Statistical Methods in Translational Medicine

Shein-Chung Chow,1,3* Siu-Keung Tse,2 Min Lin1

This study focuses on strategies and statistical considerations for assessment of translation in language (e.g. translation of case report forms in multinational clinical trials), information (e.g. translation of basic discoveries to the clinic) and technology (e.g. translation of Chinese diagnostic techniques to well-established clinical study endpoints) in pharmaceutical/clinical research and development. However, most of our efforts will be directed to statistical considerations for translation in information. Translational medicine has been defined as bench-to-bedside research, where a basic laboratory discovery becomes applicable to the diagnosis, treatment or prevention of a specific disease, and is brought forth by either a physician–scientist who works at the interface between the research laboratory and patient care, or by a team of basic and clinical science investigators. Statistics plays an important role in translational medicine to ensure that the translational process is accurate and reliable with certain statistical assurance. Statistical inference for the applicability of an animal model to a human model is also discussed. Strategies for selection of clinical study endpoints (e.g. absolute changes, relative changes, or responder-defined, based on either absolute or relative change) are reviewed. [J Formos Med Assoc 2008;107(12 Suppl):S61–S73]

Key Words: bench-to-bedside, biomarker development, lost in translation, one-way translation, two-way translation

In early 2000, the United States Food and Drug Administration (FDA) presented the Critical Path Initiative to assist sponsors to identify possible causes of the scientific challenges underlying the medical product pipeline problems. The Critical Path Opportunities List released by the FDA on 16 March 2006 identified better evaluation tools and streamlining clinical trials as the top two areas to bridge the gap between the quick pace of new biomedical discoveries and the slower pace at which those discoveries are currently developed into therapies. This has led to the consideration of the use of adaptive design methods in clinical development and the focus of translational medicine (or science/research), which attempt not only to identify the best clinical benefit of a drug product under investigation, but also to increase the probability of success. Statistical methods for the use of adaptive trial designs in clinical development have been reported.1,2 In this article, however, we will focus on statistical methods that are commonly employed in translational medicine.

Chow has classified translational medicine into three areas, namely, translation in language, translation in information, and translation in (medical) technology.3 Translation in language refers to possible lost in translation of informed consent and/or case report forms in multinational clinical trials. Lost in translation is commonly encountered because of differences in language, perception, culture and medical practices. A typical approach for
assessment of the possible lost in translation is to first translate the informed consent and/or the case report forms by an experienced expert, and then translate them back by a different experienced but independent expert. The back-translated version is then compared with the original version for consistency. If the back-translated version passes the test for consistency, then it is validated through a small scale pilot study before it is applied to the intended multinational clinical trial. Translation in information is referred to as bench-to-bedside in translational science/research, which is also known as translational medicine. Translation in technology includes biomarker development and translation in diagnostic procedures between traditional Chinese and Western medicine. In this article, we will focus on statistical methods for translation in information and technology. Note that, in practice, translational medicine is often divided into two areas, namely, discovery and clinical translational medicine. Discovery translational medicine refers to biomarker development, bench-to-bedside, and animal versus human models, while clinical translational medicine includes translation among study endpoints, translation in technology, and generalization from a target patient population to another.

In the next section on Biomarker Development, statistical methods for biomarker development, especially for optimal variable screening in microarray analysis, are outlined. Also included in this section is a cross-validation method for model selection and validation. The section on Bench-to-Bedside discusses the ideas of a one-way and two-way translation process in pharmaceutical/clinical development. Whether or not an established animal model is predictive of a human model is examined in the section on Animal vs. Human Models. The Translation in Study Endpoints section summarizes translation among different study endpoints. Issues that are commonly encountered in translation in technology are described in the section on Bridging Studies. Also included in this section is the generalization of results obtained from a target patient population to another similar but different target patient population. This is followed by some concluding remarks.

**Biomarker Development**

A biomarker is an objectively measured and evaluated characteristic that is an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention. Biomarkers can be classified into classifier, prognostic and predictive markers. A classifier marker usually does not change over the course of the study and can be used to distinguish patients who would benefit from the treatment from those who would not. A typical example is a DNA marker for population selection in the enrichment process of clinical trials. A prognostic marker informs the clinical outcomes, which is independent of treatment. A predictive marker informs the treatment effect on the clinical endpoint, which could be population-specific. That is, a predictive marker can be predictive for population A but not population B. It should be noted that correlation between biomarkers and true endpoints makes a prognostic marker. However, correlation between biomarkers and true endpoints does not make a predictive biomarker.

In clinical development, a biomarker can be used to select the right population, identify the natural course of disease, assist in early detection of disease, and help develop personalized medicine. The utilization of biomarkers may lead to a better target population, detection of a large effect size with a smaller sample size, and timely decision making. As indicated in the FDA Critical Path Initiative Opportunity List, better evaluation tools call for biomarker qualification and standards. Statistical methods for early stage biomarker qualification include, but are not limited to: (1) distance-dependent K-nearest neighbors (DD-KNN); (2) K means clustering; (3) single/average/complete linkage clustering; and (4) distance-dependent Jarvis–Patrick clustering. More information can be found at http://www.ncifcrf.gov/human_studies.shtml.
We will now review statistical methods that are commonly used in biomarker development for optimal variable screening. The selected variables will then be used to establish a predictive model through a model selection/validation process.

**Optimal variable screening**

DNA microarrays have been used extensively in medical practice. Microarrays identify a set of candidate genes that are possibly related to a clinical outcome of a disease or a medical treatment. However, there are many more candidate genes than the number of available samples (the sample size) in almost all studies, which leads to an irregular statistical problem in disease diagnosis or treatment outcome prediction. Some available statistical methods deal with one single gene at a time,\(^4\) and/or combining several similar studies is often considered to increase sample size. These approaches may, however, not be appropriate because: (1) the combined dataset may still be much too small; and (2) there may be heteroscedasticity among the data from different studies. Alternatively, Shao and Chow proposed an optimal variable screening approach for dealing with the situation in which the number of variables (genes) is much larger than the sample size.\(^5\)

Let \(y\) be a clinical outcome of interest and \(x\) a vector of \(p\) candidate genes that are possibly related to \(y\). Shao and Chow simply considered inference on the population of \(y\) conditional on \(x\), and noted that their proposed method can be applied to the unconditional analysis (i.e. both \(y\) and \(x\) are random).\(^5\) Consider the following model:

\[
y = \beta'x + \varepsilon,
\]

where \(\beta\) is a \(p\)-dimensional vector and the distribution of \(\varepsilon\) is independent of \(x\) with \(E(\varepsilon) = 0\) and \(E(\varepsilon^2) = \sigma^2\). Under model (1), assume that there is a positive integer \(p_0\) (which does not depend on \(n\)), such that only \(p_0\) components of \(\beta\) are non-zero. Furthermore, \(\beta\) is in the linear space generated by the rows of \(X'X\) for sufficiently large \(n\), where \(X\) is the \(n \times p_n\) matrix whose \(i^{th}\) row is \(x_i\).

In addition, assume that there is a sequence \(\{\zeta_n\}\) of positive numbers such that \(\zeta_n \to \infty\) and \(\lambda_n = b_0 \zeta_n\), where \(\lambda_n\) is the \(i^{th}\) non-zero eigen value of \(X'X\), \(i=1,\ldots,n\), and \(\{b_0\}\) is a sequence of bounded positive numbers. Note that in many problems, \(\zeta_n = n\). Furthermore, there exists a constant \(c > 0\) such that \(p_n/\zeta_n \to 0\). For the estimation of \(\beta\), Shao and Chow considered the following ridge regression estimator:\(^5\)

\[
\hat{\beta} = (X'X + h_nI_n)^{-1}X'Y,
\]

where \(Y = (y_1, \ldots, y_n)'\), \(I_n\) is the identity matrix of order \(p_n\) and \(h_n > 0\) is the ridge parameter. The bias and variance of \(\hat{\beta}\) are given by:

\[
\text{bias}(\hat{\beta}) = E(\hat{\beta}) - \beta = - \left(h_n^{-1}X'X + I_n\right)^{-1}\beta
\]

and

\[
\text{var}(\hat{\beta}) = \sigma^2 \left(X'X + h_nI_n\right)^{-1}X'X\left(X'X + h_nI_n\right)^{-1}.
\]

Let \(\beta_i\) and \(\hat{\beta}_i\) be the \(i^{th}\) component of \(\beta\) and \(\hat{\beta}\), respectively. Under the assumptions as described earlier, we have \(E(\hat{\beta}_i - \beta_i)^2 \to 0\) (i.e. \(\hat{\beta}_i\) is consistent for \(\beta_i\) in mean squared error) if \(h_n\) is suitably chosen. Thus, we have

\[
\text{var}(\hat{\beta}) \to 0\quad \text{for all} \quad i \quad \text{as long as} \quad h_n \to \infty.
\]

Hence, \(\text{var}(\hat{\beta}) \to 0\) and \(\hat{\beta}_i\) is more complicated. Let \(\Gamma\) be an orthogonal matrix such that

\[
\Gamma X'X \Gamma = \begin{pmatrix}
\Lambda_n & 0_{n \times (p_n - n)} \\
0_{(n-p_n) \times n} & 0_{(p_n-n) \times (p_n-n)}
\end{pmatrix},
\]

where \(\Lambda_n\) is a diagonal matrix whose \(i^{th}\) diagonal element is \(\lambda_n\) and \(0_{l \times k}\) is the \(l \times k\) matrix of 0’s. Then, it follows that

\[
\text{bias}(\hat{\beta}) = - \left[\Gamma \left(\Gamma X'X \Gamma + I_n\right)^{-1}\Gamma^\prime\right] \beta = - \Gamma \Lambda \Gamma \beta,
\]

where \(A\) is a \(p_n \times p_n\) diagonal matrix whose first \(n\) diagonal elements are \(\frac{h_n}{h_n + \lambda_n}, i = 1, \ldots, n\), and the last diagonal elements are all equal to 1. Under the above mentioned assumptions, combining the
results for variance and bias of $\hat{\beta}_i$, i.e. equations (3) and (4), it can be shown that for all $i$

$$E(\hat{\beta}_i - \beta_i)^2 = \text{var}(\hat{\beta}_i) + \left[\text{bias}(\hat{\beta}_i)\right]^2 \to 0$$

if $h_n$ is chosen so that $h_n \to \infty$ at a rate slower than $\zeta_n$ (e.g. $h_n = \frac{s^{2/3}}{n}$). Based on this result, Shao and Chow proposed the following optimal variable screening procedure:\textsuperscript{5} let $\{a_n\}$ be a sequence of positive numbers satisfying $a_n \to 0$. For each fixed $n$, we screen out the $i^{th}$ variable if and only if $|\hat{\beta}_i| \leq a_n$.

Note that, after screening, only variables associated with $|\hat{\beta}_i| > a_n$ are retained in the model as predictors. The idea behind this variable screening procedure is similar to that in the Lasso method.\textsuperscript{6} Under certain conditions, Shao and Chow showed that their proposed optimal variable screening method is consistent in the sense that the probability that all variables (genes) unrelated to $y$, which will be screened out, and all variables (genes) related to $y$, which will be retained, is $1$ as $n$ tends to infinity.\textsuperscript{5}

**Model selection and validation**

Suppose that $n$ data points are available for selecting a model from a class of models. Several methods for model selection are available in the literature. These methods include, but are not limited to, Akaike information criterion (AIC),\textsuperscript{7,8} the $C_p$\textsuperscript{9} and the Jackknife and the bootstrap.\textsuperscript{10} These methods, however, are not asymptotically consistent in the sense that the probability of selecting the model with the best predictive ability does not converge to $1$ as the total number of observations $n \to \infty$. Alternatively, Shao proposed a method for model selection and validation using the method of cross-validation.\textsuperscript{11} The idea of cross-validation is to split the dataset into two parts. The first part contains $n_c$ data points, which will be used for fitting a model (model construction), whereas the second part contains $n = n_c$ data points, which are reserved for assessing the predictive ability of the model (model validation). It should be noted that all of the $n = n_c$ data, not just $n_c$, are used for model validation. Shao showed that all of the methods of AIC, $C_p$, Jackknife and bootstrap are asymptotically equivalent to the cross-validation with $n_c = 1$, denoted by CV(1), although they share the same deficiency of inconsistency.\textsuperscript{11} Shao indicated that the inconsistency of the leave-one-out cross-validation can be rectified by using leave-$n_c$-out cross-validation with $n_c$ satisfying $n_c/n \to 1$ as $n \to \infty$.\textsuperscript{11}

In addition to the cross-validation with $n_c = 1$, denoted by CV(1), Shao also considered the other two cross-validation methods, namely, a Monte Carlo cross-validation with $n_c(n_c \neq 1)$, denoted by MCCV($n_c$), and an analytic approximate CV($n_c$), denoted by APCV($n_c$). MCCV($n_c$) is a simple and easy method that utilizes the method of Monte Carlo by randomly drawing (with or without replacement) a collection of $b$ subsets of $\{1, \ldots, n\}$ that have size $n_c$ and select a model by minimizing

$$\hat{\Gamma}_{a_n} = \frac{1}{n_c b} \sum_{i=1}^{n_c} \sum_{b=1}^{b} \left[ y_{i} - \hat{y}_{a_n} \right]^2.$$

On the other hand, APCV($n_c$) selects the optimal model based on the asymptotic leading term of balance incomplete CV($n_c$), which treats each subset as a block and each $i$ as a treatment. Shao compared these three cross-validation methods through a simulation study under the following model with five variables with $n = 40$:\textsuperscript{11}

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + e_i,$$

where $e_i$ are independent and identically distributed (i.i.d.) normal variable with mean 0 and variance 1, i.e. $N(0,1)$, $x_{ik}$ is the $i^{th}$ value of the $k^{th}$ prediction variable $x_{ik}$, $x_{ik} = 1$, and the values of $x_{ik}$, $k = 2, \ldots, 5$, $i = 1, \ldots, 40$, are taken from an example in Gunst and Mason.\textsuperscript{12} Note that there are 31 possible models, and each model is denoted by a subset of $\{1, \ldots, 5\}$ that contains the indices of the variable $x_k$ in the model. Shao indicated that MCCV($n_c$) has the best performance among the three methods under study, except for the case in which the largest model is the optimal model. APCV($n_c$) is slightly better than the CV(1) in all cases.\textsuperscript{11} CV(1) tends to select unnecessarily large models. The probability of selecting the optimal
Using the CV(1) model could be very low (e.g. < 0.5).

Remarks
In practice, it is suggested that the optimal variable screening method proposed by Shao and Chow be applied to select a few relevant variables, say 5–10. After this, apply the cross-validation method to select the optimal model based on linear model selection or non-linear model selection. The selected model can then be validated based on the cross-validation methods as described in the previous subsection. Note that the method for variable screening described in this section is very similar to the method based on least absolute shrinkage and selection operator.

The selection of biomarker should be based on the consideration of the diagnostic accuracy and time for evaluation. However, the FDA approved MammaPrint® has a very poor positive predictive value for metastatic disease. Thus, it is suggested that the development of biomarker for long-term prognosis of clinical outcomes should be based on the prediction of treatment effect.

Bench-to-Bedside
Pizzo defined translational medicine as bench-to-bedside research, wherein a basic laboratory discovery becomes applicable to the diagnosis, treatment or prevention of a specific disease, and is brought forth by either a physician–scientist who works at the interface between the research laboratory and patient care, or by a team of basic and clinical science investigators. Thus, translational medicine refers to the translation of basic research discoveries into clinical applications. More specifically, translational medicine takes the discoveries from basic research to a patient and measures an endpoint in a patient. Recently, scientists have become increasingly aware that this bench-to-bedside approach to translational research is a two-way street. Scientists provide clinicians with new tools for use with patients and for assessment of their impact, and clinical researchers make novel observations about the nature and progression of disease that often stimulate basic investigations. As indicated by Pizzo, translational medicine can also have a much broader definition, referring to the development and application of new technologies, biomedical devices, and therapies in a patient-driven environment such as clinical trials, in which the emphasis is on early patient testing and evaluation. Thus, translational medicine also includes epidemiologic and health-outcomes research and behavioral studies that can be brought to the bedside or ambulatory setting.

Mankoff et al pointed out that there are three major obstacles to effective translational medicine in practice. The first is the challenge of translating basic science discoveries into clinical studies. The second hurdle is the translation of clinical studies into medical practice and health care policy. A third obstacle to effective translational medicine is philosophical. It may be a mistake to think that basic science (without observations from the clinic and without epidemiologic findings of possible associations between different diseases) will efficiently produce the novel therapies for human testing. Pilot studies such as non-human and non-clinical studies are often used to transform therapies developed using animal models to a clinical setting. Statistical processes play an important role in translational medicine. In particular, we define a statistical process of translational medicine as a translational process for: (1) determining association between some independent parameters observed in basic research discoveries and a dependent variable observed from clinical application; (2) establishing a predictive model between the independent parameters and the dependent response variable; and (3) validating the established predictive model. As an example, in animal studies, the independent variables may include in vitro assay results, pharmacologic activities such as pharmacokinetics and pharmacodynamics, and dose toxicities, and the dependent variable could be a clinical outcome (e.g. a safety parameter).
If the information observed with basic research discoveries is translated to the clinic, it is usually referred to as a one-way translation process. However, as indicated by Pizzo, the translational process should be a two-way translation, in which information obtained in clinical study can also be translated back to basic research discoveries. Tse has described the statistical process in how to validate a translational process, either one-way or two-way, based on some probability-based criteria. Furthermore, a measure to assess the degree of lost in translation has also been proposed.\textsuperscript{16}

### Animal vs. Human Models

In translational medicine, a commonly asked question is whether an animal model is predictive of a human model. To address this question, we may assess the similarity between an animal model (population) and a human model (population). For this purpose, we first establish an animal model to bridge the basic research discovery ($x$) and clinic ($y$). For illustration purposes, consider a one-way translation process of $y = \beta_0 + \beta_1 x + \varepsilon$. Let $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ be the predictive model obtained from the one-way translation based on data from an animal population. Thus, for a given $x_0$, $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$ follows a distribution with mean $\mu_y$ and $\sigma_y^2$. Under the predictive model $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$, denoted by $(\mu_y, \sigma_y)$ the target population, and assume that the predictive model works for the target population. Thus, for an animal population, $\mu_y = \mu_{\text{animal}}$ and $\sigma_y = \sigma_{\text{animal}}$ while for a human population, $\mu_y = \mu_{\text{human}}$ and $\sigma_y = \sigma_{\text{human}}$. Assuming that the linear predictive model can be applied to both animal and human populations, we can link the animal and human models by the following:

$$\mu_{\text{human}} = \mu_{\text{animal}} + \varepsilon,$$

and

$$\sigma_{\text{human}} = C\sigma_{\text{animal}}.$$

In other words, we expect that there are differences in population mean and population standard deviation (SD) under the predictive model caused by possible differences in response between animals and humans. As a result, the effect size adjusted for SD under the human population can be obtained as follows:

$$\frac{\mu_{\text{human}}}{\sigma_{\text{human}}} = \mu_{\text{animal}} + \epsilon$$

$$\frac{\sigma_{\text{human}}}{\sigma_{\text{animal}}} = C\sigma_{\text{animal}}$$

where $\Delta = (1 + \epsilon)\mu_{\text{animal}})/C$. Chow et al refer to $\Delta$ as a sensitivity index when changing from one target population to another.\textsuperscript{17} As can be seen, the effect size under the human population is inflated (or reduced) by the factor $\Delta$. If $\epsilon = 0$ and $C = 1$, we then claim that there is no difference between the animal and human populations. Thus, the animal model is predictive of the human model. Note that the shift and scale parameters (i.e. $\epsilon$ and $C$) can be estimated by

$$\hat{\epsilon} = \hat{\mu}_{\text{human}} - \hat{\mu}_{\text{animal}}$$

and

$$\hat{C} = \frac{\hat{\sigma}_{\text{human}}}{\hat{\sigma}_{\text{animal}}}$$

respectively, in which $(\hat{\mu}_{\text{animal}}, \hat{\sigma}_{\text{animal}})$ and $(\hat{\mu}_{\text{human}}, \hat{\sigma}_{\text{human}})$ are estimates of $(\mu_{\text{animal}}, \sigma_{\text{animal}})$ and $(\mu_{\text{human}}, \sigma_{\text{human}})$, respectively. Thus, the sensitivity index can be assessed as follows:

$$\hat{\Delta} = (1 + \hat{\epsilon}/\hat{\mu}_{\text{animal}})/\hat{C}.$$

In practice, there may be a shift in population mean (i.e. $\epsilon$) and/or in population SD (i.e. $C$). Chow et al indicated that shifts in population mean and population SD can be classified into the following four cases, in which: (1) both $\epsilon$ and $C$ are fixed; (2) $\epsilon$ is random and $C$ is fixed; (3) $\epsilon$ is fixed and $C$ is random; and (4) both $\epsilon$ and $C$ are random. For the case in which both $\epsilon$ and $C$ are fixed, $(\nu)$ can be used for the estimation of $\Delta$\textsuperscript{18} Chow et al derived statistical inference of $\Delta$ for the case in which $\epsilon$ is random and $C$ is fixed by assuming that $y$ conditional on $\mu$ follows a normal distribution, $N(\mu, \sigma^2)$. That is,

$$y\mid\mu \sim N(\mu, \sigma^2),$$

where $\mu$ is distributed as $N(\mu_\mu, \sigma_\mu^2)$, and $\sigma$, $\mu_\mu$ and $\sigma_\mu$ are some unknown constants. It can be
verified that \( y \) follows a mixed normal distribution with mean \( \mu_y \) and variance \( \sigma^2 + \sigma_y^2 \). That is, \( y \sim N(\mu_y, \sigma^2 + \sigma_y^2) \). As a result, the sensitivity index can be assessed based on data collected from both animal and human populations under the predictive model.

Note that for other cases in which \( C \) is random, the above method can also be derived similarly. The assessment of sensitivity index can be used to adjust the treatment effect to be detected under a human model when applying an animal model to a human model, especially when there is a significant or major shift between an animal and human population. In practice, it is of interest to assess the impact of the sensitivity index on lost in translation and the probability of success. This, however, requires further research.

**Translation in Study Endpoints**

In clinical trials, it is not uncommon that a study is powered based on expected absolute change from baseline of a primary study endpoint, but the collected data are analyzed based on relative change from baseline (e.g. percent change from baseline) of the primary study endpoint. In many cases, the collected data are analyzed based on the percentage of patients who show some improvement (i.e. responder analysis). The definition of a responder may be based on either absolute change from baseline or relative change from baseline of the primary study endpoint. It is controversial in terms of the interpretation of the analysis results, especially when a significant result is observed based on a study endpoint (e.g. absolute change from baseline, relative change from baseline, or responder analysis). In practice, it is of interest to explore how an observed significant difference of a study endpoint (e.g. absolute change from baseline, relative change from baseline, or responder analysis) can be translated to that of the other study endpoint (e.g. absolute change from baseline, relative change from baseline, or responder analysis).

**Power analysis and sample size calculation**

The power analysis for sample size calculation has an immediate impact on the assessment of treatment effect based on different study endpoints. For example, sample size required in order to achieve the desired power based on the absolute change may be very different from that obtained based on the percent change, or the percentage of patients who show an improvement based on the absolute change or relative change at the \( \alpha \) level of significance. Denote the measurements of the \( i \)th subject before and after the treatment by \( w_{ij} \) and \( w_{j2} \), respectively. For illustration purposes, assume that \( w_{ij} \) are log normal distributed, i.e. \( \log w_{ij} \sim N(\mu, \sigma^2) \). Let \( w_{j2} = w_{ij} (\bar{w}_{ij} + 1) \), where \( \log \bar{w}_{ij} \sim N(\mu, \sigma^2) \), and \( w_{ij} \) and \( \bar{w}_{ij} \) are independent for \( 1 \leq i, j \leq n \). It follows that \( \log (w_{j2} - w_{ij}) \sim N(\mu_2 + \mu_1, 2\sigma^2) \) and \( \log \left( \frac{w_{j2} - w_{ij}}{w_{ij}} \right) \sim N(\mu_2, \sigma^2) \). Define \( X_i \) and \( Y_i \) as \( X_i = \log (w_{j2} - w_{ij}), Y_i = \log \left( \frac{w_{j2} - w_{ij}}{w_{ij}} \right) \). Then \( X_i \) and \( Y_i \) represent the logarithm of the absolute and relative changes of the measurements before and after treatment. It can be shown that both \( X_i \) and \( Y_i \) are normally distributed.

Let \( \mu_{AC} \) and \( \mu_{BC} \) be the population means of the logarithm of the absolute and relative change of a primary study endpoint of a given clinical trial, respectively. Thus, the hypotheses of interest based on the absolute change are given by \( H_{01}: \mu_{AC} \leq \delta_0 \) vs. \( H_{a1}: \mu_{AC} > \delta_0 \), where \( \delta_0 \) is the difference of clinical importance. For the relative change, the hypotheses of interest are given by \( H_{02}: \mu_{BC} \leq \Delta_0 \) vs. \( H_{a2}: \mu_{BC} > \Delta_0 \), where \( \Delta_0 \) is the difference of clinical importance. In practice, for a specific value of \( \delta \), it would be of interest to determine what value

\( S67 \)
of $\Delta$ (with $\Delta > \Delta_0$) would be equivalent to $\delta$
clinically.

In addition, if we consider that a patient is a
responder if the logarithm of his/her absolute change of the primary study endpoint is greater
than $\delta$, then it is of interest to test the following
hypotheses:

$$H_0: P_{\text{AC}} = \eta \text{ vs. } H_{\text{ai}}: P_{\text{AC}} > \eta,$$

where $P_{\text{AC}}$ is the proportion of patients whose
logarithm of absolute change of the primary study endpoint is greater than $\delta$. In practice, we may
claim superiority (clinically) of the test treatment if we reject the null hypothesis at $\eta = 50\%$, and
favor the alternative hypothesis that $P_{\text{AC}}$ is $> 50\%$. However, this lacks statistical justification. For a
noninferiority (or superiority) trial, it is often of
interest to know how the selection of a noninferi-
ority margin of $\mu_{\text{AC}}$ can be translated to the non-inferiority margin of $P_{\text{AC}}$. Similarly, for the relative
change, the hypotheses of interest are given by

$$H_0: P_{\text{RC}} = \eta \text{ vs. } H_{\text{ai}}: P_{\text{RC}} > \eta,$$

where $P_{\text{RC}}$ is the proportion of patients whose
logarithm of the relative change of the primary study endpoint is greater than $\delta$. In particular, $\alpha = 0.05$, power $= 1 - \beta = 0.8$, and $\eta = 0.5$, which provide a better understanding of the required sample size to achieve a pre-
determined power level for the above four sets of hypotheses.

**An example**

As an example, consider a clinical trial for
evaluation of possible weight reduction of a test
treatment in female patients. Weight data from 10 subjects are given in Table 1. As seen in Table 1,
mean absolute change and mean percent change

<table>
<thead>
<tr>
<th>Subject</th>
<th>Absolute Change</th>
<th>Percent Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>7</td>
<td>15</td>
</tr>
<tr>
<td>3</td>
<td>9</td>
<td>20</td>
</tr>
<tr>
<td>4</td>
<td>10</td>
<td>25</td>
</tr>
<tr>
<td>5</td>
<td>8</td>
<td>18</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>12</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Subject</th>
<th>Absolute Change</th>
<th>Percent Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>7</td>
<td>15</td>
</tr>
<tr>
<td>12</td>
<td>9</td>
<td>20</td>
</tr>
<tr>
<td>13</td>
<td>11</td>
<td>25</td>
</tr>
<tr>
<td>14</td>
<td>13</td>
<td>30</td>
</tr>
<tr>
<td>15</td>
<td>15</td>
<td>35</td>
</tr>
</tbody>
</table>

**Figure 1.** Plot of $(\delta - \delta_0)$ versus $n$.

**Figure 2.** Plot of $(\Delta - \Delta_0)$ versus $n$.

**Figure 3.** Plot of $\ln \delta$ versus $n$.

**Figure 4.** Plot of $\ln \Delta$ versus $n$. 
from pretreatment are 5.3 lbs and 4.8%, respectively. If a subject is considered a responder if there is weight reduction of >5 lbs (absolute change) or >5% (relative change), the response rates based on absolute and relative change are given as 40% and 30%, respectively. For illustration purposes, Table 2 summarizes the sample sizes required for achieving the desired power for detecting a clinically meaningful difference, say, by an absolute change of 5.5 lbs and a relative change of 5.5%, for the two study endpoints respectively.

As shown in Table 2, a sample size of 190 is required for achieving an 80% power for detecting a difference of 5.5 lbs (posttreatment absolute change from pretreatment) at the 5% level of significance, while a much larger sample size of 95 is required in order to have an 80% power for detecting a difference of 5.5% (relative change between posttreatment and pretreatment) at the 5% level of significance. Based on responder analysis, the results are different. Based on the analysis of responders, defined as subjects who have weight reduction ≥5.5 lbs, a total sample size of 54 subjects is needed for detecting a 50% improvement at the 5% level of significance. On the other hand, if we define a responder as a subject who has a weight reduction ≥5.5%, then a total sample size of 52 subjects is required for achieving a 50% improvement at the 5% level of significance.

**Remarks**

As discussed above, sample sizes required for achieving the desired power for detecting a clinically meaningful difference at the 5% level of significance may be very different depending upon the choice of study endpoint and a clinically meaningful difference. In practice, it will be more complicated if the intended trial is to establish noninferiority. In this case, sample size calculation will also depend on the selection of the noninferiority margin. To ensure the success of the intended clinical trial, the sponsor will usually evaluate carefully several clinical strategies when selecting the type of study endpoint, clinically meaningful difference, and noninferiority margin during the stage of protocol development. Commonly considered study endpoints are:

- measure based on absolute change;
- measure based on relative change;
- proportion of responders defined based on absolute change;
- proportion of responders defined based on relative change.

In some cases, investigators may consider a composite endpoint based on absolute and relative change. For example, in clinical trials for evaluation of the efficacy and safety of a compound for treating patients with active ulcerative colitis, a study endpoint utilizing the so-called Mayo score is often considered. The investigator may define a subject as a responder if he/she has a decrease from baseline in the total score of at least 3 points and at least 30%, with an accompanying decrease in the subscore for rectal bleeding of at least 1 point or absolute subscore for rectal bleeding of 0 or 1 at day 57. Note that the

<p>| Table 1. Weight data from 10 female subjects |</p>
<table>
<thead>
<tr>
<th>Pretreatment</th>
<th>Posttreatment</th>
<th>Absolute change</th>
<th>Relative change (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>110</td>
<td>106</td>
<td>4</td>
<td>3.6</td>
</tr>
<tr>
<td>90</td>
<td>80</td>
<td>10</td>
<td>11.1</td>
</tr>
<tr>
<td>105</td>
<td>100</td>
<td>5</td>
<td>4.8</td>
</tr>
<tr>
<td>95</td>
<td>93</td>
<td>2</td>
<td>2.1</td>
</tr>
<tr>
<td>170</td>
<td>163</td>
<td>7</td>
<td>4.1</td>
</tr>
<tr>
<td>90</td>
<td>84</td>
<td>6</td>
<td>6.7</td>
</tr>
<tr>
<td>150</td>
<td>145</td>
<td>5</td>
<td>3.3</td>
</tr>
<tr>
<td>135</td>
<td>131</td>
<td>4</td>
<td>3.0</td>
</tr>
<tr>
<td>160</td>
<td>159</td>
<td>1</td>
<td>0.6</td>
</tr>
<tr>
<td>100</td>
<td>91</td>
<td>9</td>
<td>9.0</td>
</tr>
<tr>
<td>120.5 (30.5)</td>
<td>115.2 (31.5)</td>
<td>5.3</td>
<td>4.8</td>
</tr>
</tbody>
</table>

| Table 2. Sample size calculation |
| Study endpoint                  | Clinically meaningful difference | Sample size required |
| Absolute change                 | 5 lb                            | 190                  |
| Relative change                 | 5%                             | 95                   |
| Responder 1*                    | 50% improvement                 | 54                   |
| Responder 2†                    | 50% improvement                 | 52                   |

*Responder is defined based on absolute change ≥5.5 lbs; †responder is defined based on relative change ≥5.5%.
Mayo scoring system for assessment of ulcerative colitis activity consists of the three domains of: Mayo score; partial Mayo score; and mucosal healing.\(^{19}\)

In addition to the four types of study endpoints which are derived from the clinical data collected from the sample patient population, clinically meaningful differences or noninferiority margins that we would like to detect or establish may be based on absolute or relative change. For example, based on responder analysis, we may wish to detect a 30% difference in response rate or to detect a 50% relative improvement in response rate. As a result, there are a total of eight clinical strategies for assessment of the treatment effect. In practice, some strategies may lead to the success of the intended clinical trial (i.e. achieve the study objectives with the desired power), while some strategies may not. A common practice for the sponsor is to choose the strategy that corresponds to their best interest. However, regulatory agencies may challenge the sponsor as to the inconsistent results. This has raised the following questions. First, how do we translate the clinical information among different study endpoints since they are obtained based on the same data collected from the same patient population? Second, which study endpoint is telling the truth? These questions, however, remain unanswered. More research is required. The current regulatory position is to require the sponsor to prespecify which study endpoint will be used for assessment of the treatment effect in the study protocol, without any scientific justification.

**Bridging Studies**

In recent years, the influence of ethnic factors on clinical outcomes for evaluation of efficacy and safety of study medications under investigation has attracted much attention from regulatory authorities, especially when the sponsor is interested in bringing forward an approved drug product from the original region (e.g. the USA or European Union) to a new region (e.g. Asia-Pacific region). To determine if clinical data generated from the original region are acceptable in the new region, the International Conference on Harmonization (ICH) has issued a guideline on ethnic factors in the acceptability of foreign clinical data.\(^{20}\) The purpose of this guideline is not only to permit adequate evaluation of the influence of ethnic factors, but also to minimize duplication of clinical studies in the new region. This guideline is known as the ICH E5 guideline.

As indicated in the ICH E5 guideline, a bridging study is defined as a study performed in the new region to provide pharmacokinetic, pharmacodynamic, or clinical data on efficacy, safety, dosage, and regimen in the new region, which will allow extrapolation of the foreign clinical data to the population in the new region. The ICH E5 guideline suggests that the regulatory authority of the new region assesses the ability to extrapolate foreign data based on the bridging data package, which consists of: (1) information including pharmacokinetic data and any preliminary pharmacodynamic and dose–response data from the complete clinical data package (CCDP), which is relevant to the population of the new region; and if needed, (2) a bridging study to extrapolate the foreign efficacy and/or safety data to the new region. The ICH E5 guideline indicates that bridging studies may not be necessary if the study drugs are insensitive to ethnic factors. For drugs that are characterized as insensitive to ethnic factors, the type of bridging studies (if needed) will depend on experience with the drug class and on the likelihood that extrinsic ethnic factors will affect the drug’s safety, efficacy and dose response. On the other hand, for drugs that are ethnically sensitive, a bridging study is usually needed since the populations in the two regions are different. In the ICH E5 guideline, however, no criteria for assessment of the sensitivity to ethnic factors for determining whether a bridging study is needed are provided. Moreover, when a bridging study is conducted, the ICH guideline indicates that the study is readily interpreted as capable of bridging the foreign data if it shows that dose response, safety, and efficacy in the new region are similar...
to those in the original region. However, the ICH does not clearly define the similarity.

Shih has interpreted this as consistency among study centers by treating the new region as a new center of multicenter clinical trials.\(^{21}\) Under this definition, Shih proposed a method for assessment of consistency to determine whether the study is capable of bridging the foreign data to the new region. Alternatively, Shao and Chow proposed the concepts of reproducibility and generalizability probabilities for assessment of bridging studies.\(^{22}\) If the influence of the ethnic factors is negligible, then we may consider the reproducibility probability to determine whether the clinical results observed in the original region are reproducible in the new region. If there is a notable ethnic difference, the generalizability probability can be assessed to determine whether the clinical results in the original region can be generalized in a similar, but slightly different patient population as a result of a difference in ethnic factors. In addition, Chow et al proposed assessment of bridging studies based on the concept of population (or individual) bioequivalence.\(^{17}\)

Along the same lines, Hung, alone and with others, considered the assessment of similarity based on testing for noninferiority between a bridging study conducted in the new region, compared with the previous study conducted in the original region.\(^{23,24}\) This led to arguments regarding the selection of a noninferiority margin.\(^{25}\) Note that other methods such as the use of the Bayesian approach have also been proposed.\(^{26}\)

Test for consistency

For assessment of similarity between a bridging study conducted in a new region and studies conducted in the original region, Shih considered all of the studies conducted in the original region as a multicenter trial and proposed to test consistency among study centers by treating the new region as a new center of a multicenter trial.\(^{21}\)

Suppose that there are \(K\) reference studies in the CCDP. Let \(T_i\) denote the standardized treatment group difference, i.e.

\[
T_i = \frac{\bar{x}_i - \bar{x}_m}{s_i \sqrt{\frac{1}{m_{ci}} + \frac{1}{m_{ci}}}},
\]

where \(\bar{x}_i(\bar{x}_m)\) is the sample mean of \(m_{ci}(m_{cm})\) observations in the treatment (control) group, and \(s_i\) is the pooled sample SD. Shih considered the following predictive probability for testing consistency:\(^{21}\)

\[
p(T \mid T_i, i = 1, \ldots, K) = \left\{ \frac{2\pi(K + 1)}{K} \right\}^{-K/2} \exp\left[ -K(T - \bar{T})^2 / 2(K + 1) \right].
\]

**Test for reproducibility and generalizability**

When the ethnic difference is negligible, Shao and Chow suggested assessing reproducibility probability for similarity between clinical results from a bridging study and studies conducted in the CCDP.\(^{22}\) Let \(x\) be a clinical response of interest in the original region. Let \(y\) be similar to \(x\), but a response in a clinical bridging study conducted in the new region. Suppose the hypotheses of interest are:

\[
H_0: \mu_i = \mu_0 \quad \text{vs.} \quad H_a: \mu_i \neq \mu_0.
\]

We reject \(H_0\) at the 5% level of significance if and only if \(|T| > t_{n-2}\), where \(t_{n-2}\) is the \((1 - \alpha/2)^{th}\) percentile of the \(t\) distribution with \(n - 2\) degrees of freedom, \(n = n_1 + n_2\),

\[
T = \bar{y} - \bar{x} \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n - 2}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}},
\]

and \(\bar{x}, \bar{y}, s_1^2\) and \(s_2^2\) are sample means and variances for the original region and the new region, respectively. Thus, the power of \(T\) is given by:

\[
p(\theta) = P(|T| > t_{n-2}) = 1 - \Phi_{\frac{-1}{\sqrt{n_1} + \sqrt{n_2}}}(-t_{n-2}\theta) + \Phi_{\frac{1}{\sqrt{n_1} + \sqrt{n_2}}}(t_{n-2}\theta),
\]

where

\[
\theta = \frac{\mu_i - \mu_0}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},
\]

and \(\Phi_{\frac{1}{\sqrt{n_1} + \sqrt{n_2}}}(\lambda|\theta)\) denotes the cumulative distribution function of the non-central \(t\) distribution with
\( n-2 \) degrees of freedom and the non-centrality parameter \( \theta \). Replacing \( \theta \) in the power function with its estimate \( T(x) \), the estimated power

\[
\hat{p} = P(T(x)) = 1 - F_{n-2}(t_{n-2} \mid T(x)) + F_{n-2}(-t_{n-2} \mid T(x))
\]

is defined as a reproducibility probability for a future clinical trial with the same patient population. Note that when the ethnic difference is notable, Shao and Chow recommended assessing the so-called generalizability probability for similarity between clinical results from a bridging study and studies conducted in the CCPD.\(^{22}\)

**Test for similarity**

Using the criterion for assessment of population (individual) bioequivalence, Chow et al proposed the following measure of similarity between \( x \) and \( y \)\(^{17}\):

\[
\theta = \frac{E(x - y)^2 - E(x - x')^2}{E(x - x')^2 / 2},
\]

where \( x' \) is an independent replicate of \( x \), and \( y, x \) and \( x' \) are assumed to be independent. Since a small value of \( \theta \) indicates that the difference between \( x \) and \( y \) is small (relative to the difference between \( x \) and \( x' \)), similarity between the new and original regions can be claimed if and only if \( \theta < \theta_{\text{ul}} \) where \( \theta_{\text{ul}} \) is a similarity limit. Thus, the problem of assessing similarity becomes a problem of testing the following hypotheses:

\[
H_{0\text{c}}: \theta \geq \theta_{\text{ul}} \quad \text{vs.} \quad H_{1\text{c}}: \theta < \theta_{\text{ul}}.
\]

Let \( k = 0 \) indicate the original region and \( k = 1 \) indicate the new region. Suppose that there are \( m_k \) study centers and \( n_k \) responses in each center for a given variable of interest. For simplicity, we only consider the balanced case in which centers in a given region have the same number of observations. Let \( z_{ijk} \) be the \( i^{\text{th}} \) observation from the \( j^{\text{th}} \) center of region \( k \), let \( b_{jk} \) be the between-center random effect, and let \( e_{ijk} \) be the within-center measurement error. Assume that

\[
z_{ijk} = \mu_k + b_{jk} + e_{ijk}, i = 1, \ldots, n_k, j = 1, \ldots, m_k, k = 0, 1,
\]

where \( \mu_k \) is the population mean in region \( k \), \( b_{jk} \sim N(0, \sigma_{b_k}^2), e_{ijk} \sim N(0, \sigma_{e_k}^2) \) and \( \{b_{jk}\} \) and \( \{e_{ijk}\} \) are independent. Under the above model, the criterion for similarity becomes

\[
\theta = (\mu_0 - \mu_1)^2 + \sigma_{b_k}^2 - \sigma_{e_k}^2, \quad \sigma_{e_k}^2 = \sigma_{b_k}^2 + \sigma_{e_k}^2
\]

where \( \sigma_{e_k}^2 = \sigma_{b_k}^2 + \sigma_{e_k}^2 \) is the total variance (between-center variance plus within-center variance) in region \( k \). The above hypotheses are equivalent to

\[
H_{0\text{c}}: \zeta \geq 0 \quad \text{vs.} \quad H_{1\text{c}}: \zeta < 0,
\]

where \( \zeta = (\mu_0 - \mu_1)^2 + \sigma_{b_k}^2 - (1 + \theta_{\text{ul}})\sigma_{e_k}^2 \).

**Remarks**

Liu et al proposed a Bayesian approach to use a normal prior to taking up the strength from CCPD for the evaluation of similarity between the new and original regions.\(^{26}\) Their approach, however, has been criticized in cases where there is a serious imbalance in the information provided between the new and original regions. Alternatively, Hsiao et al considered a mixture model for the prior information (which is a weighted average of a non-informative prior and a normal prior) based on the concept of positive treatment effect.\(^{27}\) Note that a group sequential method and a two-stage design for assessment of similarity between the new and original regions has also been proposed.\(^{28,29}\)

**Concluding Remarks**

Translational medicine is a multidisciplinary entity that bridges basic scientific research with clinical development. As the expense in developing therapeutic pharmaceutical compounds continues to increase and the success rates for getting such compounds approved for marketing and to the patients needing these treatments continues to decrease, a focused effort has emerged in improving the communication and planning between basic and clinical science. This will likely lead to
more therapeutic insights being derived from new scientific ideas, and more feedback being provided back to research so that its approaches are better targeted. Translational medicine spans all the disciplines and activities that lead to making key scientific decisions as a compound traverses across the difficult preclinical–clinical divide. Many argue that improvement in making correct decisions on which dose and regimen should be pursued in the clinic, the likely human safety risks of a compound, the likely drug interactions, and the pharmacologic behavior of the compound, are likely to be the most important decisions made in the entire development process. Many of these decisions and the path for uncovering this information within later development are defined at this specific time within the drug development process. Improving these decisions will likely lead to a substantial increase in the number of safe and effective compounds available to combat human diseases.

References