimating equations: An orientation. *Am J Epidemiol* 2003;157:364-375.

Yan Ma, Ph.D.
Chisa Hidaka, M.D.
Stephen Lyman, Ph.D.

SECTION 14

Key Statistical Principles: The Nature of Data

The goal of clinical research is to use information collected from a sample of patients to answer questions about all patients with that condition or who receive that treatment. The statistical inference methods we use to do this requires that (1) the selected sample is representative of the population of interest and (2) we know something about the distribution of the data. If a variable's distribution approximates that of a known probability distribution, like the normal distribution, that has well-described parameters, then we can use our knowledge of these distributions to calculate the probability that our hypotheses are valid using parametric tests (described in section 13). When a variable does not follow such a distribution, we need to use different methods that do not rely on parameters to help us, using nonparametric tests (also described in section 13). This makes understanding our data important in developing the proper analysis plan for our study, and the point of this chapter is to:

1. describe the tools used to summarize data and learn about the distributions (descriptive statistics),
2. develop methods for estimating association between two variables (measures of Association),
3. describe how we use our knowledge of the parameters of probability distributions to confirm or negate our hypotheses (confidence intervals and *P* values).

DESCRIPTIVE STATISTICS

Every variable has an underlying distribution function that describes how the observations are spread over all possible values of the variable. This distribution is influenced by the type of variable: categorical (discrete) or continuous. In brief, continuous variables are those that can take on any value in a range of possible values, whereas categorical variables can only take on a specific set of values. Categorical variables can be further classified as ordinal, where the categories have a defined order (e.g., disagree, neutral, agree), or nominal, where there is no intrinsic order (e.g., gender [male, female]). The purpose of descriptive statistics is to generate a few measures that give us an idea of the particular features of the distribution of the variable of interest.

FREQUENCIES AND PERCENTAGES

A simple way to describe the distribution of a variable is to list all of the different values and the frequency of each value (i.e., the number of times the value occurs in the data set) (Table 28).

We can see that presenting frequencies and percentages in a table is an effective way to describe categorical data because there are a limited number of possible values. This method does not work as well for numerical variables because the range of possible values is often much larger and it is not practical to display all the observed values. One way to resolve this problem, if it makes sense for the analysis, is to group the values of the variable into a smaller number of defined categories and calculate the frequencies and percentages for these created categories. This is often done with variables such as age (e.g., ≤59 years, 60 to 79 years, and ≥80 years) and clinical laboratory values such as serum vitamin D level (where <20 ng/mL is deficient, 20 to 31 ng/mL is insufficient, and ≥32 ng/mL is sufficient), where groupings of the numeric data make sense clinically. If it does not make sense to categorize the variable, other methods are necessary to summarize the data.

NUMERICAL SUMMARY METHODS

Measures of Location

These are methods to describe the center of a distribution of continuous data. The mean and median are most common, although the mode is rarely used in special circumstances.

Mean: The mean of a variable is the sum of all values of the variable divided by the number of observations. In statistical notation this is represented by the following:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

So, using the variable age from our example data in Table 29, the mean age is as follows:

$$\text{Mean age} = (49 + 45 + 63 + 41 + 50 + 29 + 37 + 40 + 40 + 47)/10 = 441/10 = 44.1$$

The mean is simple to compute and has nice theoretical properties in terms of statistics that make it widely used. However, the mean is sensitive to extreme values, especially when the number of observations is small.

Median: The median of a variable is the central value when the data points are arranged in rank order, so that half of the data values are higher than the

TABLE 28. *Sex Distribution in Hypothetical Study Population*

	N	%
Sex		
Female	437	53.2
Male	384	46.8
Total	821	100.0

Table 29. *Age for 10 Hypothetical Research Subjects*

Subject	Age (yr)	Squared Deviation From Mean $(x_i - \bar{x})^2$
1	49	24.01
2	45	0.81
3	63	357.21
4	41	9.61
5	50	34.81
6	29	228.01
7	37	50.41
8	40	16.81
9	40	16.81
10	47	8.41
	Mean ($\bar{x}$), 44.1	Sum, 746.9

median value and the other half are lower than the median value. When there are an even number of observations in a data set, the median is defined as the midpoint between the 2 middle values. In our age example, the median is calculated as follows:

Age sorted lowest to highest: 29, 37, 40, 40, 41, 45, 47, 49, 50, 63

This data set has an even number of observations, so the 2 middle values are 41 and 45. The median is (41 + 45)/2 = 43.

The median requires the data to be sorted, so it is not as simple to compute as the mean, especially with larger sample sizes. It is not as sensitive to outlying values, though, so it may be a better measure of central tendency, especially for smaller samples.

Mode: The mode of a variable is the value that occurs most frequently. A multimodal variable has more than 1 value that meets this criterion. In our age example the mode is 40, because it occurs twice and all other values occur only once. This statistic is not often reported because it is usually not useful for describing continuous variables, which may have all unique values so that every value in the data set is a mode.

Measures of Variability/Dispersion

Range: The range is the difference between the largest and smallest values of a variable (Table 29). A wider range indicates more variability in the data. In our age data, the range is 63 – 29 = 34. The minimum and maximum values of a variable are more often reported than the range, however, because these 2 values also provide some information about the location of the extremes of a variable.

Variance: The variance of a data set, denoted s^2, is a measure of variability around the sample mean. The equation for variance is listed below. In words, the variance is the average of the squared deviations from the mean:

$$\text{Variance}(s^2) = \frac{1}{(n-1)}\sum_{i-1}^{n}(x_i - \bar{x})^2$$

$$= \frac{1}{(20-1)}\sum_{i-1}^{n}(x_i - \bar{x})^2$$

$$\text{Variance} = \frac{1}{(9)}(764.9) = 83.0$$

Standard Deviation: The standard deviation is the square root of the variance. This value is more often reported in descriptive tables because it is measured in the same units as the variable, and the mean and standard deviation together can tell us a lot about a distribution of values.

MEASURES OF ASSOCIATION

The objective of the study designs discussed in previous chapters is to compare outcomes in 2 or more groups, such as new treatment versus old treatment, exposure versus no exposure, and so on. The numerical summary methods we have discussed above are useful in describing individual variables, but now we need to define some measures of association that will let us compare groups.

Relative Risk (Risk Ratio)

What Is Relative Risk? In epidemiology, the incidence of an event (e.g., disease diagnosis or surgical complication) is the frequency of new events that occur during a specified time interval. What we call the "risk" of an event is the incidence rate, which is the incidence divided by the population at risk during that time interval.

$$\text{Incidence rate} = \frac{\text{Number of new events}}{\text{Population at risk for event}}$$

A related concept is the prevalence of the event, which is the sum of events that have already occurred plus new incident events divided by the population total. So we can see that the incidence rate is a measure of the risk of disease whereas prevalence shows the burden of disease in a population.

As formulas, incidence and prevalence can be described as follows:

Incidence = Number of knee replacement
surgeries performed this year

Table 30. *Two-by-Two Frequency Table*

	Outcome		
	Yes	No	
Treatment A	a	b	a + b
Treatment B	c	d	c + d
	a + c	b + d	

$$\text{Incidence rate} = \frac{\text{Number of Knee arthroplasties performed this year}}{\text{Number of people in population}}$$

$$\text{Prevalence rate} = \frac{\text{Number of people living with knee arthroplasty}}{\text{Number of people in population}}$$

What we usually want to do in clinical research is to compare the risk between groups, so an easy way to do this is to simply determine the ratio of the risk (whether incidence or prevalence) for the 2 groups:

$$\text{Relative risk} = \frac{\text{Risk in group 1}}{\text{Risk in group 2}}$$

How to Calculate the Relative Risk in a Cohort Study: When preparing data for a comparison of 2 treatments, we can most easily calculate relative risk by creating a 2 × 2 table (Table 30). The term "2 × 2" refers to the numbers of rows and columns in the table—2 possible outcomes and 2 possible treatments in the example in Table 30.

We can calculate the likelihood (or risk) of the outcome for each treatment group as follows:

$$\text{Risk}_{\text{TXA}} = \frac{A}{a + b}$$

$$\text{Risk}_{\text{TXA}} = \frac{C}{c + d}$$

We can then calculate a relative risk of the outcome in patients receiving treatment A versus treatment B:

$$\text{Relative risk} = \frac{\text{Risk}_{\text{TXA}}}{\text{Risk}_{\text{TXB}}} = \frac{a/(a + b)}{c/(c + d)}$$

How to Interpret the Relative Risk: If the relative risk equals 1, then the risk is the same in both groups and there does not appear to be an association. When the relative risk is greater than 1, the risk in group 1 is greater than that in group 2; this is usually described as evidence of an increased risk, or positive association. If the relative risk is less than 1, the risk in group 1 is less than that in group 2; this is usually described as indicating a negative association, or a decreased risk.

For example, if we had a cohort study that wanted to determine whether a new form of pain management reduced the incidence of postoperative pain, we would generate a contingency table from our data (Table 31). The relative risk is less than 1, which indicates that the new pain management technique reduces the incidence of postoperative pain.

Odds Ratio (Relative Odds)

Calculating relative risk requires us to know the incidence rate for a population, which is not possible for some study designs. In a case-control study, for example, the 2 groups are based on outcome status, so we do not know the population at risk. Thus we need another measure of association that will work for both cohort and case-control studies. For this type of study, we can calculate a different measure of association, using the odds.

What Is an Odds Ratio? The odds of an event is defined as the ratio of the probability that an event occurs to the probability that the event does not occur. If we represent the probability that event A occurs by P, then the probability that event A does not occur is 1 – P. So the odds of event A is as follows:

$$\text{Odds} = \frac{P}{1 - P}$$

For example, when rolling a die, the probability of rolling a 1 or 2 is 2/6 = 1/3 = 33.3%, so the odds of rolling a 1 or 2 is as follows:

$$\text{Odds} = \frac{33.30\%}{66.70\%} = 0.50$$

It is important to note that the probability of rolling a 1 or 2 (33.3%) and the odds of rolling a 1 or 2 (0.50) are 2 distinct measures.

How to Calculate the Odds Ratio: Now suppose we have a study as in the contingency table for a

Table 31. *Sample Data for Relative Risk*

	Postoperative Pain	No Postoperative Pain	Total
New pain management	4	21	25
Old pain management	9	16	25
Total	14	36	

NOTE. Risk in patients with new technique = 4/25 = 0.16. Risk in patients with old technique = 9/25 = 0.36. Relative risk = 0.16/0.36 = 0.44.

cohort study as shown in Table 32. In a cohort study we are comparing the odds of event A in the exposed group with the odds of event A in the non-exposed group.

First, we need to calculate the probability (P) of event A for group 1:

$$P = \mathrm{a}/(\mathrm{a} + \mathrm{b})$$

Next, we will calculate the odds of event A for group 1:

$$\text{Odds} = [\mathrm{a}/(\mathrm{a} + \mathrm{b})]/[\mathrm{b}/(\mathrm{a} + \mathrm{b})] = \mathrm{a}/\mathrm{b}$$

Similarly, the odds of event A for group 2 equals c/d. Finally, the odds ratio for group 1 versus group 2 is (a/b)/(c/d) = ad/bc (Table 33).

In a case-control study, first, we need to calculate the odds that a case had a history of exposure ($\text{Odds}_{\text{cases}}$):

$$\text{Odds}_{\text{cases}} = [\mathrm{a}/(\mathrm{a} + \mathrm{c})]/[\mathrm{c}/(\mathrm{a} + \mathrm{c})] = \mathrm{a}/\mathrm{c}$$

Next we calculate the odds that a control had a history of exposure ($\text{Odds}_{\text{controls}}$):

$$\text{Odds}_{\text{controls}} = [\mathrm{b}/(\mathrm{b} + \mathrm{d})]/[\mathrm{c}/(\mathrm{b} + \mathrm{d})] = \mathrm{b}/\mathrm{d}$$

$$\text{Odds ratio} = \text{Odds}_{\text{cases}}/\text{Odds}_{\text{controls}} = (\mathrm{a}/\mathrm{c})/(\mathrm{b}/\mathrm{d}) = \mathrm{ad}/\mathrm{bc}$$

$$\text{Odds ratio} = \frac{(\mathrm{a/c})}{(\mathrm{b/d})} = \frac{\mathrm{ad}}{\mathrm{bc}}$$

Note that the formula for the odds ratio is the same for both cohort and case-control studies.

How to Interpret the Odds Ratio: Similar to the interpretation of the relative risk, an odds ratio of 1 indicates that the exposure is not related to the event.

If the odds ratio is larger than 1, then the exposure is positively associated with the event, and if the odds ratio is less than 1, the exposure is negatively associated with the event.

Using the Odds Ratio to Estimate Relative Risk: The odds ratio is itself a useful measure of association, but there may be situations when reporting the relative risk is preferred. In a case-control study, although the relative risk cannot be directly calculated, the odds ratio is a good approximation of the relative risk when the cases and controls are representative samples of the populations from which they are drawn and the outcome is infrequent.

TABLE 32. *Contingency Table for a Cohort Study*

	Event A		
	Yes	No	
Exposed	a	b	a + b
Not exposed	c	d	c + d
	a + c	b + d	

TABLE 33. *Odds of Event A for Group 1*

	Event A		
	Cases	Controls	
History of exposure	a	b	a + b
Not exposed	c	d	c + d
	a + c	b + d	

We will use examples of a cohort study, where both the relative risk and odds ratio can be directly calculated, to see when the odds ratio is a good estimate of the relative risk.

When event is infrequent:

	Event A		
	Yes	No	
Exposed	25	975	1,000
Not exposed	10	990	1,000
	35	1,965	

$$\text{Relative risk} = \frac{25/1,000}{10/1,000} = \frac{25}{10} = 2.50$$

$$\text{Odds ratio} = \frac{25 \times 990}{10 \times 975} = \frac{24,750}{9,750} = 2.54$$

When event is frequent:

	Event A		
	Yes	No	
Exposed	250	750	1,000
Not exposed	100	900	1,000
	350	1,750	

$$\text{Relative risk} = \frac{250/1,000}{100/1,000} = \frac{250}{100} = 2.50$$

$$\text{Odds ratio} = \frac{25 \times 990}{10 \times 750} = \frac{24,250}{9,900} = 3.00$$

MEASURES OF PROBABILITY

Understanding the properties of a distribution allows us to apply this knowledge to the first steps of understanding statistical inference, which is the process of drawing conclusions about an entire population based on the information from a sample of that population. Recall that it is our goal to describe or make an educated estimate of some characteristic of a continuous variable using the information from our sample of observations.

There are 2 ways of estimating these characteristics. "Point estimation" involves taking the sample data and calculating a single number, such as the mean, to estimate the parameter of interest. However, the inherent problem of calculating 1 mean from 1 sample of a population is that drawing a second sample and calculating its mean may yield a very different value. The point estimate does not take into account the inherent variability that exists between any combinations of samples that are drawn from all populations. To account for this variability, a second technique, called "interval estimates," provides a reasonable range of values that are intended to contain the parameter of interest with a certain degree of confidence. This range in values is called a confidence interval (CI).

The CI allows us to evaluate the precision of a point estimate by calculating an interval that contains the true population mean with a planned degree of certainty. For the 95% CI, we are 95% confident that the true population mean lies somewhere between the upper and lower limits calculated. Another way to understand the 95% CI is as follows: if we were to select 100 random samples from a population and use these samples to calculate 100 different intervals for these samples, 95 of these intervals would cover the true population mean (whereas 5 would not).

A few ways not to interpret the CI is to state that the probability of the calculated mean lies between the upper and lower limits of the calculated interval. In addition, it would be incorrect to state that there is a 95% chance that the mean is between the upper and lower limits of the calculated interval.

Suppose, for example, that we were looking to find the CI for serum cholesterol for all men in the United States who are hypertensive and smoke. If the mean serum cholesterol level in a sample of 12 hypertensive men is 217 mg/100 mL with a standard deviation of 46 mg/100 mL, what is the 95% CI for this calculated mean? To calculate the upper and lower bounds, we first use the equation for the interval for a continuous variable:

$$\bar{X} \pm 1.96\left(\frac{\sigma}{\sqrt{n}}\right)$$

where X is the calculated mean, σ is the standard deviation, and n is the sample population. When the values are plugged into the equation, we end up with a lower limit of 191 and an upper limit of 243. Although the interval values calculated appear to indicate a fairly precise mean, what would you imagine would happen if we were able to increase the sample size of the sample we collected? Imagine that all that changed from this sample was only the number of our sample. If the sample size was increased from 12 to 50, the CI now changes to 204 in the lower limit and 230 in the upper limit. As you can see, the sample size plays an important role in the precision of our estimates. The more people we have in our study, the more narrow the range, which in turn increases our accuracy.

As stated earlier, the bounds of the CI give us an important indicator of the precision of the calculated mean. Therefore the more narrow the CI, the more precise the estimate. The CI is an important and extremely helpful way of evaluating an estimate and, if possible, should always be reported whenever an estimate is provided in the results. Whereas the standard deviation gives the reader an idea of the spread of the values around the mean, the CI provides the reader the precision of the estimate.

CONCLUSIONS

Up until now, the chapters of this book have focused on designing studies. This chapter begins to explore what to do with the data once they have been collected. The first step should be to describe each variable. For continuous variables, we calculate the appropriate measures of central tendency and spread or dispersion. For categorical variables, we create frequency tables and calculate percentages for each stratum within each variable. Next, we calculate measures of association through either a relative risk if we know the underlying distribution or an odds ratio if we have conducted a case-control study. Finally, we calculate measures of probability. This can be done through hypothesis testing as described in section 13 but also through the calculation of CIs, which gives us a different perspective on the probability underlying our data.

SUGGESTED READING

Hennekens CH, Buring JE, Mayrent SL. *Epidemiology in medicine.* Philadelphia: Lippincott Williams & Wilkins, 1987.

Rothman KJ, Greenland S, Lash TL. *Modern epidemiology.* Ed 3. Philadelphia: Lippincott Williams & Wilkins, 2008.

Hulley SB, Cummings SR, Browner WS, Grady D, Hearst N, Newman TB. *Designing clinical research.* Ed 2. Philadelphia: Lippincott Williams & Wilkins, 2001.

Dorey F, Nasser S, Amstutz H. The need for confidence intervals in the presentation of orthopaedic data. *J Bone Joint Surg Am* 1993;75:1844-1852.

Gordis L. *Epidemiology.* Philadelphia: Elsevier Health Sciences, 1996.

Morshed S, Tornetta P III, Bhandari M. Analysis of observational studies: a guide to understanding statistical methods. *J Bone Joint Surg Am* 2009;91: 50-60 (Suppl 3).

Petrie A. Statistics in orthopaedic papers. *J Bone Joint Surg Br* 2006;88:1121-1136.

Varkevisser C, Pathmanathan I, Brownlee A. Designing and conducting health systems research projects, volume 2. Data analyses and report writing. KIT, IDRC: 2003.

Pagano M, Gauvreau K. *Principles of biostatistics.* Pacific Grove, CA: Duxbury Press, 2000.

Huong T. Do, M.A.
Joseph Nguyen, M.P.H.
Stephen Lyman, Ph.D.

SECTION 15

Survival Analysis in Orthopaedic Surgery: A Practical Approach

Survival analysis is an effective statistical tool for evaluating and comparing outcomes of orthopaedic procedures. This method is based on constructing a life-table of a cohort of patients after certain orthopaedic procedures. The life-table contains all the data relevant for the determination of the cohort at regular follow-up periods. The main outcome value in the life-table is the cumulative survival of the study group at each time interval with provision of 95% CIs of distribution of cumulative survival values. The calculation of these values is based on the recognition of a number of patients who were lost to follow-up and determination of the uniform criteria for patients with failed outcome. If the latter parameters are similar in different studies, a comparison of survival values can be performed by the log-rank test.

To evaluate the clinical outcome of orthopaedic procedures, 2 important and unique characteristics should be addressed: the relatively limited number of patients (<100 patients in most studies) and the term of follow-up (usually several years). These requirements might challenge the effectiveness of traditional statistical tools for comparison of medical or surgical treatments used in other clinical areas, with involvement of large cohorts of patients with clear short-term outcomes that remain unchanged for long time periods. To answer this specific need, orthopaedic procedures are evaluated and compared by use of survival analysis, which has been especially adapted to the field of orthopaedic surgery. Initially, this method was developed for the long-term follow-up of prosthetic implants,[205] but it can also be used for other orthopaedic procedures.[206]

There are 2 main methods for survivorship analysis. In the classic "product limit method" according to Kaplan and Meier, the survival (i.e., the success of the procedure) changes immediately after clinical failure.[207] Using this method in relatively small groups of evaluated patients, the CIs at the change points of the survivorship might be misleadingly overestimated or even show values above 100%.[208] Therefore, for more reliable evaluation of orthopaedic procedures with relatively small groups of patients who are followed up at constant time intervals, for example, on an annual basis in the arthroplasty follow-up, a need for special adaptation of this method is apparent. Exactly for this purpose, Murray et al.[208] popularized a method of survivorship analysis based on construction of a "life-table" with the assumption that all the procedures were performed at the same time 0 and the patients

TABLE 34. *Life-Table of Patients Operated on in 1989-1994 With BioModular Uncemented Total Shoulder Prosthesis*[212]

						Withdrawn at Last Review				
Postoperative Year	No. at Start	Success	Lost	Died	Failed	No. at Risk	Proportion Failing (%)	Proportion Succeeding (%)	Cumulative Survival (%)	95% Confidence Interval
1	90	0	0	1	5	89.5	5.6	94.4	94.4	87.6-97.6
2	84	0	0	1	3	83.5	3.6	96.4	91	83.1-95.4
3	80	0	0	0	4	80	5.0	95.0	86.5	77.6-92.2
4	76	0	0	0	5	76	6.6	93.4	80.7	70.9-87.8
5	71	0	0	2	2	70	2.9	97.1	78.4	68.1-86.0
6	67	0	0	2	2	66	3.0	97.0	76.1	65.5-84.3
7	63	1	1	4	2	60	3.3	96.7	73.5	62.4-82.2
8	55	7	0	2	0	50.5	0	100	73.5	62.1-82.4
9	46	4	0	1	1	43.5	2.3	97.7	71.8	59.9-81.3
10	40	18	0	0	1	31	3.2	96.8	69.5	56.8-79.8
11	21	9	0	1	0	16	0	100	69.5	55.3-80.7

NOTE. Postoperative years 1, 7, and 8 represent the data discussed in the text.

re-evaluated at constant intervals, taking into consideration patients who were lost to follow-up, thus establishing a cumulative success rate for each time interval. Subsequently, according to these considerations, 95% CIs of survival were determined. In this method 95% CIs are more appropriate for a small group of patients and never exceed 100% of survivorship.

VARIABLES

As an example of a life-table (Table 34), we use data published on survival analysis of 90 patients after total shoulder arthroplasty.[209] According to the method presented here, the main outcome values are the cumulative survival rates for each time period with 95% CI distribution of these values. The survival values can be presented graphically as survival curves. In addition to these final outcome values, the life-table includes all the parameters that are required for the calculation of the main outcome values; thus it contains all the data for independent evaluation of survivorship outcome, enabling critical review by readers and an ability to compare outcomes with other studies. The calculation method is shown in rows 1, 7, and 8 in Table 34.

TIME PERIODS OF FOLLOW-UP

In the first column of the life-table, the follow-up periods are given. As has been noted, the main characteristic of the presented survival analysis is the constant periods between patient evaluations according to the nature of the surgical procedure. In the presented example, because the life-table deals with the outcome of shoulder arthroplasty, 1 year between follow-up evaluations is a commonly used practice. Because the purpose of the survival analysis, among others, is a comparison between different cohorts of patients, the use of the established follow-up period for the particular procedure is recommended. An additional basic assumption of this method is that all the patients were treated at time 0. This does not mean that all the patients actually underwent surgery on the same date, but the date of the surgery for each patient is considered as time 0, after which all the calculations are performed. Accordingly, in row 1 of the life-table, the first column contains the values of 1 year; in row 7, the value of 7 years; and in row 8, the value of 8 years (i.e., 1, 7, and 8 years of follow-up).

NUMBER OF PATIENTS REMAINING FOR FOLLOW-UP AT EACH PERIOD (NUMBER AT START)

The number of patients at the start represents the number of patients who were available for evaluation at each time period. This value is a product of subtraction of the number of patients who were withdrawn from the number of patients at the start in the previous time period. Note that the number at the start in the first row (i.e., in the first time period) represents the total number of patients enrolled in the study. The number of patients withdrawn for each time period is the sum of values given in columns 3, 4, 5, and 6 (success, lost, died, and failed). The method to determine these values is given in the next section. Therefore, in our example, in year 1, the number of patients at the start was 90 (the entire cohort). In year 7, this value is 63, when the 4 patients "withdrawn at last review" (0 + 0 + 2 + 2 = 4) were subtracted from the original number; there were 67 patients in row 6. Similarly, in row 8, the original number of patients is the product of subtraction of 8 patients (1 + 1 + 4 + 2 = 8 "withdrawn at last review" in row 7) from 63 patients, which is the original number of patients in row 7, giving a value of 55 patients.

WITHDRAWN AT LAST REVIEW

This section requires special attention because it is based on assumptions that can influence the entire life-table and can be manipulated according to special characteristics of the study group. This section contains 4 subsections (4 columns)—success, lost, died, and failed—which will be discussed separately.

Success

This might be misleading terminology, but it means that the patients reached their maximal follow-up time period and should be considered for withdrawal in the discussion of the next time period of the survival analysis. For example, in row 7, 1 patient reached the maximal follow-up of 7 years; therefore he cannot be discussed as part of the group of patients in row 8. In addition, from inspection of the life-table, the "success" column indicates the minimal follow-up time in the studied group and the number of patients who did not reach the maximal follow-up period, excluding those who were lost to follow-up and died, and at what quantitative extent. By looking at our example, we see that only 9 patients reached the whole 11-year period of follow-up, as indicated in row 11, and the minimal follow-up time was 7 years, because the first "success" is indicated in row 7.

Lost

The patients who were lost to follow-up are the main factor of uncertainty of a life-table and survival analysis. The designers of this method reasonably argued that this group might have a higher proportion of unsatisfied persons with failed procedures.[210] We will address this topic in the following sections.

Died

Two factors are crucial in the estimation of this group. It must be verified at the highest possible extent that the cause of death is unrelated to the procedure for which survival analysis is performed, because in that case the patient should be included in the "failed" group. In addition, maximal effort should be exerted to verify that the persons who have died are not included in the "lost-to-follow-up" group. The reason for the latter is that the proportion of failures in patients who died might be overestimated.[210] This might affect the other parameters of the life-table, as will be discussed later.

Failed

The way these data are filled is determined by the survival analysis constructor and has the highest potential to be biased. Unfortunately, because different authors consider different criteria for determination of failure of the studied procedure, their life-tables might be difficult for meaningful comparison. The minimalistic approach for determination of failure and the most often used is eventual revision surgery. The maximalistic approach might involve clinical signs on imaging modalities, such as radiographic signs of prosthesis loosening, a certain level of pain, restricted range of movements, and so on, without surgery. These signs can also be the reason for the decision on revision surgery[209] and become part of the minimalistic approach. Therefore a clear definition of the criteria of "failure" should be provided. It is also possible to perform a survival analysis with different failure definitions on the same group of patients to compare life-tables from different sources.

NUMBER OF PATIENTS AT RISK

This variable reflects the number of patients who are actually considered for evaluation in the certain period of time, according to the life-table design. These patients were available for follow-up at a certain time period and therefore were determined as a product of subtraction of unavailable patients, meaning those who died, were lost, or reached the end of their follow-up (success), from the total number of patients at the start of this time period. These patients at risk can reach clinical failure as discussed before, and would be removed from further follow-up, or could be considered as successes and be followed up in the next time period. The fact that not all of the subtracted individuals were exposed to the risk during the total time period should be taken into consideration. It will be impossible to know the exact fraction of these patients; therefore a reasonable estimation of 50% is used, and subtraction of only half of the withdrawn patients is implemented for the life-table. In the example in Table 34, the number at risk in row 1 was 89.5 after subtraction of 0.5 [(0 success + 0 lost + 1 died)/2 = 0.5] from 90 (number at start).

PROPORTION OF FAILING

This is a proportional value of failed cases from the number at risk. It is usually represented in percentages. In our example (Table 34), in postoperative year 7, the proportion of failing was 3.3% (2 [failed]/60 [number at risk] × 100 = 3.3%).

PROPORTION OF SUCCEEDING

Naturally, the proportion of succeeding will be the remainder value from the proportion of failing to 100%. So, during the seventh postoperative year, the proportion of succeeding is 96.7% (100% – 3.3% [proportion failing] = 96.7%).

CUMULATIVE SURVIVAL

This is the main outcome value of the life-table and can be later represented graphically as a survival estimation for the given time period.[206] Because it is cumulative in definition, this value is calculated by multiplying the proportion succeeding in the given time period by the cumulative survival proportion in the previous time period, expressed in percentages. In the first time period, the cumulative survival proportion is equal to the proportion of succeeding, because we consider the initial cumulative survival of the procedure as 100%, as expressed in the example in Table 34. Another example is the cumulative survival of 73.5% in postoperative year 8 (1 [proportion of succeeding in year 8] × 0.735 [cumulative survival in year 7] × 100 = 73.5%).

95% CONFIDENCE INTERVAL

The last column to be filled in the life-table contains the CIs of the cumulative survival and represents distribution of 95% of these values for every time period. The calculation of the CI for a given time interval is based on determination of the "effective number of risk" (M), which contains information on the number of patients at risk from the previous time intervals according to the following formula:

$$M = i/\sum 1/n_i$$

where i is the time interval and n is the number of patients at risk in the time interval i.[208,210]

Accordingly, the confidence limits (CL) are calculated according to the following formula[209,211]:

$$CL = \frac{M}{M + 1.96^2}\cdot\left[P + \frac{1.96^2}{2\cdot M} \pm 1.96\sqrt{P\cdot\frac{(1-P)}{M} + \left(\frac{1.96^2}{4\cdot M^2}\right)}\right]$$

when M is an effective number at risk and P is cumulative survival at the given time interval (expressed as proportion and not as percentage). This mathematical expression is based on the theoretical assumption presented by Rothman[212] and popularized by Murray et al.[208] The mathematical basis of these assumptions will not be discussed in this presentation, which is more of a practical nature. The interested reader is referred to these extensive statistical reports that are given in the "References" section.

As an example of the calculations of the CIs, we will refer to time interval 8 (i = 8 [postoperative year 8]) (Table 34). The M value is 69.739 according to the following calculation:

$$\frac{8}{\frac{1}{89.5}+\frac{1}{83.5}+\frac{1}{80}+\frac{1}{76}+\frac{1}{70}+\frac{1}{66}+\frac{1}{60}+\frac{1}{50.5}}$$
$$= 69.739$$

The values of the CI are calculated as follows (M = 69.739, P = 0.735).

For the upper limit,

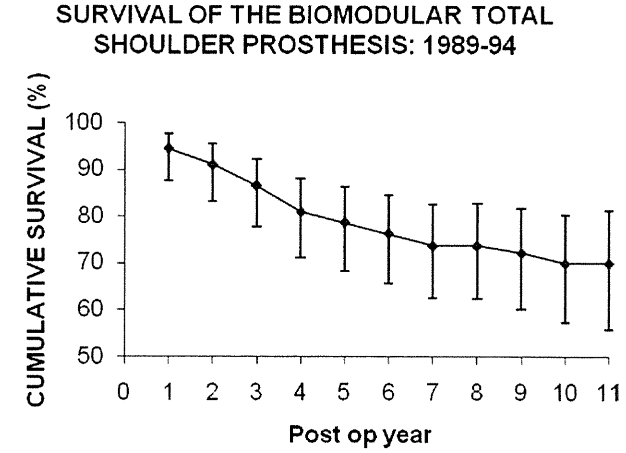

Figure 11. Graphic representation of outcome values of survival analysis given in Table 34. Vertical bars represent 95% confidence intervals of the cumulative survival rates.

TABLE 35. *Life-Table of Patients With Shoulder Osteoarthritis Operated on in 1989-1994 With BioModular Uncemented Total Shoulder Prosthesis*[212]

Postoperative Year	No. at Start	Success	Lost	Died	Failed	No. at Risk	Proportion Failing (%)	Proportion Succeeding (%)	Cumulative Survival (%)	95% Confidence Interval
						Withdrawn at Last Review				
1	48	0	0	1	4	47.5	8.4	91.6	91.6	80.3-96.7
2	43	0	0	0	3	43	7	93	85.2	72.1-92.8
3	40	0	0	0	3	40	7.5	92.5	78.8	64.9-88.2
4	37	0	0	0	3	37	8.1	91.9	72.4	57.4-83.6
5	34	0	0	0	1	34	2.9	97.9	70.9	55.5-82.7
6	33	0	0	0	2	33	6.1	93.9	66.6	50.8-79.4
7	31	0	0	3	1	29.5	3.4	96.6	64.3	48.2-77.7
8	27	3	0	1	0	25	0	100	64.3	47.7-78.1
9	23	1	1	0	1	22	4.5	95.5	61.4	44.4-76
10	20	10	0	0	0	15	0	100	61.4	43.4-76.7
11	10	2	0	1	0	8.5	0	100	61.4	41.6-78

$$\frac{M}{M+1.96^2}\cdot\left[P+\frac{1.96^2}{2\cdot M}+1.96\sqrt{P\cdot\frac{(1-P)}{M}+\frac{1.96^2}{4\cdot M^2}}\right]=0.824$$

For the lower limit,

$$\frac{M}{M+1.96^2}\cdot\left[P+\frac{1.96^2}{2\cdot M}-1.96\sqrt{P\cdot\frac{(1-P)}{M}+\frac{1.96^2}{4\cdot M^2}}\right]=0.621$$

Therefore the 95% CI for the cumulative survival of 73.5% in postoperative year 8 (Table 34) is between 62.1% and 82.4%.

At this stage, when all the data are entered into the life-table, the main outcome values, cumulative survival and its 95% CIs, can be presented graphically (Fig 11).

COMPARISON BETWEEN SURVIVAL ANALYSES

The last step of the process of evaluating the results in the life-table is the ability to compare it with the results of other survival analyses. It is clear that the prerequisite for such comparison will be the same determination for "failure" in the compared life-tables, similar numbers of "lost to follow-up," and a similar method of life-table construction.

For comparison of 2 life-tables with a relatively small number of patients with low failure rates, the log-rank test is usually used.[206] The null hypothesis of this type of comparison is the same proportion of failures in every time interval for 2 compared treatments. Using this test, we will be able to compare the occurrence of failures in the 2 survival analyses in question. For this purpose, a χ^2 statistic is calculated. For comparing 2 life-tables, the χ^2 distribution of values with 1 degree of freedom is assumed.[213] In this case the value of χ^2 above 3.841 indicates a P value below .05; when the value of χ^2 is above 6.635, the P value is below .01; and when the value of χ^2 is above 10.828, the P value is below .001.[213] We will demonstrate the calculations by using 2 life-tables (Tables 34 and 35).

For calculation of the χ^2 statistic according to the log-rank test, additional variables are determined and summarized (Table 36). "Postoperative year," "Number at risk" and "Observed failure" are taken from the life-tables that are compared.

Total number at risk" is the sum of "Number at risk" from the 2 life-tables for each postoperative year. For example, for year 7, this value is 89.5 (60 [Table 34] + 29.5 [Table 35]).

"Expected failure" for each of the life-tables for every postoperative year is calculated according to the following formula:

(Observed failure)
× (Number at risk)/(Total number at risk)

In our example, in postoperative year 7 in Table 35, this value is 0.33 (1 [observed failure] × 29.5 [number at risk]/89.5 [total number at risk]).

After the previously described variables are determined, the χ^2 statistic can be calculated for each of the life-tables according to the formula (Observed failures

Table 36. *Variables Required for Comparison of Survival Data in Tables 34 and 35*

Postoperative Year	No. at Risk: Table 34	Observed Failure: Table 34	No. at Risk: Table 35	Observed Failure: Table 35	Total No. at Risk	Expected Failure: Table 34	Expected Failure: Table 35
1	89.5	5	47.5	4	137	3.27	1.39
2	83.5	3	43	3	126.5	1.98	1.02
3	80	4	40	3	120	2.67	1.00
4	76	5	37	3	113	3.36	0.98
5	70	2	34	1	104	1.35	0.33
6	66	2	33	2	99	1.33	0.67
7	60	2	29.5	1	89.5	1.34	0.33
8	50.5	0	25	0	75.5	0.00	0.00
9	43.5	1	22	1	65.5	0.66	0.34
10	31	1	15	0	46	0.67	0.00
11	16	0	8.5	0	24.5	0.00	0.00

– Expected failures)^2/Expected failures, summing up to the postoperative year in question. In comparing 2 life-tables, χ^2 is equal to the sum of the results of this formula for each in the example above. If we compare the 11-year survival from Tables 34 and 35, χ^2 equals 26.07 according to the following calculation: (25.00 [sum of observed failures until year 11 in Table 34] – 16.63 [sum of expected failures until year 11 in Table 34])^2/16.63 [sum of expected failures until year 11 in Table 34] + (18.00 [sum of observed failures until year 11 in Table 35] – 6.05 [sum of expected failures until year 11 in Table 35])^2/6.05 [sum of expected failures until year 11 in Table 35]) = (25.00 – 16.63)^2/16.63 + (18.00 – 6.05)^2/6.05 = 26.07.

This value of χ^2 is higher than 10.828, giving a P value $< .001$. Therefore the difference in the 11-year survival of the implanted shoulder prostheses between these 2 groups of patients is highly significant.

CONCLUSIONS

A method for constructing and comparing survival analyses of orthopaedic procedures by use of the life-table method is presented. The method requires simple arithmetical calculations and can be further simplified by use of basic computer software, such as commonly used spreadsheet software packages. The main issue that should be addressed in this method of survival analyses is a determination of the endpoint criteria for "failures."

Nahum Rosenberg, M.D.
Michael Soudry, M.D.

SECTION 16

Outcome Measures in Multicenter Studies

Small communication errors between different project teams can result in a catastrophic failure: for example, the loss of radio contact between NASA and its Mars Climate Orbiter in 1999 led to a loss of more than US $125 million.[214] The metric/US customary unit mix-up that destroyed the craft was caused by human error in the software development and therefore severe communication problems associated with a lack of control. This nonmedical case exemplifies the need for appropriate harmonization, communication, and subsequent control if more than one group is involved in a complex research project.

Orthopaedic multicenter studies are complex by nature. They are difficult to organize, complex to manage, and hard to analyze. However, there are good reasons to face these challenges:

1. The larger sample size enables testing hypotheses with greater statistical power. It also allows a more precise estimation of population parameters.[215] Especially in low-prevalence disorders, multicenter studies represent the sole option to generate a large enough sample size.

2. The findings and observations of multicenter studies are more generalizable than those of 1 single-center only.[216] The heterogeneity in patient demographics, clinical characteristics, and treatment differences contributes to the variance in study outcome. Even if the treatment is uniformly delivered, it may result in different outcomes at different sites (e.g., European sites compared with Asian sites).
3. The study protocol as a result of a consensus process of experts from different sites is more likely to represent the general opinion in a field and has a better chance for acceptance in the scientific community after the study.[215] This has been recently demonstrated in a large cross-sectional survey of 796 surgeons. The majority of them agreed that they would change their practice based on the results of a large randomized trial.[217]

CHALLENGES IN MULTICENTER STUDIES

The advantages of multicenter studies represent a number of challenges at the same time. The inclusion of more study sites increases the complexity. Slight differences in treatment modalities have to be considered. Working processes that may work locally without extensive infrastructure (e.g., patient monitoring) are not feasible at another site. In most studies differences in infrastructure between various sites require an independent system for data acquisition and processing. The inclusion of several study sites also requires strict monitoring to obtain a defined level of data quality. In summary, it has to be ensured that all sites measure the same variable with the same instrument and the same quality.

Although the inclusion of sites with different cultural background makes the study more representative, this is one of the greatest challenges in multicenter studies. It leads to a number of confounding variables such as socioeconomic environment or different patient expectations. Inclusion of non–English-speaking sites requires cross-cultural adaptation with translation and validation of questionnaires. Differences in cultural background have to be considered during interpretation of data.

Finally, different legal and ethical boundary conditions aggravate study preparation, performance, and analysis. Necessary applications to local ethics committees are becoming more and more complex, time-consuming, and expensive. Different legal restrictions add another challenge in multicenter studies.

The necessary infrastructure and manpower lead to increased costs and time compared with single-center studies. All these challenges have to be considered during planning and performance of multicenter studies to avoid major pitfalls and to produce valuable data.

In summary, there are 2 main challenges related to outcome measures in multicenter trials:

1. Measuring the same data. This means that at 1 site, exactly the same variable is measured as at the other site.
2. Obtaining the same data. This means that varying infrastructure as well as different legal, socioeconomic, and cultural boundary conditions may influence parameters locally, which aggravates further data processing and analysis.

This article should help to identify key components related to outcome measures in multicenter studies. Examples will be used to illustrate possible pitfalls but also strategies to avoid them.

OBJECTIVE OUTCOME MEASURES IN MULTICENTER STUDIES

Although objective outcome parameters are considered as investigator independent, there are a number of factors that may increase variability or introduce sources of unsystematic or systematic errors in multicenter studies. If parameters are measured with different devices., different protocols, or different setups, further data processing may be aggravated.

Range of Motion

Active range of motion and passive range of motion are the most widely used orthopaedic measures in daily clinical practice as well as in clinical studies. Despite their widespread use, there exists a great variability in recording methods. Whereas one group quantified standard errors of measurement between 14° and 25° (interrater trial) and between 11° and 23° (intrarater trial) when comparing 5 methods for assessing shoulder range of motion,[218] other authors concluded in a systematic review that "inter-rater reliability for measurement of passive physiological movements in lower extremity joints is generally low."[219] If objective instruments are used, the interrater reliability of passive physiologic range of motion at the upper extremity can be improved.[220] For example, sufficient intrarater and interrater reliability could be demonstrated when a Fastrak measurement system (Polhemus, Colchester, VT) was used for measuring cervical spine flexion/extension, lateral flexion, and rotation and shoulder flexion/extension, abduction, and external rotation in healthy subjects.[221] However, these systems require handling know-how, are costly, and are often not available at all study sites. Therefore recording of active and passive range of motion with simple methods like the goniometer has to be standardized across study sites. This includes exact definition of measurement planes, starting position, the neutral position for the joint, and the description of the course of movement. The values should be recorded in the neutral-zero method. If only the complete range or a deficit in one plane is reported, information about changes in the neutral position are lacking.[222]

Active and passive range of motion should be recorded with the neutral-zero method. Each movement should be described exactly in the study protocol.

Performance Tests

Physical function and its impairment due to disease activity can be quantified with performance tests. Most tests include the recording of the time needed for a patient to perform the requested activity. In addition, several observational methods have been described that use ratings from observers to assess the quality of physical function.[223] However, Terwee et al.[223] found a number of methodologic shortcomings during a review of the measurement properties of all performance-based methods that have been used to measure the physical function of patients with osteoarthritis of the hip or knee. Most of the tests in this study showed low values for reliability, which represents a challenge for multicenter studies. Impellizzeri and Marcora[224] propose that physiologic and performance tests used in sports science research and professional practice should be developed following a rigorous validation process, as is done in other scientific fields, such as clinimetrics. If performance tests are used in multicenter studies, they have to be described in detail (e.g., with photographic illustrations), should be demonstrated during onsite visits, and should be controlled during study monitoring.

Exact protocols including detailed test descriptions are required for performance tests. Similar testing procedures should be ensured during onsite visits.

Strength Measurements

Muscle strength tests in typical positions belong to the most common clinical outcome parameters. They not only reflect muscle power but also indicate absence of pain, which enables active force generation. Although they are considered "objective," they underlie a number of influencing factors such as fear of injury, pain, medications, work satisfaction, and other motivational factors with an influence on sincerity of active testing.[225] These factors may vary among study sites depending on cultural background, Workers' Compensation, and other socioeconomic factors.

Another challenge is presented by the variety of measurement devices. For example, shoulder abduction strength, as required for the calculation of the Constant score,[226] can be measured with a number of devices, e.g., spring balance, Isobex (Medical Device Solutions AG, Oberburg, Switzerland), or dynamometer. They all operate on a different working principle and subsequently measure different parameters. If not specified before the study, this may lead to a situation in which data pooling is not feasible. Therefore specification of the measurement device is mandatory in each multicenter study. More information about the measurement protocol is necessary, however, to ensure comparability of data. Positioning of the patient may influence the result. This has been shown for grip strength as well as for hip abduction strength. For example, maximal hip abductor strength is significantly higher in the side-lying position compared with the standing and supine positions.[227] In addition, information about the number of repetitions, as well as further data processing, is required to avoid additional bias. Strength measurements are typically repeated in triplicate. Then, it has to be specified whether the maximum, mean, or median value will be processed according to the research question.[228]

For strength measurements, the exact measurement device, including manufacturer, positioning of the patient, number of repetitions, and selection process of measurements have to be defined to ensure data comparability across study sites.

Sophisticated Functional Tests

More sophisticated functional tests such as in-shoe plantar pressure measurements, gait analysis (instrumented walkway), or force-plate analysis may contribute additional information.[229] For specific research questions, these laboratory methods are considered to be the most accurate measurement methods, and clinicians and scientists tend to include them in clinical trials. For instance, high reliability could be shown for various methods of instrumented foot-function tests.[230] However, a number of issues have to be considered to avoid pitfalls when used in multicenter trials: not only does the technology of the chosen test have to be available at each site, but the know-how to operate it is also crucial. For example, a sophisticated motion-capture system requires skilled staff who can install, calibrate, and run it. Laboratory space, logistics for patient handling, computational resources, and experiences in patient testing are necessary. If such a method is to be used in a multicenter study, exact definitions of the system and of all laboratory parameters applicable to all sites are mandatory, as well as careful training. If the specific laboratory test is not feasible at all sites, it is an option to perform the test in a study subgroup only at clinics with the required infrastructure and resources.

> *Sophisticated functional tests may provide additional information for a given research question but require specific infrastructure, know-how, and resources. If not available at all sites, these methods can be limited to a subset of selected sites to collect the additional information.*

Radiographic Evaluation

Radiographic parameters are part of almost all orthopaedic studies. However, despite the widespread use, only little consensus exists about radiographic grading. Interrater agreement measured with the κ coefficient ranges from 0.4 for sclerosis to 0.95 for joint-space narrowing as shown by Lane et al.[231] This broad range was recently re-emphasized in another study investigating the reliability and agreement of measures used in radiographic evaluation of the adult hip.[232] The authors also stated that direct measurements (femoral head diameter) were more reliable than measurements requiring estimation on the part of the observer (Tönnis angle, neck-shaft angle). Agreement between repeated measurements showed many parameters with low absolute reliability. The same problem was reported from the quantification of fracture classification,[233] reduction,[234] and healing.[235,236]

However, the information of an image is stored in the radiograph. Central radiograph reading may help to extract the required data and to avoid subjective judgment by the treating surgeon on the one hand, and it is more reliable in detecting all suspicious findings and less biased by the surgeon's perspective on the other hand. Establishing a radiology review board for a multicenter study is a worthwhile method to increase data quality.[237] The images should be collected centrally, and a minimum of 2 experienced investigators should evaluate the blinded radiographs independently. It is recommended to collect the digital radiographs in DICOM (Digital Imaging and Communications in Medicine) format for later image processing. Clear definitions of each radiologic parameter documented in the study plan or an image-reading manual is mandatory.[238] An initial training session may help to improve interrater agreement.

> *Central image reading by 2 independent, experienced observers and consensus finding help to increase data quality. Strict radiologic definitions are mandatory; an initial training session may help to improve agreement.*

Bone Density Measurements

Local bone density and systemic osteoporosis status both came into focus in several studies.[239,240] A typical example is the change in local bone density around joint replacements as a reaction to different prosthesis designs.[241] Although many authors refer to predefined areas like Gruen zones,[242] they may vary from group to group depending on the exact definition. In a multicenter study, the measurement method (e.g., peripheral quantitative computed tomography or dual-energy absorptiometry), the exact device, and the imaging parameters, as well as the processing algorithm, have to be specified. Especially the differences between different devices for dual-energy absorptiometry introduce a large source of variability in studies with several study sites. These devices are often calibrated with cohorts provided by the manufacturer only. Therefore pooling of absolute values is unfeasible; only relative spatial or temporal changes can be compared or pooled.[243] Limitation to one device type only reduces the number of potential recruitment sites in many studies.

However, if peripheral quantitative computed tomography is feasible within a multicenter study, cross-calibration with a standardized phantom (e.g., the European forearm phantom) improves data quality[244] and allows pooling of the absolute values. Study protocols including bone density assessment should include documentation of precision accuracy and stability at one site as well as comparisons between different sites. A protocol for the circulation and testing of a calibration phantom helps to ensure the required data quality.

> *Quantification of local and systemic bone density has to be defined in detail including measurement site and area, measurement device, imaging protocol, and (cross-)calibration.*

PATIENT-REPORTED OUTCOMES IN MULTICENTER STUDIES

Patient-reported outcomes (PROs) are subjective parameters that come directly from the patient. In contrast to objective parameters, they should exclusively reflect the patient's health condition (e.g., function of the knee, ability to walk, pain, and HRQL) without any space for interpretation by a clinician. They can be used to obtain information on the actual status of a sign or symptom of the patient (e.g., on the preoperative status of an arthritic joint) or to see changes of a sign or symptom over the time—for example, to assess the effect of a medical treatment or the success of a surgery.

Choosing the Conceptual Framework

The 4 target domains that contribute to functional outcomes can be viewed as physical, mental, emotional, and social in nature.[245] In treating patients with impingement, for example, there is a need to facilitate clinical decisions where surgeons must weigh, either explicitly or implicitly, the expected benefits of a particular intervention, whether surgical, medical, or rehabilitative, against the potential harm and cost.[246,247] The choice of an appropriate disability conceptual framework to classify different domains and instruments is fundamental because there is a lack of consistent language and uniform definitions when defining physical function. However, without a common metric to measure these targets, we would be unable to compare results across trials and guide clinical decision making.

The main purpose of the International Classification of Functioning, Disability and Health (ICF) of the World Health Organization to provide a common language to describe disability concepts has made the framework widely popular.[248] Functioning and disability are described in the ICF in terms of the dynamic interaction between health condition, personal factors, and the environment. The ICF is not only a classification system for the impact of disease, it is also a theoretical framework for a relation between variables. The ICF places the emphasis on function rather than condition or disease. The ICF provides a description of situations with regard to human functioning and its restriction. The information is organized into 2 parts: part 1 deals with functioning and disability, whereas part 2 covers contextual factors. Each part has 2 components: The body component comprises 2 classifications, 1 for functions of body systems and 1 for body structures. Activities may be limited in nature, duration, and quality.[249] Activity limitations are referred to as disabilities and are scaled by difficulties and whether assistance is needed to carry out the activity. The ICF has been identified by the American National Committee on Vital and Health Statistics as the only viable code set for reporting functional status.[250]

The design and conduct of good comparative studies in this context rely on the choice of valid instruments that are reliable and responsive.[251] Should the instrument assessing functional outcomes prove to have good psychometric properties, the value of the published literature would be enhanced.[252] However, pragmatic qualities such as the applicability of such instruments in trials examining specific populations, for instance, femoroacetabular impingement and hip labral pathology, should also be considered in addition to the psychometric properties. For example, logistical choices for use of functional outcome instruments should take into consideration the burden to administer, require additional training, and have an adequate score distribution as well as format compatibility.[253] To obtain comparable results, it is necessary that all participating centers use the same version of an outcome measure and perform it in the same way (e.g., direct distribution or telephone interview). This is especially important for those instruments where different versions exist, e.g., the HRQL instrument SF-36 (version 1 or 2, 1 week's recall or 4 weeks' recall) or the Constant score at the shoulder.

> - *Use a framework to classify health concepts, whether impairment or activity participation.*

- *Use both disease-specific and generic health measures.*
- *Use instruments with tested psychometric properties.*

Cross-Cultural Challenges

The cultural background can be an important confounding factor in international multicenter studies and also in national studies including migration populations of different cultures. For example, illness behavior and perceptions of pain are different between Americans and Asians.[254]

Have you ever thought about how to assess the same item in patients from different countries, with different cultural backgrounds and different functional demands?

For example, East Asian people use different functions of the hand when eating with chopsticks than Western people. In many cultures, kneeling is an important function regularly practiced during eating or praying, with highest functional demands because of the maximum flexion of the knee.

When using a PRO, it is important that it is available in the national language of the target population because it should be answered by the patient in context with his or her cultural background. Availability in another language does not mean that it has simply been translated by one interpreter or even by a doctor during the interview with the patient. An instrument that should allow reliable comparisons with other studies (e.g., comparing treatment effects) or will be used in an international multicenter study should undergo a careful methodologic process of cross-cultural adaptation and validation such as or comparable to the process described by Guillemin et al.[255] and Beaton et al.[256] (Fig 12). The questionnaire must be correctly translated not only for all questions and answers but also for all instructions for the patient and for the scoring method. For all steps of such a process, careful written documentation that highlights difficulties to reach equivalence between the original questionnaire and new-language questionnaire is necessary.

The first step is the translation into the target language. This should be done independently by 2 bilingual translators with the target language as their mother tongue. One of the translators should be aware of the concept of the questionnaire and should have a medical background. The second translator should have no medical background and be uninformed regarding the concept of the questionnaire. Both translators produce 2 forward-translations, versions T1 and T2.

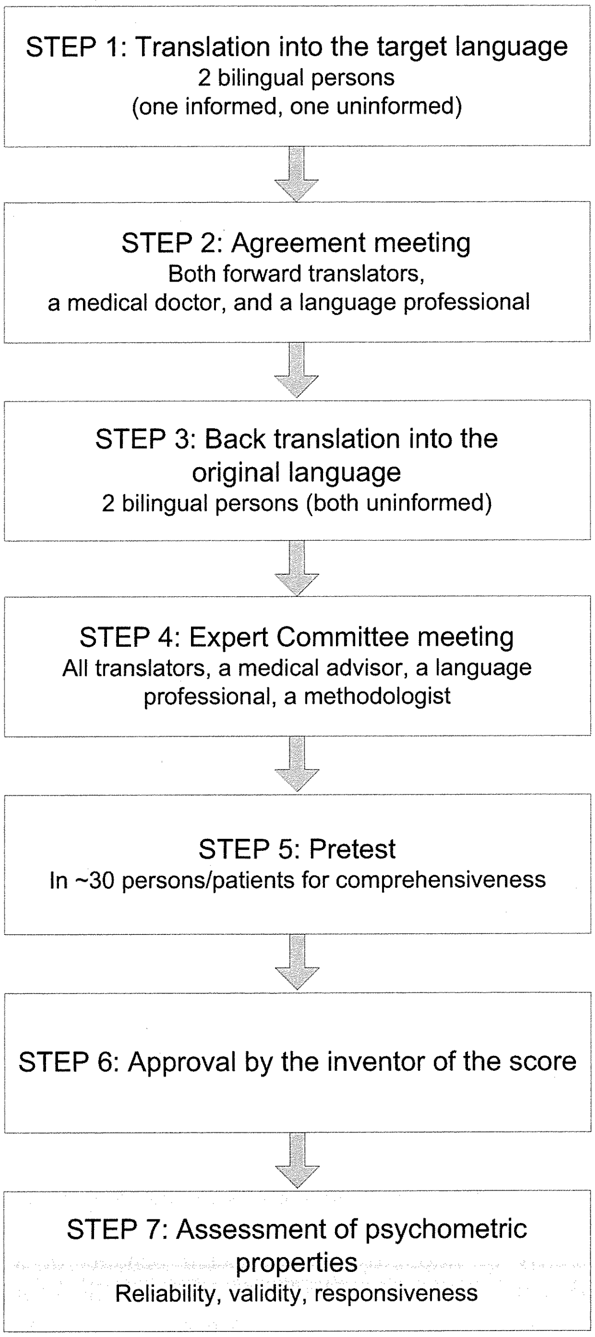
STEP 1: Translation into the target language
2 bilingual persons
(one informed, one uninformed)
STEP 2: Agreement meeting
Both forward translators,
a medical doctor, and a language professional
STEP 3: Back translation into the original language
2 bilingual persons (both uninformed)
STEP 4: Expert Committee meeting
All translators, a medical advisor, a language professional, a methodologist
STEP 5: Pretest
In ~30 persons/patients for comprehensiveness
STEP 6: Approval by the inventor of the score
STEP 7: Assessment of psychometric properties
Reliability, validity, responsiveness

FIGURE 12. Steps of cross-cultural adaptation. Adapted from Beaton et al.[256]

The second step is an agreement meeting, where both forward-translators find an agreement on the translations and produce a synthesis version (T12). The discussion should be led by a third person acting as mediator, e.g., a medical doctor familiar with the questionnaire and its concept. A language professional