

# Statistical Analysis of Quantitative MRI Data: Basic Methods

Professor A. John Petkau  
Department of Statistics  
University of British Columbia  
Vancouver, British Columbia, Canada

## 1. Introduction

We discuss basic concepts and methods for the statistical analysis of quantitative MRI data, with a focus on the types of MRI data most commonly utilized as outcome measures for clinical trials in multiple sclerosis (MS). The discussion is in the context of a randomized controlled clinical trial with a placebo (*Pl*) and an active treatment (*Rx*) arm. Much more detailed presentations of these basic statistical concepts and methods are available in excellent introductory statistical intended primarily for medical researchers [1-4].

## 2. Quantitative MRI Data in MS Clinical Trials

MS is a chronic disease that affects the central nervous system. During an MS exacerbation inflammation of the myelin sheath disrupts messages passed along the motor and sensory nerves. Areas of demyelination may result reflected by scar tissue or lesions. MRI imaging allows visualization of these lesions, leading to the quantitative MRI data most commonly used as outcome measures for clinical trials in MS: the counts and volumes of these lesions.

MRI imaging has been included in most clinical trials in MS since the pivotal trial establishing the benefit of interferon beta-1b in relapsing-remitting MS [5, 6]. Phase III trials are designed with a clinical endpoint but it has become customary to obtain MRI images from all patients at screening, at baseline and at least annually during the trial. A “frequent MRI imaging” cohort is sometimes also included consisting of patients with MRI images obtained on a more frequent basis, typically monthly, although perhaps only for the first six months or year of follow-up. In contrast, Phase II trials are often designed with a MRI endpoint based on frequent MRI imaging as the primary outcome and with much shorter follow-up, often no more than one year.

Both  $T_1$  and  $T_2$ -weighted scans are usually obtained. The  $T_1$ -weighted scan is used with an injection of gadolinium, a contrast agent. Lesions appear on the  $T_1$  image as bright areas (enhancements) where gadolinium has been able to cross the blood/brain barrier and on the  $T_2$  image as white areas. The volume of  $T_2$  lesions, the *lesion burden*, indicates the volume of brain tissue affected by the MS disease process. Comparison with previous images allows identification of *new*, *recurrent* and *persistent* lesions on the  $T_1$  image and *new*, *recurrent* and *enlarging* lesions on the  $T_2$  image. The different lesion counts are often combined into an overall count of unique lesions that are active, the *combined unique activity* count.

These MRI outcomes are available at each scheduled MRI time point during follow-up. Typically, this longitudinal MRI data for each patient is summarized over time for statistical analysis. The most common summary of the repeated lesion burden readings is the change in lesion burden from baseline to the end of study. For the lesion count outcomes, the most common summary is the accumulated count over all the scans obtained for this patient during follow-up. A simpler overall summary is the proportion of active scans for each patient. Hence, statistical methods for count and continuous responses are of greatest relevance.

### 3. Basic Statistical Concepts

The objective of statistical science is to convert data into information. To accomplish this, statistical scientists develop and apply efficient methods for collecting data (study design), for summarizing and presenting data (descriptive statistics), and for drawing conclusions from data (inferential statistics). Clinical trials are carried out to assess the efficacy of therapies, so the discussion here will focus on aspects of inferential statistics.

All techniques of statistical inference have the same objective: to use data collected on a sample of patients to draw conclusions about the population from which those patients were sampled. The responses of sample patients convey information about what the responses of the population of patients would be if one measured them. The extreme importance of appropriate study design to guarantee that the sample genuinely represents the population is self-evident.

The eligibility criteria for a clinical trial identify the target population of patients about which conclusions are desired. Randomization ensures that the assignment of patients to the *Pl* and *Rx* arms is not subject to bias. This generates random samples of patients from the two conceptual populations of interest: the *treatment population* consisting of the target population with patients subjected to *Rx* and the *placebo population* consisting of the target population with patients subjected to *Pl*. Comparison of the responses of the patients in the two random samples then allows conclusions to be drawn concerning the differences that would be observed in these two conceptual populations (e.g., in the mean response) if responses were obtained from all the patients in the target population.

### 4. Basic Statistical Methods for Continuous and Count Data

The basic statistical methods appropriate for continuous responses such as the change in lesion burden can also be used for count responses such as the accumulated combined unique activity. Suppose then that a randomization scheme has resulted in simple random samples of  $n_{Rx}$  and  $n_{Pl}$  patients on the *Rx* and *Pl* arms of a clinical trial and an investigator wishes to use a continuous or count outcome measure to assess whether *Rx* has an effect differing from that of *Pl*.

A *histogram* or *boxplot* provides an immediate visual impression of a data set's key features:

- the *location* of the center of the data set,
- the degree of *symmetry* exhibited by the data set,
- the amount of *spread* in the data set,
- the presence of any *outlying values* in the data set.

The location of the center and the amount of spread in a data set is usually described by the sample average (*ave*) and standard deviation (*SD*). When serious departures from symmetry or extreme outliers are present, alternative summaries such as the median for the center and the interquartile range for the spread might better describe these features.

For sets of data without serious departures from symmetry or extreme outliers, the *ave* and *SD* provide a rough but informative description of the data set via the *empirical rule*:

- a majority of the data values (about 68%) will lie within one *SD* of *ave*,
- most of the data values (about 95%) will lie within two *SDs* of *ave*,
- essentially all of the data values (about 99.7%) will lie within three *SDs* of *ave*.

These exact percentages apply only if the histogram is well approximated by a *normal curve*, but this empirical rule provides a useful qualitative description for most data sets.

## Measurement Error

As for any measurement process, MRI imaging is subject to multiple sources of error. At any point in time, a patient has a certain value of lesion burden, call it  $X$ . From a scan of this patient, the value obtained for the patient's lesion burden is not  $X$  itself, but rather  $X$  contaminated by these sources of error. This leads to a measured lesion burden  $Y$  that can be represented as

$$Y = X + \varepsilon_X,$$

where  $\varepsilon_X$  is the error made in measuring  $X$ , often referred to as the *noise* in the measurement process. Although impossible in practice, it is useful to imagine measuring this patient's lesion burden repeatedly under exactly the same conditions at exactly the same instant of time; each repeated measurement involves a new MRI scan and a new analysis of the resulting image to obtain a new measured value of the patient's lesion burden. Each repetition yields its own value of  $\varepsilon_X$ , so the resulting measured values are all different.

The histogram of the  $\varepsilon_X$ 's describes the properties of the measurements of this patient's lesion burden  $X$ . The *ave* of this histogram describes the *bias* in the measurement process; if *ave* is 0, the process is said to be *accurate* as no systematic errors are being made in measuring this patient's lesion burden. The *precision* of the measurement process is described by the amount of spread in the histogram: the smaller the *SD*, the more precise. Measurement processes should be not only accurate, but also highly precise.

In the clinical trials context, lesion burden is measured on many patients (on each arm) at each visit. If the measurement process is *homogeneous* so that the properties of the  $\varepsilon_X$ 's do not vary with the value of  $X$  being measured, then the *SD* of the measured  $Y$  values is given by

$$SD(Y) = SD(X)\sqrt{1+R^2},$$

where  $R = SD(\varepsilon)/SD(X)$ . This describes how lack of precision in the measurement process is reflected in increased variability of the measured lesion burden values. Provided the *SD* of the measurement errors is small relative to the *SD* of the lesion burdens across patients in the target population, the resulting increase in variability of the measured values will be negligible.

## Point Estimation and Confidence Intervals

The investigator's objective is to use the data to draw conclusions about  $Rx$ . She is interested in  $ave_{Rx}$  and  $SD_{Rx}$  because these summary statistics provide information about the responses that would be observed if this outcome measure was collected on the entire treatment population, the conceptual population consisting of the target population with all patients subjected to  $Rx$ .

Suppose the investigator wishes to describe the mean response in the treatment population; this unknown *population parameter* is denoted by  $\mu_{Rx}$ . Her best guess, or *sample estimate*, of this population mean response is  $ave_{Rx}$ , the average response in the random sample of  $n_{Rx}$  patients.

The investigator needs to describe how precisely determined  $ave_{Rx}$  is as an estimate of  $\mu_{Rx}$ ; this information is provided by the *standard error (SE)*. For the average of a simple random sample,

$$SE(ave) = \sigma/\sqrt{n},$$

where  $\sigma$  is the standard deviation in the population.  $SE(ave)$  is directly proportional to  $\sigma$  and decreases as the sample size  $n$  increases. More precisely,  $SE(ave)$  varies inversely as  $\sqrt{n}$  so, for example, doubling the precision requires four times as large a sample.

The *SE* describes the variability of an estimate in repeated sampling. If one imagines drawing a large number of simple random samples of size  $n$  from the same target population, then the collection of values of *ave* in these simple random samples can be described as follows:

- a majority of the values (about 68%) will lie within one *SE* of the population mean  $\mu$ ,
- most of the values (about 95%) will lie within two *SEs* of  $\mu$ ,
- essentially all of the values (about 99.7%) will lie within three *SEs* of  $\mu$ .

A more detailed description is that the histogram of this collection of values of *ave* can be approximated as a normal curve, with a mean of  $\mu$  and a standard deviation of  $SE(ave)$ .

The *Central Limit Theorem* establishes that this description of the sampling distribution of the sample average will apply irrespective of the form of the histogram of the data in the target population, provided that the sample size  $n$  is reasonably large. The only limitation to the application of this result is precise interpretation of what is required for the sample size  $n$  to be “reasonably large”. If the histogram of the data in the target population is highly skewed (as is sometimes the case for both change in lesion burden and accumulated combined unique activity data) or otherwise poorly behaved, then a larger sample size is required before this description will provide a good approximation. But, for many of the outcome measures collected in clinical trials, this description will provide an adequate approximation even for quite modest values of  $n$ .

In the context of a clinical trial, the population of interest is a conceptual population, so the population standard deviation  $\sigma_{Rx}$  is unknown and must be estimated by  $SD_{Rx}$ , the standard deviation in the sample. This leads to the *estimated standard error* of the estimate  $ave_{Rx}$  as

$$\hat{SE}(ave_{Rx}) = SD_{Rx} / \sqrt{n_{Rx}} .$$

The investigator can now identify a range of plausible values for  $\mu_{Rx}$ , the unknown mean response in the treatment population, by constructing a confidence interval (CI). The approximate  $1 - \alpha$  CI for  $\mu_{Rx}$  is given by

$$ave_{Rx} \pm z_{\alpha/2} \hat{SE}(ave_{Rx}) ,$$

where the cut-off value  $z_{\alpha/2}$  is such that the area to the right under the *standard normal curve* is equal to  $\alpha / 2$ . For the common choices of 90%, 95% and 99% levels of confidence, these cut-off values are 1.65, 1.96 and 2.58 respectively. The width of this approximate CI increases as the variability increases, as the sample size decreases, and as the level of confidence increases.

But the investigator’s primary objective is to draw conclusions concerning differences between the treatment and placebo populations to identify the benefit future patients would derive from *Rx* over and above any benefit derived from *Pl*. If a beneficial effect of *Rx* corresponds to lowering the mean response (as for change in lesion burden and accumulated combined unique activity data), then the parameter of interest might be  $\mu_{Pl} - \mu_{Rx}$ , the difference in population means that can be attributed to the effect of *Rx*. This unknown population parameter would be estimated by  $ave_{Pl} - ave_{Rx}$ , and an approximate  $1 - \alpha$  CI for  $\mu_{Pl} - \mu_{Rx}$  would be given by:

$$ave_{Pl} - ave_{Rx} \pm z_{\alpha/2} \hat{SE}(ave_{Pl} - ave_{Rx}) .$$

For independent simple random samples from two distinct populations, as in a randomized controlled clinical trial, the *SE* for the difference of the sample averages is given by:

$$\hat{SE}(ave_{Pl} - ave_{Rx}) = \sqrt{\hat{SE}(ave_{Pl})^2 + \hat{SE}(ave_{Rx})^2} .$$

This CI provides a range of plausible values for  $\mu_{PI} - \mu_{Rx}$ , the unknown population parameter of interest, thereby indicating the possible magnitude of the treatment effect. In particular, if the calculated CI contains the value 0, then at the  $1 - \alpha$  level of confidence, the data fail to provide a clear indication of an effect due to  $Rx$  (relative to  $PI$ ).

### Testing Hypotheses

The usual approach taken to address the key question of whether the data provide convincing evidence of an effect due to  $Rx$  is based on hypothesis testing. When the parameter of interest is  $\mu_{PI} - \mu_{Rx}$ , the investigator carries out a *test of significance* of the null hypothesis,  $H_0: \mu_{PI} - \mu_{Rx} = 0$ , or equivalently,  $H_0: \mu_{PI} = \mu_{Rx}$ . To do so, the investigator calculates the  $Z$ -statistic,

$$Z = (ave_{PI} - ave_{Rx}) / \hat{SE}(ave_{PI} - ave_{Rx}).$$

This re-expresses the estimate of the parameter of interest,  $ave_{PI} - ave_{Rx}$ , in terms of its number of estimated  $SE$ 's away from 0, the value of the parameter of interest under  $H_0$ .

The investigator then evaluates the probability, under the assumption  $H_0$  is true, of observing a value of this  $Z$ -statistic as extreme or more extreme than the value observe; this is the *P-value* corresponding to the observed value of this  $Z$ -statistic. In the clinical trials context, two-sided alternatives to the  $H_0$  are usually contemplated (a beneficial effect due to  $Rx$  is anticipated but the possibility that  $Rx$  may be detrimental should not be overlooked), so both negative and positive values of the  $Z$ -statistic as extreme or more extreme than the value observed must be considered when calculating the  $P$ -value.

Based on the Central Limit Theorem, an approximate  $P$ -value can be evaluated by comparing the observed value of this  $Z$ -statistic to the standard normal curve. Small  $P$ -values correspond to extreme values of the  $Z$ -statistic; for example, values of  $Z = \pm 1, \pm 2$  and  $\pm 3$  lead to approximate two-sided  $P$ -values of 0.32, 0.05 and 0.003 respectively. There are two possible explanations for an observed extreme value of the  $Z$ -statistic: either  $H_0$  is true but the samples led to an extreme value of  $ave_{PI} - ave_{Rx}$ , an event of low probability, or  $H_0$  is false, in which case a value of  $ave_{PI} - ave_{Rx}$  quite different from 0 is not surprising. Events of low probability are not expected to occur, so small  $P$ -values are interpreted as evidence that  $H_0$  is false; the smaller the  $P$ -value, the more convincing the evidence is considered.

The strength of evidence represented by a  $P$ -value is on a continuum of possible values between 0 and 1, so sharp dividing points such as the popular 0.05 are totally arbitrary and have no meaning;  $P$ -values of 0.051 and 0.049 represent evidence of essentially identical strength. There can be no absolute standard for how small a  $P$ -value should be considered a definitive refutation of  $H_0$ . For any context, this depends on the consequences of the two possible errors:

- *Type 1 error*: rejecting  $H_0$  when it is true (false positive finding),
- *Type 2 error*: failing to reject  $H_0$  when it is false (false negative finding).

Reports of experimental results should always present the actual  $P$ -values so readers can decide for themselves if they find the evidence against  $H_0$  convincing.

The logic of a test of significance deserves careful attention. The approach is directed towards rejecting  $H_0$  to substantiate a claim that a difference exists. The  $P$ -value provides an indication of the strength of evidence against  $H_0$ . A test of significance can result in a definitive conclusion that  $H_0$  is false, but failure to reject  $H_0$  indicates only an absence of evidence of a difference and provides no indication how large a difference might still exist (a CI is better suited to this task). If the study was not designed with adequate sensitivity to detect clinically

meaningful differences, for example, then it is quite unlikely to lead to a rejection of  $H_0$ . Thus, failure to reject  $H_0$  should not be interpreted as establishing that  $H_0$  is true.

### Sample Size Calculations

How many patients should be randomized in a clinical trial designed to assess the effect of  $R_x$  relative to  $Pl$ ? This question is usually addressed by specifying desired levels for the chances of making Type 1 and Type 2 errors with a test of significance of the relevant null hypothesis. If the investigator will measure the magnitude of the treatment effect by  $\mu_{Pl} - \mu_{R_x}$ , the difference in mean response in the placebo and treatment populations, this becomes  $H_0: \mu_{Pl} = \mu_{R_x}$ .

Formally, it is assumed  $H_0$  will be rejected if the  $P$ -value from the test of significance is less than some pre-specified value, called the *level of significance* of the test and denoted by  $\alpha$ ; the chance of making a Type 1 error is then equal to  $\alpha$ . The common choice of  $\alpha = 0.05$  means there is a 1 in 20 chance of deciding there is a difference in the mean responses of the placebo and treatment populations (rejecting  $H_0$ ) when no such difference exists.

Specification of the desired chances of Type 2 error involves two choices. First, the investigator must identify a meaningful difference in the values of  $\mu_{Pl}$  and  $\mu_{R_x}$ , say  $\mu_{Pl} - \mu_{R_x} = \Delta$ . In addition, she must specify the desired level for the chance of making a Type 2 error when  $\mu_{Pl} - \mu_{R_x} = \Delta$  (failing to reject  $H_0$  when the mean response on the placebo population exceeds that on the treatment population by  $\Delta$ ). The chance of making a Type 2 error is denoted by  $\beta$  and  $1 - \beta$  is called the *power* of the test (at the alternative  $\mu_{Pl} - \mu_{R_x} = \Delta$ ). Common choices for  $1 - \beta$  are 0.80, 0.90 and 0.95. These correspond to chances of 1 in 5, 1 in 10 and 1 in 20 of failing to detect the specified meaningful difference  $\Delta$  in the values of  $\mu_{Pl}$  and  $\mu_{R_x}$ , so a clinical trial designed to have a power of 0.90 or 0.95 is much more desirable than one designed for a power of 0.80.

With a specified level of significance, increasing the sample size increases the power of a test (equivalently, decreases the chances of a Type 2 error). For a two-sided test of significance based on the  $Z$ -statistic described above, the sample size per arm required to achieve the desired levels of both the Type 1 and Type 2 errors is

$$n \geq (z_{\alpha/2} + z_{\beta})^2 (\sigma_{Pl}^2 + \sigma_{R_x}^2) / \Delta^2,$$

where  $\Delta$  is the value of  $\mu_{Pl} - \mu_{R_x}$  that is to be detected with the specified power of  $1 - \beta$ .

The first term in this expression for the required sample size involves only the specified levels of the chances of Type 1 and Type 2 error. For the common choice of  $\alpha = 0.05$ ,  $z_{\alpha/2} = 1.96$ , while  $z_{\beta} = 0.84, 1.28$  and  $1.65$  for  $1 - \beta = 0.80, 0.90$  and  $0.95$ , leading to

$$(z_{\alpha/2} + z_{\beta})^2 \approx 7.8, 10.5 \text{ and } 13.0 \text{ respectively.}$$

Hence, increasing the power of a level  $\alpha = 0.05$  test from 0.80 to 0.90 requires an increase in the sample size of roughly one-third, while increasing the power from 0.80 to 0.95 requires an increase in the sample size of roughly two-thirds. These are substantial increases in the required sample size but each one-third increase results in halving the chances of a Type 2 error, so these higher values deserve serious consideration as choices for the power.

The expression for the required sample size makes it clear that  $\Delta$ , the difference to be detected in the mean responses of the placebo and treatment populations, plays a critical role. In particular, if a difference one-half as large is to be detected, then the required sample size is four times as

large. Similarly, a context in which the standard deviations in both the placebo and treatment populations are twice as large will require a sample size four times as large.

The main difficulty in determining required sample sizes is usually lack of adequate knowledge about the population standard deviations  $\sigma_{Pl}$  and  $\sigma_{Rx}$ , particularly  $\sigma_{Rx}$ . Natural history data on patients eligible for the planned clinical trial provides an indication of the mean response and standard deviation to be anticipated on the *Pl* arm. Based on knowledge of the size of the treatment effect that would be clinically important, the investigator should be able to decide on the magnitude of the difference in the mean responses to be detected. But often there is little information available on how variable the responses of patients on the *Rx* arm are likely to be. This considerable uncertainty implies there can be no single answer to the sample size question. Rather, required sample sizes should be evaluated for ranges of plausible values of all uncertain input parameters. Careful judgment needs to be exercised to come to a final decision on an appropriate sample size for a contemplated clinical trial.

### Qualifying Comments

The methods discussed above are approximate: their validity relies on the approximation of the sampling distribution of the sample average by a normal curve as justified by the Central Limit Theorem. The quality of this approximation depends on the form of the histogram of the responses in the populations from which the samples are drawn and on the size of the samples. For any sample size, the better behaved the histogram of the population (closer to a normal curve) the more accurate will be the approximation. The sample sizes employed in clinical trials in MS ensure such approximations are adequate for most inferences desired in such contexts.

There are some exceptions, however. Accurate approximations cannot be expected for very small *P*-values. But the appropriate inference is clear with a very small *P*-value, so its exact value is not so critical. This explains the usual practice of citing any *P*-value which turn out to be less than 0.001 (say) when calculated using such an approximation, as simply  $P < 0.001$ .

Exceptions can also occur in subgroup analyses. For example, the frequent MRI cohort of the pivotal Phase III clinical trial of interferon beta-1b consisted of all the patients randomized at a single center [6]. These 52 patients can also be viewed as representing random samples from conceptual populations, but the samples are rather small so there could be some doubt about the quality of the approximations the Central Limit Theorem would provide in this case.

Two strategies are available for dealing with such cases. The first is to transform the data to make the histogram better behaved. The methods described are then applied to the transformed data. The main limitation of this strategy is that the choice of transformation can seem arbitrary. In fact, theoretical results and experience make this a very effective strategy. For example, transforming count responses by taking square roots and time-to-event data by taking logs often works well. Both transformations greatly reduce large values in the data set and therefore can be effective at reducing positive skewness. But neither deals effectively with extreme outliers.

The second strategy involves replacing the data by their ranks and then using a rank-based nonparametric procedure, such as the *Mann-Whitney-Wilcoxon statistic* to carry out the desired inference. This greatly reduces the impact of any outliers in the data set and thus leads to more robust inferences. A limitation of this strategy is that nonparametric procedures tend to be focused on testing hypotheses and may not readily lend themselves to estimating the magnitude of a treatment effect. Also, such procedures discard much of the information on magnitude in

the data and therefore, in circumstances where parametric procedures are appropriate, require larger sample sizes to achieve the same inferential goal.

To conclude this subsection, special circumstances are described in which slight variations of these procedures should be used. Suppose it is known that the standard deviations in the placebo and treatment populations are the same,  $\sigma_{pl} = \sigma_{Rx} = \sigma$  say. Although this would be a rare circumstance in the clinical trials context, then  $SD_{pl}$  and  $SD_{Rx}$  both estimate the common population standard deviation  $\sigma$ , so a better estimate is obtained by combining these two estimates. Statistical theory establishes that this should be done by forming  $SD_{pooled}$ , where

$$SD_{pooled} = \sqrt{[(n_{pl} - 1)SD_{pl}^2 + (n_{Rx} - 1)SD_{Rx}^2] / (n_{pl} + n_{Rx} - 2)} ;$$

this estimate is said to be based on  $n_{pl} + n_{Rx} - 2$  degrees of freedom. This pooled estimate is then used to estimate the values of  $SE(ave_{pl})$  and  $SE(ave_{Rx})$ , leading to

$$\hat{SE}(ave_{pl} - ave_{Rx}) = SD_{pooled} \sqrt{1/n_{pl} + 1/n_{Rx}} .$$

The approximate procedures described above then apply as before.

Suppose that, in addition, the histograms of the placebo and treatment populations are known to be well approximated by normal curves. Using *Student's t curve* (with  $n_{pl} + n_{Rx} - 2$  degrees of freedom) to obtain the cut-off values for forming CI's and calculating *P-values* then yields exact inferences. For a 95% level of confidence, the Student's *t* cut-off values are 2.23, 2.09, 2.01 and 1.98 for 10, 20, 50 and 100 degrees of freedom. Each is slightly larger than the corresponding standard normal cut-off value of 1.96, so use of Student's *t* results in slightly wider CI's and larger *P-values*. But these differences are substantial only for small values of the degrees of freedom; that is, only for a small total sample size  $n_{pl} + n_{Rx}$ . Except perhaps for subgroup analyses, this variation has little impact for clinical trials in MS.

In practice, these procedures based on the use of  $SD_{pooled}$  and Student's *t* cut-off values are often used even when the assumptions required to justify their use cannot be strictly verified. This works very well provided the sample sizes on the two arms are roughly equal – as would typically be the case in the clinical trials context.

### More General Contexts

The discussion has focused on the special case where the parameter of interest is  $\mu_{pl} - \mu_{Rx}$ , the difference in the mean responses in the placebo and treatment populations. But these methods can be applied in a great variety of problems arising in clinical trials comparing two arms. In general, there is a parameter of interest,  $\theta$  say, that forms the basis of the comparison of the two corresponding conceptual populations. From the data collected in the clinical trial, an estimate  $\hat{\theta}$  of the parameter of interest would be available, together with  $\hat{SE}(\hat{\theta})$ , an estimated standard error for that estimate. An approximate  $1 - \alpha$  CI for the parameter of interest is then

$$\hat{\theta} \pm z_{\alpha/2} \hat{SE}(\hat{\theta}) .$$

Similarly, the *Z*-statistic for the test of significance of the null hypothesis,  $H_0: \theta = 0$ , is

$$Z = \hat{\theta} / \hat{SE}(\hat{\theta}) ,$$

thereby allowing the evaluation of an approximate *P-value*.

It is usually clear how  $\hat{\theta}$  should be calculated. The only technical obstacle is determination of the appropriate formula for  $\hat{SE}(\hat{\theta})$ . Implementation of these inferential techniques is then straightforward. But the necessary details of carrying out sample size calculations depend upon the choice of the parameter of interest and are often not so straightforward. Handbooks provide guidance for the most common situations arising in the clinical trials context [7].

For the designs usually employed in clinical trials, expressions for  $\hat{SE}(\hat{\theta})$  for most estimates of interest can be found in statistical texts [1-4]. For some estimates, available formulae are themselves approximations, typically relying on the availability of large samples, but in some instances depending in addition upon assumptions concerning the histograms of the responses for the target population. The *bootstrap method* is an alternative computer-intensive method that can be used routinely to estimate  $\hat{SE}(\hat{\theta})$  for most estimates [8, 9].

The fundamental requirement for the use of these approximate methods in such more general contexts is that the sampling distribution of  $\hat{\theta}$  should be reasonably well approximated by a normal curve. Provided both sample sizes,  $n_{Pl}$  and  $n_{Rx}$ , are reasonably large this would almost always be the case although the advice of a statistician should be sought to insure that technical issues concerned with how best to make such approximations are carefully addressed.

## **5. Analyses Incorporating Covariates**

To this point, discussion has been limited to the simple situation where the only characteristics of the patients taken into account are the arms to which they are randomized. Such unadjusted comparisons may be required for certain purposes (e.g., regulatory approval), but incorporating covariates of potential importance is essential for an efficient and comprehensive analysis.

For continuous outcome measures like the change in lesion burden such analyses can be implemented with *ordinary regression analysis*. This approach might also be used for lesion count responses provided the responses are relatively large for at least most of the patients, although this would often be more suitable after some transformation (e.g., square root) of the count responses. *Poisson regression analysis* provides a more generally applicable approach, particularly in the general form allowing for *overdispersion* in the count responses. *Logistic regression analysis* provides corresponding methodology for binary responses.

Such covariance analyses have multiple purposes. One is to induce closer equivalence between the *Rx* and *Pl* arms. Despite randomization, some degree of imbalance on covariates will always exist and a covariance analysis can account for such imbalances. Incorporating into the analysis covariates identified as important in advance is to be preferred to the common approach of carrying out tests of homogeneity across the arms on each covariate and ignoring those where no difference is detected. In particular, any covariates used in stratification of the randomization scheme should be incorporated into the analysis of the resulting data.

The use of covariance analysis to carry out interaction tests examining the consistency of treatment effects across subgroups of patients (e.g., subgroups defined by gender) provides a more coherent approach than the common practice of comparing the *Rx* and *Pl* arms within such subgroups. Such subgroup analyses should be undertaken only if the corresponding interaction test indicates an effect and, in any case, should be viewed as exploratory.

In the context of clinical trials, covariance analysis is frequently used to clarify the extent to which detected differences are due to *Rx* rather than other factors associated with response. But, covariance analysis also has the potential to provide a more powerful comparison of the *Rx* and *Pl* arms through the reduction of variability that results when covariates highly associated with the response are taken into account. Thus, for example, adjustment for the baseline combined unique activity count (if available) should be expected to lead to a more sensitive analysis of the accumulated combined activity data. Covariance analyses deserve to be much more fully utilized in the analysis of data collected in MS clinical trials.

## **6. Longitudinal Analyses**

The discussion to this point has not utilized the longitudinal nature of the data collected in the clinical trials context. We have assumed that a single value, suitably summarizing the data collected over time for each patient, is utilized as the response. But this prevents any critical examination of the patterns over time that may contain important information concerning the action of *Rx*. This is equivalent to throwing away some of the information collected in the trial. Further, the longitudinal data could allow improved estimation of change over time and thereby a more sensitive assessment of the differences between the arms.

The wish to make better use of the longitudinal data is reflected in the common approach of carrying out separate analyses of the data summarized from baseline to each time point of interest. For example, if only annual T2 MRI scans were collected in a trial with two years of follow-up, the change in lesion burden from baseline to Year 1, from baseline to Year 2 and from Year 1 to Year 2 might be analyzed separately. This simple approach provides only an indirect assessment of the patterns over time but is reasonable if there are only a few time points of interest. On the other hand, serious difficulties of *multiple testing* are encountered when there are many time points of interest as in a frequent MRI setting. What is needed are methods that allow simultaneous analyses of the data at all the individual time points.

Classical statistical methodology for longitudinal data, such as *repeated measures analysis* and *growth curves analysis*, is not well suited to the clinical trials context. Both require a rigid schedule of data collection and strong assumptions on the variance and correlation structure in the data. Further, neither can easily handle the missing data that inevitably arise in the context of clinical trials. More flexible methodology designed specifically for such analyses is now readily available and deserves to be much more widely used to provide more comprehensive analyses of the data collected in clinical trials and other longitudinal studies.

One general approach, *generalized estimating equations* (GEE), is based on extensions of the equations used for regression analysis [10, 11]. The focus in this approach is on the relationship of the outcome measure for the patients on each arm, as a group, to covariates (including group membership and time on study), so it involves *population-average modeling*. The approach is *semiparametric* in that it does not rely on strong assumptions about the joint distribution of the repeated responses on individual patients. This limits the types of inferences that can be made (e.g., it is not suited to making predictions for individuals) but makes it particularly attractive for the clinical trials context. The GEE approach provides extensions of ordinary regression for continuous responses such as lesion burden, of Poisson regression for count responses such as active lesion counts and of logistic regression for binary responses such as active scan. It provides a unified approach to analysis of longitudinal quantitative MRI data [12].

*Fully parametric* approaches rely on specification of the joint distribution of the repeated responses of individual patients. With *random effects models* [13, 14], the focus is on the relationship of each patient's outcome measure to the values of that patient's covariates, so this is also referred to as *subject-specific modeling*. Another approach is based on *hidden Markov models* [15, 16]. A fully parametric model that adequately captures the detailed structure of MRI lesion count data over time in both *Rx* and *Pl* arms is an essential tool for addressing questions such as the optimal scheduling of MRI imaging for a Phase II clinical trial in MS.

Excellent texts are available that provide detailed developments of methods for the analysis of longitudinal data [17, 18]. Analyses based on summary measures over time may suffice for the primary analyses of the results of many clinical trials, but longitudinal analyses deserve a much more prominent role in the secondary analyses that are an essential part of the comprehensive review of the data collected in MS clinical trials. Indeed, some critical issues can be addressed properly only through the use of longitudinal analyses [19, 20].

### **Acknowledgments**

Parts of the text are based on the author's paper, Statistical methods for evaluating multiple sclerosis therapies, *Seminars in Neurology*, 18:351-375, 1998. This work was partially supported by a research grant from the NSERC of Canada.

### **References**

1. Matthews DE, Farewell V: *Using and Understanding Medical Statistics*. Basel: Karger, 1985.
2. Armitage P, Berry G: *Statistical Methods in Medical Research, 2<sup>nd</sup> edition*. Oxford: Blackwell, 1987.
3. Altman DG: *Practical Statistics for Medical Research*. London: Chapman & Hall, 1991.
4. van Belle G, Fisher LD, Heagerty PJ, Lumley T: *Biostatistics: A Methodology for the Health Sciences, 2<sup>nd</sup> edition*. New York: John Wiley & Sons, 2004.
5. The IFNB Multiple Sclerosis Group: Interferon beta-1b is effective in relapsing-remitting multiple sclerosis. I. Clinical results of a multicenter, randomized, double-blind, placebo-controlled trial. *Neurology* 43:655-661, 1993.
6. Paty DW, Li DKB, the UBC MS/MRI Study Group, the IFNB Multiple Sclerosis Group: Interferon beta-1b is effective in relapsing-remitting multiple sclerosis. II. MRI analysis results of a multicenter, randomized, double-blind, placebo-controlled trial. *Neurology* 43:662-667, 1993.
7. Machin D, Campbell MJ: *Statistical Tables for the Design of Clinical Trials*. Oxford: Blackwell, 1987.
8. Efron B, Tibshirani RJ: Statistical data analysis in the computer age. *Science* 253:390-395, 1991.
9. Efron B, Tibshirani RJ: *An Introduction to the Bootstrap*. New York: Chapman & Hall, 1993.
10. Zeger SL, Liang KY: An overview of methods for the analysis of longitudinal data. *Statistics in Medicine* 11:1825-1839, 1992.
11. Liang KY, Zeger SL: Regression analysis for correlated data. *Annual Review of Public Health* 14:43-68, 1993.

12. D'yachkova Y, Petkau J, White R: Longitudinal analyses for magnetic resonance imaging outcomes in multiple sclerosis clinical trials. *Journal of Biopharmaceutical Statistics* 7:501-531, 1997.
13. Laird NM, Ware JH: Random-effects models for longitudinal data. *Biometrics* 38:963-974, 1982.
14. Gibbons RD, Hedeker D, Elkin I, Waternaux C, Kraemer HC, Greenhouse JB, Shea MT, Imber SD, Sotsky SM, Watkins JT: Some conceptual and statistical issues in analysis of longitudinal psychiatric data. *Archives of General Psychiatry* 50:739-750, 1993.
15. MacDonald IL, Zucchini W: *Hidden Markov Models and Other Models for Discrete-Valued Time Series*. London: Chapman and Hall, 1997.
16. Altman RM, Petkau AJ: Application of hidden Markov models to multiple sclerosis lesion count data. *Statistics in Medicine* 24:2335-2344, 2005.
17. Diggle PJ, Heagerty P, Liang KY, Zeger SL: *Analysis of Longitudinal Data, 2<sup>nd</sup> edition*. New York: Oxford University Press, 2002.
18. Fitzmaurice GM, Laird NM, Ware JH: *Applied Longitudinal Analysis*. Hoboken: John Wiley & Sons, 2004.
19. Polman C, Kappos L, White R, Dahlke F, Beckmann K, Pozzilli C, Thompson A, Petkau J, Miller D, for the European Study Group in Interferon Beta-1b in Secondary Progressive Multiple Sclerosis: Neutralizing antibodies during treatment of secondary progressive multiple sclerosis with interferon beta-1b. *Neurology* 60:37-43, 2003.
20. Petkau AJ, White RA, Ebers GC, Reder AT, Sibley WA, Lublin FD, Paty DW, the INFB Multiple Sclerosis Study Group: Longitudinal analyses of the effects of neutralizing antibodies on interferon beta-1b in relapsing-remitting multiple sclerosis. *Multiple Sclerosis*, 10:126-138, 2004.