Chapter 15
Overview of Drug Development and Statistical Tools for Manufacturing and Testing

John Peterson and Stan Altan

Abstract An essential part of the application for marketing approval of a new drug or therapeutic biological product submitted to regulatory authorities is the Chemistry Manufacturing and Controls (CMC) section. It presents the sponsor company's documentation of sufficient scientific and engineering knowledge to manufacture the product with consistent quality that provides defined clinical efficacy with an acceptable safety profile. The CMC section includes three main parts: (1) Chemical Development (synthesis of a new molecular entity (NME), purification of a new biologic entity (NBE); (2) Pharmaceutical Development (comprised of formulation and process development); (3) Analytical Development (analytical methods for physical, chemical, biological characterization). Specifications are established during development and constitute an important part of the CMC section in defining the product's quality requirements. This chapter provides an overview of the drug development process, and some statistical tools useful in support of CMC studies. This chapter aims to set the stage for the subsequent 11 chapters in the CMC section of this book which delve in greater detail into important CMC related statistical topics.

Keywords Chemistry • Manufacturing and Controls • Quality by design • Bioassay • Probability of conformance • Bayesian method • Drug development

15.1 Introduction to Drug Development, Manufacturing and Analytical Testing

The Chemistry Manufacturing and Controls (CMC) aspects of drug development is an essential component of the overall process necessary for sponsor companies

J. Peterson (✉)
Statistical Sciences Group, GlaxoSmithKline Pharmaceuticals, Upper Providence, PA 19426, USA
e-mail: john.peterson@gsk.com

S. Altan
Nonclinical Statistics and Computing, Janssen R&D LLC, Raritan, NJ 08869, USA

to obtain approval to market a drug in the United States and elsewhere in the world. In the United States, such approval requires the filing of a New Drug Application (NDA) or a Biologics License Applications (BLA) to the Federal Drug Administration (FDA). We will not delve into the details of the drug approval process and associated interactions with regulatory agencies but suffice it to say that the CMC section is a critical part of the NDA submission. It includes three main areas: (1) Chemical Development (synthesis of a new molecular entity (NME) or new chemical entity (NCE), purification of a new biologic entity (NBE), also referred to as an Active Pharmaceutical Ingredient (API) or biological API or Drug Substance),[1] (2) Pharmaceutical Development (comprised of formulation and process development), and (3) Analytical Development (analytical methods for physical, chemical/biological characterization).

Early CMC studies follow on the heels of the decision to develop a NME based on an expectation that the NME has an important therapeutic effect and is safe in humans. These early studies occur prior to the filing of the Investigational New Drug Exemption (IND) application with the FDA, which is necessary to study the NME in humans. At this point, initial toxicology and animal pharmacology studies are initiated to characterize the safety profile of the compound and its physiologic effects in animal models. At the same time, the early formulation investigations are taking place, to develop a mixture of the NME plus additional materials, known as excipients, which will provide the scientific understanding to find a unique formula and dosage form (e.g. solids such as a tablet or capsule, injectables such as a prefilled syringe, nasal spray and inhalation solution) to maintain its chemical stability and bioavailability suitable for human dosing in the early clinical trials. Such a mixture is referred to as the drug product, distinct from the Drug Substance which is the raw NME.

These early CMC studies, following the filing of the IND, lead to more extensive studies to refine the formulation, and to characterize its chemical and physical properties in increasingly greater detail. In addition, engineering and process studies are carried out to determine an optimal choice of manufacturing conditions as well as the development of state of the art chemical and biological analytical tools to measure the potency, purity and other characteristics important to assuring the identity of the drug product. Equally important, parallel clinical studies are being carried out in humans referred to by their stage of development: Phase 1 (single and multiple dose tolerance studies), Phase 2 (initial efficacy and pharmacokinetic studies) and Phase 3 (active controlled efficacy and long term safety studies) clinical trials. The drug product that patients are dosed with during these three

---

[1]A division of NMEs into large and small molecules is commonly found in the CMC literature. Small molecules are the classical drug mainly chemically synthesized. They are easily manufactured as tablets or capsules. Biological NMEs such as monoclonal antibodies are proteins that are either identical to or closely match endogenous proteins that play a key role in a disease process. They are manufactured as an injection or an infusion typically. Although small molecule drugs still make up 90 % of the drugs on the market, large molecule biological products are becoming increasingly more important.

phases of clinical studies are manufactured at small scale sizes following Good Manufacturing Practices (GMP), a set of government mandated rules governing the design, monitoring and control of pharmaceutical products to assure quality and enforced through the FDA. The analytical testing of the drug products are required to follow Good Laboratory Practices (GLP) rules to assure identity and support safety and quality. The latter is required under Code of Federal Regulations Title 21 (FDA 2015c). Both of these requirements point to the heavy regulation under which pharmaceutical products are manufactured and tested. It is impossible to be a CMC statistician and not appreciate the role of government regulation in the manufacture and testing of drug products, arguably the most heavily regulated industry in the world. There are corresponding regulatory rules which apply to the conduct of the clinical studies as well. Regulatory aspects of drug manufacture are discussed in greater detail in Chaps. 2 and 19.

The question of quality of a drug product falls squarely within the purview of CMC studies. What exactly does quality mean in this context? What is the role of Statistics in defining quality? What statistical tools are appropriate to answer these questions? In varying degrees of granularity, each of the chapters in this section of the book discuss statistical tools intended to address some important topic in the CMC arena: assay development, quality control, experimental design especially in the context of Quality by Design (QbD), process validation, statistical process and quality control, acceptance testing, stability modeling, dissolution, content uniformity, chemometrics and comparability studies. What they share in common is that in the background, there is a specification driving the need for the specific statistical tools discussed. It is beyond the scope of this chapter to discuss specifications and how they are derived, there is greater detail on this in Chaps. 18, 19 and 20, but one can say that conformance to specifications can be understood as the fundamental requirement for assuring the identity and quality of the drug product. The conformance is both a regulatory as well as a commercial requirement.

The CMC scientific studies and the statistical tools employed to design and elucidate the results of the CMC studies are intended to help define quality of the drug product through control of important drug product characteristics known as critical quality attributes (CQA). The study of these CQAs through appropriate statistical modeling leads to a greater scientific understanding of the product and process, as envisioned by the Quality by Design (QbD) paradigm set forth in the International Congress of Harmonization (ICH) guidances: Q8(R2) Pharmaceutical Development (2009), Q9 Quality Risk Management (2006) and Q10 Pharmaceutical Quality System (2009).

Modern statistical modeling approaches exploiting a Bayesian perspective have established a clear basis to answer the fundamental question of risk control as a probability statement in relation to process manufacturing factors (critical process parameters) and formulation changes linked to CQA outcomes (Peterson 2008). A set of CQA boundaries would be defined referred to as *design space* within which a manufacturer could move without having to file a regulatory supplement

seeking formal permission to modify the process. These concepts are completely consistent with the modernization of the pharmaceutical industry called for by the cGMPs for the twenty-first century initiative laid out in 2004 and later regulatory guidances (FDA 2015a, b). More on this topic is discussed in Chap. 18.

The follow-up to QbD and Design Space is process validation as elaborated on in great detail in the 2011 FDA guidance (FDA 2011). Two components of a formal process validation protocol are referred to as Stage 2-Process Performance Qualification (PPQ) and Stage 3-Continuous Process Verification (CPV). The notion of a life cycle approach to the manufacture of drug products is the intent of the 2011 guidance and proposes a more extensive role for statistical justification of moving from one stage of process validation to the next. An extensive discussion on the technical steps involved in Process Validation is given in Chap. 19.

Notions of quality in relation to pharmaceutical products are being increasingly related to "fit for use" concepts first elaborated on by Juran (1988). This calls for clear linkages of quality attributes to customer requirements, or ultimately in the context of pharmaceutical products, patient safety and efficacy. Much discussion has been held on establishing these linkages, especially in a QbD context, but so far, the necessary steps to provide such a clear connection have remained elusive. "Clinically relevant specifications" is a common topic of industry –regulatory agency conferences (Marroum 2012; Sharp 2012; Lostritto 2014) but exactly how we establish these in relation to CMC product attributes is still a work in progress.

One essential question relating to quality of a drug product concerns the amount of API being delivered to the patient. One might ask how close do individual dosage units deliver API to the stated amount. This is referred to as a content uniformity question. It is discussed in detail in Chap. 24. Analytical methods are developed and validated in connection with this question. Method development and validation are discussed in Chaps. 16 and 17. For small molecules, HPLC methods for potency assessment are typical, and for large molecules, a cell based bioassay would be developed. In both cases, an experimental design such as a Gauge R&R (also referred to as a "Gage" R&R) study could be used to study the total variability present in the measurement system. This is an important effort since the uncertainty associated with any analytical determination is a critical consideration (see ASTM E2655-14 2014) and has consequences on inferential statements related to drug product quality. In the case of a bioassay, average potency of a batch of biological product would be estimated using standard modeling methods relying on dose response curves and comparing the test sample to a standard curve. The use of a Gauge R&R design or a suitable block design (Altan and Shoung 2008) including analyst, run and plates and possibly other fixed and random effects provides a coherent approach to calculating estimates of the uncertainty associated with a reportable value based on 1 or more analytical determinations. An understanding of the uncertainty associated with such reportable values is an important consideration in both proposing and assessing drug product specifications.

The new Office of Pharmaceutical Quality (OPQ) established in 2014 has been charged with the responsibility of establishing quality metrics by which companies can evaluate their degree of "quality culture". The thinking is that notions of quality

have to imbue a company from top down and every employee engaged in the manufacture of a drug product has to regard their job as in some way related to the furtherance of quality. The notion of quality metrics as a regulatory and commercial enterprise is still in its infancy, with considerable ongoing discussions being held at various joint industry FDA conferences (International Society of Professional Engineers 2015). It's clear that this major initiative by the FDA will evolve over the coming years and entail considerable accommodation by industry to pursue heightened standards in manufacturing quality. Readers should consult the current regulatory and scientific literature to stay abreast of its developments.

In the following sections of this chapter, we turn to a selection of statistical tools by way of introduction to their application to CMC studies, and pave the way for subsequent chapters to elaborate in greater detail on the topics opened up here.

15.2 Overview of Traditional CMC Statistical Methods

The traditional collection of statistical methods used to support drug product development and manufacturing are essentially the same as what might be called "industrial statistics". These involve statistical methods for process optimization, process validation, process monitoring, acceptance sampling, and manufacturing risk assessment.

While most of the statistical methods we touch on in this section have been around for decades, introduction of the concept of ICH Q8 "design space" and more modern manufacturing practices has created a need for statistical methods with better multivariate predictive capability. The concept of ICH Q8 "design space" has generated a need for improved assessment of the probability of conformance to specifications, particularly for processes with multiple responses. The introduction of modern pharmaceutical manufacturing and process monitoring technologies has produced a greater need to be able to handle large data sets comprised of many correlated measurements. An overview of the statistical methods that can address these two needs are given in Sects. 15.2.2 and 15.2.3, respectively. Subsequent Sects. 15.2.4–15.2.9 touch on important topics related to traditional categories of industrial statistical methods that have found good use in pharmaceutical development, manufacturing and testing. These sections serve as an introduction to subsequent chapters of the CMC section of this book which expand in much greater detail the topics touched on here.

15.2.1 Factorial and Fractional Factorial Designs

Often a pharmaceutical process may have many controllable factors that could *potentially* contribute to manipulation that will lead to process improvement. However, it is often the case that some factors will have a small or negligible effect

over the range of experimental conditions to be considered. It is sometimes the case that only a small subset of factors acting together will produce substantial process improvement, but we may not know which these are among the many potential factors. As such, it is prudent to begin an experimental campaign with an efficient factor screening design to determine, in part, the key factors for use in process optimization techniques, such as response surface methodology. These factor screening experiments need to be efficient so that there are sufficient resources remaining for process optimization, once the critical factors are identified. In fact, some industrial statisticians recommend that no more than 25 % of available experimental resources should be used for the first set of experiments (Montgomery 2009, p. 556).

Initially, experimenters will want to make a list of potential factors that could affect a process quality attribute (or attributes), and then perform an experiment (or experiments) to screen out those factors that have little or no effect upon the process. If the number of potential factors is not too large (2 to 4 say) an experimenter may consider doing a full factorial design, provided he or she has the required resources. For a screening design, it is often best to use only 2 (or possibly 3) levels for each factor. A popular factorial screening design is the 2^k design. Here, each factor level is assigned only two levels. For k factors, the design has 2^k combinations of runs that must be executed. A full factorial design in k factors can model all linear and interaction terms up to the one k-way interaction. Often though, only linear and pairwise interactions are modeled, unless there is prior concern about the presence of a higher-order interaction.

The factorial design analysis typically involves doing a preliminary fit of the full model and then checking model assumptions. This will involve a residual analysis to check for possible outliers and normality of the residuals. Statistically insignificant factors should be considered for removal from the model, but care must be taken, particularly if the residuals show large variability in the data. Low power to detect practically significant factors could then be a concern. Typical regression model selection criteria (such as AIC or BIC), can also be used to discard unimportant factors. Of course, practical scientific background knowledge also plays a part in such decisions.

Often, the experimenter will also perform one or more "center point" runs. These are runs executed at the center point of all factors that are quantitative. If all the factors are quantitative, there would be only one center point whose coordinates are at the center of each factor (halfway between the factor's lower and upper limits). Factor ranges are an important consideration in sufficiently characterizing the linear or quadratic surface. A design with three quantitative factors and one qualitative factor (e.g. catalyst type) might have center points, one at each level of the qualitative factors. The purpose of the center point runs are to provide information about possible lack-of-fit from the type of model that is fit for the factorial design. Such designs typically support models with only first order and pairwise interactions terms such as

$$y = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{i<j} \beta_{ij} x_i x_j + e. \tag{15.1}$$

If the investigator is concerned about some kind of departure from the model in (15.1), the center points can provide a test for such a departure, although it provides little information as to the exact type of departure (del Castillo 2007, pp. 411–412).

As the number of factors to be studied increases, full factorial designs quickly become too costly. If we are willing to assume that higher order interactions are negligible, then one should consider fractional factorial designs. Negligible higher order interactions are associated with a certain degree of smoothness on the underlying response surface. If we believe that the underlying process is not very volatile as we change the factors across the experimental region, then the assumption of negligible higher order interactions may be reasonable. Fractional factorial designs are often certain fractions of a 2^k factorial design. They have the design form 2^{k-p}, where k is the number of factors and p is associated with the particular fraction of the 2^k design. For example, a 2^{5-1} design denotes a half-replicate of a 2^5 design, while 2^{6-2} denotes a one-fourth replicate of a 2^6 design.

For fractional factorial designs, some interaction terms are confounded (or aliased) with possibly first order or other interaction terms. This means that certain interaction terms cannot be estimated apart from other terms. If (negligible) higher order interactions terms are confounded with first order or lower order interaction terms, then we may still be able to make practical inferences from a fractional factorial design analysis. Fractional factorial designs can be categorized into how well they can "resolve" factor effects. These categories are called design resolution categories. For a resolution III design, no main effects (first order terms) are confounded with any other main effect, but main effects are aliased with two-factor interactions and two-factor interactions are aliased with each other. For a resolution IV design, no main effect is aliased with any other main effect or with any two-factor interaction, but two-factor interactions are aliased with each other. For a resolution V design, no main effect or two-factor interaction is aliased with any other main effect or two-factor interaction, but two-factor interactions are aliased with three-factor interactions. Resolution III and IV designs are particularly good for factor screening as they efficiently provide useful information about main effects and some information about two-factor interactions (Montgomery 2009, chapter 13).

Recently, a special class of screening design has gathered some popularity. It is called a "definitive screening design". It requires only twice the number of runs as there are factors in the experiment. The analysis of these designs is straightforward if only main effects or main and pure-quadratic effects are active. The analysis becomes more challenging if both two-factor interactions and pure-quadratic effects are active because these may be correlated (partially aliased) (Jones and Nachtsheim 2011). If a factor's effect is strongly curvilinear, a fractional factorial design may miss this effect and screen out that factor. If there are more than six factors, but three

or fewer are active, then the definitive screening design is capable of estimating a full quadratic response surface model in those few active factors. In this case, a follow-up experiment may not be needed for a response surface analysis.

15.2.2 Response Surfaces: Experimental Design and Optimization

Factorial and fractional factorial designs are useful for factor screening so that experimenters can identify the key factors that affect their process. It may be the case the results of a factorial or a fractional factorial design produce sufficient process improvement that the pharmaceutical scientist or chemical engineer will decide to terminate any further process improvement experiments. However, it is important to understand that a response surface follow up experiment may not only produce further process improvement, but it will also produce further process understanding by way of the response surface. If a process requires modification of factor settings (e.g. due to an uncontrollable change in another factor) it may not be clear how to make such a change without a response surface, particularly if the only experimental information available is from a fractional factorial design.

The results of the factor screening experiments may provide some indication that the process optimum may be within the region of the screening experiment. This could be the case, for example, if the response from center-point runs are better than all of the responses at the factorial design points. However, it is probably more likely that the results of the screening design point towards a process optimum that is towards the edge, or outside of, the factor screening experimental region. If this is the case, and if the results of the screening experiment provide no credible evidence of response surface curvature within the original region, then the experimenter should consider the method of "steepest ascent/descent" for moving further towards the process optimum. The classical method of steepest ascent/descent uses a linear surface approximation to move further towards the process optimum. Often, the original screening design may provide sufficient experimental points to support a linear surface. Clearly, statistical and practically significant factor interactions and/or a significant test for curvature indicates that a second-order response surface model needs to be built as the linear surface is not an adequate approximation. The method of steepest ascent/descent is a form of a "ridge analysis" (in this case for a linear surface), to be reviewed briefly below. See Myers et al. (2009, chapter 5) for further details on steepest ascent.

Special experimental designs exist for developing second-order response surface models. If it appears from the screening design that the process optimum may well be within the original experimental region, then a central composite design may be able to be developed by building upon some of the original screening design points (Myers et al. 2009, pp. 192–193). The Box–Behnken design (Box and Behnken 1960) is also an efficient design for building second-order response surfaces.

Once a second-order response surface model is developed, analytical and graphical techniques exist for exploring the surface to determine the nature of the "optimum". The basic form for the second-order response surface is

$$y = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{j=1}^{k} \sum_{i=1}^{k} \beta_{ij} x_i x_j + e, \tag{15.2}$$

where y is the response variable, $x_1, \ldots, x_k$ are the factors. It is often useful to express the equation in (15.2) in the vector–matrix form as

$$y = \beta_0 + \mathbf{x}' \boldsymbol{\beta} + \mathbf{x}' \mathbf{B} \mathbf{x} + e, \tag{15.3}$$

where $\mathbf{x} = (x_1, \ldots, x_k)'$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)'$, and $\mathbf{B}$ is a symmetric matrix whose ith diagonal element is β_{ii} and whose (i, j)th-off diagonal element is $\frac{1}{2}\beta_{ij}$. The form in (15.3) is useful in that the matrix $\mathbf{B}$ is helpful in determining shape characteristics of the quadratic surface. If $\mathbf{B}$ is positive definite (i.e. all eigenvalues are positive), then the surface is convex (upward), while if $\mathbf{B}$ is negative definite (i.e. all eigenvalues are negative), then the surface is concave (downward). If $\mathbf{B}$ is positive or negative definite, then the surface associated with (15.3) has a stationary point that is a global minimum or maximum, respectively. If, however, some of the eigenvalues of $\mathbf{B}$ change sign, then the surface in (15.3) is that of a "saddle surface", which has a stationary point, but no global optimum (maximum or minimum). The stationary point, $\mathbf{x}_0$, is that point at which the gradient vector of the response surface is stationary (i.e. all elements of the gradient vector are zero). It follows then that

$$\mathbf{x}_0 = -\frac{1}{2}\mathbf{B}^{-1}\boldsymbol{\beta}.$$

Further insight into the nature of a quadratic response surface can be assessed by doing a "canonical analysis" (Myers et al. 2009, chapter 6). A canonical analysis invokes a coordinate transformation that replaces the equation in (15.2) with

$$\widehat{y} = \widehat{y}_s + \sum_{i=1}^{k} \lambda_i w_i^2, \tag{15.4}$$

where $\hat{y}_s$ is the predicted response at the estimated stationary point, $\widehat{\mathbf{x}}_0 = -\frac{1}{2}\widehat{\mathbf{B}}^{-1}\widehat{\boldsymbol{\beta}}$, and $\lambda_1, \ldots, \lambda_k$ are the eigenvalues of $\widehat{\mathbf{B}}$. The variables, $w_1, \ldots, w_k$, are known as the canonical variables, where $\mathbf{w} = (w_1, \ldots, w_k)' = \mathbf{P}'(\mathbf{x} - \widehat{\mathbf{x}}_0)$ and $\mathbf{P}$ is such that $\mathbf{P}'\widehat{\mathbf{B}}\mathbf{P} = \boldsymbol{\Lambda}$ and $\boldsymbol{\Lambda} = diag(\lambda_1, \ldots, \lambda_k)$. Here, $\mathbf{P}$ is a $k \times k$ matrix of normalized eigenvectors associated with $\widehat{\mathbf{B}}$.

One can see from (15.4) that if $|\lambda_i|$ is small, then moving in the direction vector given by $(0, .., 0, \lambda_i, 0, \ldots, 0)'$ (in the $\mathbf{w}$-space) will result in little change from the

stationary point. Such movements can be useful if, for example, process cost can be reduced for conditions somewhat away from the stationary point, but on or close to the line intersecting $\widehat{x}_0$ along the $(0, .., 0, \lambda_i, 0, \ldots, 0)'$ direction in the w-space.

However, the response surface shape may indicate that a global optimum is not within the regions covered by experimentation thus far. Or, it may be that operation at the estimated stationary point, $\widehat{x}_0$, is considered too costly. In such cases, it may be useful to conduct a "ridge analysis". This is a situation where we want to do constrained optimization, staying basically within the experimental region. For a classical ridge analysis (Hoerl 1964), we optimize the quadratic response surface on spheres (of varying radii) centered at the origin of the response surface design region. In fact, one can consider ridge analysis to be the second-order response surface analogue of the steepest ascent/descent method for linear response surfaces. Draper (1963) provides an algorithmic procedure for producing a ridge analysis for estimated quadratic response surfaces. Peterson (1993) generalizes the ridge analysis approach to account for the model parameter uncertainty and to also make it apply to a wider class of models which are only linear in the model parameters, e.g. many mixture experiment models (Cornell 2002).

15.2.3   Ruggedness and Robustness

Once an analytical assay, or other type of process, has been optimized, one may want to assess the sensitivity of the assay or process to minor departures from the set process conditions. Here, we call this analysis a "ruggedness" assessment. Some of these conditions may involve factors for which the process is known to be sensitive, while other conditions may involve factors that (through testing or process knowledge) are known to be less influential. In a ruggedness assessment, the experimenter has to keep in mind how much control is possible for each factor. If one has tight control on all of the sensitive factors, then the process may indeed be rugged against deviations that commonly occur in manufacturing or in practical use (e.g. as with a validation assay). Ruggedness evaluations are often done on important assays that are used quite frequently.

Much of the classical literature on ruggedness evaluations for assays involve performing screening designs with many factors over carefully chosen (typically small) ranges. See, for example, Vander Heyden et al. (2001). The purpose of such an experiment is to see if any factor effects are statistically (and practically) significant. However, such experiments are typically not designed to have pre-specified power to detect certain effects with high probability. Furthermore, a typical factorial analysis of variance does not capture the probability that future responses from a process over this ruggedness experimental region will be within specifications. To address this issue, Peterson and Yahyah (2009) apply a Bayesian predictive approach to quantify the maximum and minimum probabilities that a future response will meet specifications within the ruggedness experimental region.

If the maximum probability is too small, the process is not considered rugged. However, if the minimum probability is sufficiently large, then the process is considered rugged.

This problem of process sensitivity to certain factors can sometimes be addressed by moving the process set conditions to a point that, while sub-optimal, produces a process that is more robust to minor deviations from the set point. This can sometimes be achieved by exploiting factor interactions, or from an analysis of a response surface. In some situations, however, the process is sensitive to factors which are noisy. This may be particularly true in manufacturing where process factors cannot be controlled as accurately as on the laboratory scale.

When a process will be ultimately subject to noisy factors (often called noise variables), a robustness analysis can be employed known as "robust parameter design" (Myers et al. 2009, chapter 10). Robust parameter design has found productive applications in industries involving automobile, processed food, detergent, and computer chip manufacturing. However, it has to date, not been widely employed in the pharmaceutical industry, although applications to the pharmaceutical industry are starting to appear in the manufacturing literature (Cho and Shin 2012; Shin et al. 2014).

The basic idea behind robust parameter design is that the process at manufacturing scale has both noise factors and controllable factors. For example, noise factors might be temperature (deviation from set point) and moisture, while controllable factors might be processing time and a set point associated with temperature. If any of the noise variables interact with the one or more of the controllable factors, then it may be possible to reduce the transmission of variation produced by the noise variables. This may result in a reduction in variation about a process target, which in turn will increase the probability of meeting specification for that process. There is quite a large literature on statistical methods associated with robust parameter design involving a univariate process response. (See Myers et al. 2009, chapter 10 for references.) However, there are fewer articles on robust parameter design method for multiple response processes (e.g. Miró-Quesada et al. 2004). Miró-Quesada et al. (2004) introduce a Bayesian predictive approach that is widely applicable to both single and multiple response robust parameter design problems.

15.2.4 Process Capability

During or directly following process optimization, the experimenter should also consider some aspect of process capability analysis. Such an analysis involves assessing the distribution of the process response over its specification interval (or region) and the probability of meeting specifications. Assessment of process capability involves a joint assessment of both the process mean and variance, as well as the variation of the process responses about the quality target, which in turn is related to the probability of meeting specifications. Further process capability should also be assessed during pilot plant and manufacturing as part of the process

monitoring activities. This is because the distribution of process responses may involve previously unforeseen temporal effects associated with sequential trends, the day of the week, etc. Process capability is clearly important because a process may be optimized, but it may not be "capable", i.e. it may have an unacceptable probability of meeting specifications.

Process capability indices have become popular as a way to succinctly quantify process capability. The C_p index has the form

$$\frac{USL - LSL}{6\sigma},$$

where USL = "upper specification limit", LSL = "lower specification limit", and σ equals the process standard deviation. The C_p index is estimated by substituting the estimated standard deviation, $\widehat{\sigma}$, for σ. As one can clearly see, the larger the C_p index, the better is the process capability. However, the C_p index does not take into account where the process mean is located relative the specification limits. A more sensitive process capability index is denoted as C_{pk} which has the form

$$\min\left(\frac{USL - \mu}{3\sigma}, \frac{\mu - LSL}{3\sigma}\right).$$

The C_{pk} index is estimated by substituting the estimates for μ and σ for their respective population parameters appearing in the C_{pk} definition. The magnitude of C_p relative to C_{pk} is a measure of how off center a process is relative to its target.

Statistical inference relative to process capability indices require critical care. Such index estimates may have rather wide sampling variability. In addition, the process capability index can be misleading if the process is not in control (i.e. stable and predictable). It is hoped that before or during the process optimization phase that a process can be brought into control. However, validation of such control may need to be confirmed during the actual running of the process over time using statistical process monitoring techniques (Montgomery 2009, p. 364). In addition, process capability indices have in the past received criticism for trying to represent a multifaceted idea (process capability) as one single number (Nelson 1992; Kotz and Johnson 2002). While this criticism has some merit (and is applicable to any other single statistic), it is only valid if such indices are reported to the exclusion of other aspects of the distribution of process responses.

15.2.5 Measurement System Capability

Measurement capability is an important aspect of any quality system. If measurements are poor (e.g. very noisy and/or biased) process improvement may be slow and difficult. As such, it is important to be able to assess the capability of a

measurement system, and improve it if necessary. This section will review concepts and methods for measurement system capability assessment and improvement.

Two important concepts in measurement capability analysis are "repeatability" and "reproducibility". Repeatability is the variation associated with repeated measures on the same unit under *identical* conditions, while reproducibility is the variation associated when units are measures under *different* natural process conditions, such as different operators, time periods, etc. A measurement system with good capability is able to easily distinguish between good and bad units.

A simple model for measurement systems analysis (MSA) is

$$Y = T + e,$$

where Y is the observed measurement on the system, T is the true value of the system response (for a single unit, e.g. a batch or tablet), and e is the error difference. (It is assumed here that Y, T and e are stochastically independent.) In MSA the total variance of this system is typically represented by

$$\sigma^2_{\text{Total}} = \sigma^2_{\text{Process}} + \sigma^2_{\text{Gauge}},$$

where $\sigma^2_{\text{Total}} = \text{var}(Y)$, $\sigma^2_{\text{Process}} = \text{var}(T)$, and $\sigma^2_{\text{Gauge}} = \text{var}(e)$. Clearly, accurate measuring devices are associated with a small "gauge" variance. It is also possible to (additively) decompose the gauge variance into two natural variance components, $\sigma^2_{\text{Repeatability}}$ and $\sigma^2_{\text{Reproducibility}}$. Here, "reproducibility" is the variability due to different process conditions (e.g. different operators, time periods, or environments), while "repeatability" is the variation due to the gauge (i.e. measuring device) itself. The experiment used to measure the components of σ^2_{Gauge} is typically called a "gauge R&R" study.

The estimation of variance components associated with a gauge R&R study are applied not only to pharmaceutical manufacturing process, but also to important assays. If a process or assay shows unacceptable variation, a careful variance components analysis may help to uncover the key source (or sources) of variation responsible for poor process or assay performance.

15.2.6   Statistical Process Control

In pharmaceutical manufacturing or in routine assay utilization, such processes will tend to drift over time and will eventually fall out of a state of control. All CMC processes have some natural (noise) variation to which they are subject. If this source of underlying variation is natural and historically known, this is typically called "common cause" variation. Such variation is "stationary" in the sense that it is random and does not drift or change abruptly over time. A process that is in control will be subject only to common cause variation. However, other types of variation

will eventually creep in and affect a process. These sources of variation are known as "special cause" variation. Typical sources of special cause variation are: machine error, operator error, or contaminated raw materials. Special cause variation is often large when compared to common cause variation. Special cause variation can take on many forms. For example, it may appear as a one-time large deviation or as a gradual trend that eventually pushes a process out of control. Statistical process control (SPC) is a methodology for timely detection of special cause variation and more generally for obtaining a better understanding of process variation, particularly with regard to temporal effects. See Chap. 20 for additional discussion on this and related topics.

As a practical matter, it is important to remember that SPC only provides detection of special cause variation. Operator or engineering action will be needed to eliminate these special causes of variation, so that the process can be brought back into a state of control. Identification of an assignable cause for the special cause variation may require further statistical analysis or experimentation (e.g. a variance components analysis).

An SPC chart is used to monitor a process so that efficient detection of special cause variation can be obtained. An SPC chart consists of a "center line" that represents the mean of the quality statistic being measured when the process is in a state of control. The quality statistic is typically not just one single measured process response but rather the average of a group of the same quality responses chosen within a close time frame. It is expected that the common cause variation will induce random variation of this statistic about the center line. The SPC chart also has upper and lower control limits. These two control limits are chosen (with respect to the amount of common cause variation) so that nearly all of the SPC statistic values over time will fall between them. If the statistic values being measured over time vary randomly about the center line and stay within the control limits, then the process is in state of control and no action is necessary. In fact, any action to try to improve a process that is subject only to common cause variation may only increase the variation of that process. If however the SPC statistic values start to fall outside of the control limits or behave in a systematic or nonrandom manner about the center line, then this suggests that the process is out of control.

SPC methodology involves a variety of statistical tools for developing a control chart to meet the needs of the process and the manufacturer. Specification of the control limits is one of the most important decisions to be made in creating a control chart. Such specification should be made relative to the distribution of the quality statistic being measured, e.g. the sample mean. If the quality statistic being measured is a sample mean, then it is common practice to place the control limits at an estimated "3 sigma" distance from the center line. Here, sigma refers to the standard deviation of the distribution of the sample mean, not the population of individual quality responses. Three-sigma control limits have historically performed well in practice for many industries.

Another critical choice in control chart development involves collection of data in "rational subgroups". The strategy for the selection of rational subgroups should be such that data will be sampled in subgroups, so that if special causes are present,

the chance for differences between subgroups to appear will be maximized, while the chance for differences within a subgroup will be minimized. The strategy for one's rational subgroup definition is very important and may depend upon one or more aspects of the process. How the rational subgroups are defined may affect the detection properties of the SPC chart. For example sampling several measurements at a specific discrete time points throughout the day will help the SPC chart to detect monotone shifts in the process. However, randomly sampling all process output across a sampling interval will result in a different rational group strategy, which may be better at detecting process shifts that go out-of-control and then back in between prespecified time points.

In addition to rational subgroups, it is important to pay attention to various patterns on the control chart. A control chart may indicate an out-of-control situation when one or more points lie outside of the control limits or when a nonrandom pattern of points appears. Such a pattern may or may not be monotone within a given run of points. The problem of pattern recognition associated with an out-of-control process requires both the use of statistical tools and knowledge about the process. A general statistical tool to analyze possible patterns of non-randomness is the runs test (Kenett and Zacks 2014, chapter 9).

As one might expect, having multiple rules for detecting out-of-control trends or patterns can lead to an increase in false alarms, particularly if a process is assumed to be out-of-control if at least one, out of many such rules, provides such indication. It may be possible to adjust the false positive rate for the simultaneous use of such rules, but this may be difficult as many such rules are not statistically independent. For example, one rule may be "one or more points outside of the control limits" while another rule may be "six points in a row with a monotone trend". This is a situation where computer simulation can help to provide some insights regarding the probabilities of false alarms when using multiple rules for detecting out-of-control trends or patterns.

Typical control chart development is divided into two phases, Phase I and Phase II. The purpose of Phase I is to develop the center line and control limits for the chart, as well as to ascertain if the process is in control when operating in the sequential setting as intended. Phase I may also be a time when the process requires further tweaking to be brought into a state of control. The classical Shewhart control charts are generally very effective for Phase I because they are easy to develop and interpret, and are effective for detecting large changes or prolonged shifts in the process. In phase II, the process is now assumed to be reasonably stable so that phase II is primarily a phase of process monitoring. Here, we expect more subtle changes in the process over time, and so more refined SPC charts may be employed such as cumulative sum and Exponentially Weighted Moving Average (EMWA) control charts (Montgomery 2009, chapter 9).

The notion of "average run length" (ARL) is a good measure for evaluation a process in Phase II. The ARL associated with an SPC chart or an SPC method is the expected number of points that must be plotted before detecting an out-of-control situation. For the classical Shewhart control chart, the $ARL = 1/p$, where p is the probability that any point exceeds the control limits. This probability, p, can often

be increased by taking a larger sample at each observation point. Another useful, and related measure, is the "average time to signal" (ATS). If samples are taken that are t hours apart, the ATS $=$ ARLt. We can always reduce the ATS by increasing the process sampling frequency. In addition, the ARL and ATS can be improved by judicious use of more refined control chart methodology (in some cases through cumulative sum or EWMA control charts). For some control charts (e.g. Shewhart) the run time distribution is skewed so that the mean of the distribution may not be a good measure. As such, some analysts prefer to report percentiles of the run length distribution instead.

Many processes involve multiple quality measurements. As such, there are many situations where one needs to monitor multiple quality characteristics simultaneously. However, statistical process monitoring to detect out-of-control processes or trends away from target control requires special care, and possibly multiple SPC charts of different types. The naive use of a standard SPC chart for each of several measured quality characteristics can lead to false alarms as well as missing out-of-control process responses. False alarms can happen more often than expected with a single SPC chart because the probability of *at least* one false alarm out of several can be noticeably greater than the false alarm rate on an individual SPC chart. In addition, multivariate outliers can be missed with the use of only individual control charts. Because of this, special process monitoring methods have been developed for multivariate process monitoring and control.

One of the first methods to address multivariate process monitoring is the Hotelling T^2 control chart. This chart involves plotting a version of the Hotelling T^2 statistic against an upper control limit chi-square or F critical value. The Hotelling T^2 statistic can be modified to address either the individual or grouped data situation. However, an out-of-control signal by the Hotelling T^2 statistic does not provide any indication of what particular quality response or responses are responsible. In addition to SPC charts for individual quality responses, one can plot the statistics, $d_j = T^2 - T^2_{-j}$, where T^2_{-j} is the value of the T^2 statistic but with the jth quality response omitted.

Several other univariate control chart procedures have been generalized to the multivariate setting. As stated above, EWMA charts were developed to better detect small changes in the process mean (for a single quality response type). Analogously, multivariate EWMA chart methods have been developed to detect small shifts in a mean vector. In addition, some procedures have been developed to monitor the multivariate process variation by using statistics which are functions of the sample variance-covariance matrix. See Montgomery (2009, pp 516–517) for details.

When the number of quality responses starts to become large (more than 10 say), standard multivariate control chart methods start to become less efficient in that the ARL also increases. This is because any shifts in one or two response types become diluted in the large space of all of the quality responses. In such cases, it may be helpful to try to reduce the dimensionality if the problem by projecting the high-dimensional data into a lower dimensional subspace. One approach is to use principal components. For example, if the first two principal components account for a large proportion of the variation, one can plot the principal component scores

labeled by their order within the process as given by the recorded data vectors. This is called a trajectory plot. Another approach is to collect the first few important principal components and then apply the multivariate EWMA chart approach to them (Scranton et al. 1996).

15.2.7 *Acceptance Sampling*

When lots of raw materials arrive at a manufacturing plant, it is typical to inspect such lots for defects or some measure related to raw material quality. In addition, pharmaceutical companies often inspect newly manufactured lots of product before making a decision about whether or not to release the product lot or patch for further processing or public consumption.

However, it is useful to note that the primary purpose of acceptance sampling is to sentence lots as acceptable or not; it is not to create a formal estimate of lot quality. In fact, most acceptance sampling procedures are not designed for estimation of lot quality Montgomery (2009). Acceptance sampling should not be a substitute for process monitoring and control. Nonetheless, the use of acceptance sampling plans over time produces a history of information which reflects on the quality of the process producing the lots or batches. In addition, this may provide motivation for process improvement work if too many lots or batches are rejected. See Chap. 20 for additional discussions on this and related ideas.

Acceptance sampling can be divided into two categories related to item description, attribute plans and variable plans. Attributes are quality characteristics that have discrete "accept" or "reject" levels (e.g. "defective" or "acceptable"). Variables sampling plans involve a quality characteristic that is measured on a continuous scale. Acceptance sampling plans can also be categorized according to their sequential nature. For example, a "single sampling" plan takes one sample of n units from a lot and a decision is made based on that one sample to sentence the lot as acceptable or not. A "double sampling" plan works in two stages. A first sample is taken and a decision is made to (i) accept the lot, (ii) reject the lot, or (iii) take a second sample from the lot. If the second sample is taken, the information from both the first and second sample is used to accept or reject the lot. It is also possible to have multi-stage sampling plans that are generalizations of a two-stage sampling plan, whereby more than two samples from the lot are required to make a decision. It should be noted that acceptance sampling plans (aside from 100 %) inspection usually always involve some sort of random sampling.

A useful characterization of the performance of an acceptance sampling plan is the operating-characteristic (OC) curve. This curve is a plot of the "probability of accepting the lot" vs. the "lot fraction defective". Clearly, this should be a monotone decreasing curve. The location and shape of this curve displays the discriminatory power of the sampling plan. To better understand how this works, consider the attributes sampling situation where the lot size, N, is very large so that for a random sample of size n ($n \ll N$) the number of defectives has approximately a binomial

Fig. 15.1 Operating
characteristic (OC) curve for
a sample size of $n = 50$ and
$d^* = 2$ acceptance number
for defectives
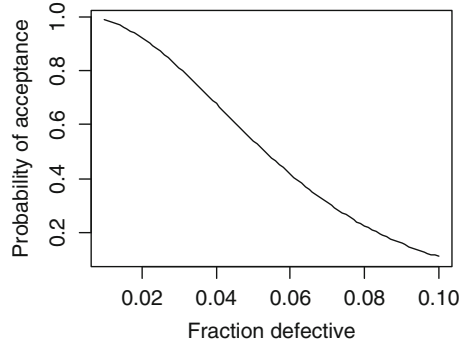
distribution. We assume here that π is the fraction of defective items in the lot. If
D is the number of defective items found out of a sample of size n, the D has a
binomial distribution with probability function,

$$\Pr\left(D = d; n, \pi\right) = \binom{n}{d} \pi^d (1 - \pi)^{n-d}.$$

Suppose that we accept the lot if $D \leq d^*$. We call d^* the acceptance number. Then
the probability of accepting the lot is

$$p\left(\pi\right) = \Pr\left(D \leq d^*; n, \pi\right) = \sum_{d=0}^{d^*} \binom{n}{d} \pi^d (1 - \pi)^{n-d}.$$

For specific values of n and d^*, one can plot $p(\pi)$ vs. π to obtain the OC curve. For
example if $n = 50$ and $d^* = 2$, then the OC curve is shown in Fig. 15.1 below.

If the sample size, n, increases (keeping d^* proportional to n), then the slope (in
the neighborhood of the curve inflection) will become steeper. This indicates more
discriminatory power for the sampling plan.

Often the statistician or quality engineer will focus on certain points on the
OC curve. For example, they may be interested knowing what level of lot quality
(fraction defective) would be associated with a high probability of acceptance
(e.g. 0.95). Or, they (and consumers) may be interested in the level of lot quality
associated with a low probability of acceptance. A sampling plan is often established
with regard to an acceptable quality level (AQL). The AQL is the lowest level
of quality for the manufacturer's process that would be consider acceptable as a
process average. The sampling plan will typically be designed so that the OC curve
shows a high probability of acceptance at the AQL. It is important to note that
the AQL is not usually intended to be a specification on the product Montgomery
(2009). It is simply a standard against which to sentence lots. See Montgomery
(2009) for an overview of OC curves and AQL's associated with double or multiple
sampling plans.

With regard to variable sampling plans, there are advantages and disadvantages of which one should be aware. The main advantage of the variables sampling plan is that essentially the same OC curve can be obtained, but with a smaller sample size, than for the corresponding attributes sampling plan. This is because there is more information in variables sampling than in attribute sampling. However, for a variables sampling plan, the probability distribution of the quality characteristic being used must be known, at least up to a good approximation. If the true distribution of the quality characteristic for the variables sampling plan deviates enough from the assumed distribution, then serious departures of the computed OC from the real underlying one may arise, leading to rather biased decisions about lot sentencing. A typical assumption is that the quality characteristic for a variables sampling plan has a normal distribution. But, this assumption should be checked carefully. See Montgomery (2009) for an overview of variables sampling plans.

15.2.8   Failure Mode and Reliability Assessment

Process and product reliability assessment involve two basic activities: (i) identification of failure modes and (ii) quantification of reliabilities associated with these failure modes. Identification of all key failure modes is important. "Failure modes" means the ways, or modes, in which something might fail. Their identification is important. Suppose all but one key failure model is identified, and much effort has gone into quantifying the reliabilities of the (identified) potential modes of failure. Nonetheless, a process can still be likely to produce a seriously flawed product because one of the key failure models was not identified.

A popular approach to identification of failure modes and their subsequent reliability quantification is called "failure mode and effects analysis" (FMEA). Teamwork involving key representatives related to different aspects of a manufacturing process is important for executing an FMEA (Breyfogle 2003). As part of the FMEA process, a team works to identify potential failure modes for design functions or process requirements. The teams then assign a severity measure to each identified failure mode. They also assign a measure of the frequency of occurrence to the failure mode, along with a measure of the likelihood of detection. After this analysis, the team typically calculates a "risk priority number" (RPN), which is the product of three numbers: the severity measure, the frequency measure, and the likelihood of detection measure. Typically, the RPN values are used to prioritize process improvement efforts, with teams tackling the potential failure models with the highest RPN's. However, RPN's have been criticized as sometimes misleading because they do not strictly follow the rules of arithmetic (Wheeler 2011).

The problem with RPN's is that the severity measure, the frequency measure, and the likelihood of detection measure are based upon ordinal measures, i.e. they are based upon ordered rankings. For example, the severity score may have levels such as: "no discernible effect", "very minor", "minor", "very low, "low", "moderate", "severe', "very severe". The frequency-of- failure score may have levels such as:

"unlikely", "relatively few failures", "occasional failures", "frequent failures", and "persistent failures". Likewise, the likelihood-of-detection score may have levels such as: "almost certain", "very high", "high", "moderately high", "moderate", "low", "very low", "remote", "cannot detect failure". However, multiplication of these ordinal values is improper because the measures are not on an "interval ratio scale". This does not mean that the failure modes identification process is flawed; it only means that the RPN's can be misleading. If ordinal scales must be used, it is better to prioritize the situations first by severity, then by occurrence within each level of severity, and finally by detectability within each combination of severity and occurrence (Wheeler 2011). More formally, it is better to model the reliability of a process by using the identified failure modes and the laws of probability. This can be done using a Bayesian network (Garcìa and Gilabert 2011).

The quantification of reliability probabilities for a process can be a complex and/or fragile undertaking. By a "reliability probability" we simply mean the probability of desired event (e.g. meeting one or more process specifications). For a complex, multi-stage process the quantification of the overall reliability probability can involve sophisticated modeling of conditional and unconditional distributions. See for example Barlow and Proschan (1975). A more subtle point is that even if sophisticated probability analyses are completed, the actual process reliability and the estimated one may in some situations be rather different. If a reliability analysis is based upon carefully designed experiments, providing adequate information, some probability-based reliability calculations may be fairly accurate. However, for practical reasons, probability based reliability calculations may be conditional on the properties of raw materials from a specific supplier or upon the functioning of a specific piece of equipment. If either or both of these change, the reliability probability may change abruptly. Hence, reliability probabilities are at best only good for the near future, and require re-validation from time to time.

15.2.9   Experimental Design and Modeling Considerations in Bioassay Potency Testing

As mentioned previously, experimental design considerations and modeling strategies are a critical part of potency testing carried out through bioassays, frequently using in vitro or cell based bioassays using microtiter plates or in some cases, in vivo or live animal bioassays (Finney 1978). The balanced incomplete block design (BIBD) given in Table 15.1 is an example of a statistical design which could be used to estimate variance components attributable to analyst, run and plate for the given bioassay. One standard microtiter plate consists of wells located across 8 rows and 12 columns, 96 wells total, with each well providing a measurement in relation to a known or unknown concentration of a biologic material. The layout of a 96 well plate with rows and columns marked as A–H and 1–12 respectively is given in Fig. 15.2.

Table 15.1 BIBD design for a Gauge R&R study of a biologic

Analyst	Run	Plate	Sample concentration (%)					
			25	50	67	100	150	200
1	1	1	X	X	X			
		2	X				X	X
	2	3			X	X	X	
		4		X	X			X
	3	5	X	X		X		
		6		X			X	X
	4	7			X	X		X
		8	X			X		X
	5	9		X		X	X	
		10	X		X		X	
2	1	1				X	X	X
		2		X	X	X		
	2	3	X	X				X
		4	X			X	X	
	3	5			X		X	X
		6	X		X	X		
	4	7	X	X			X	
		8		X	X		X	
	5	9	X		X			X
		10		X		X		X

Fig. 15.2 96-Well plate layout
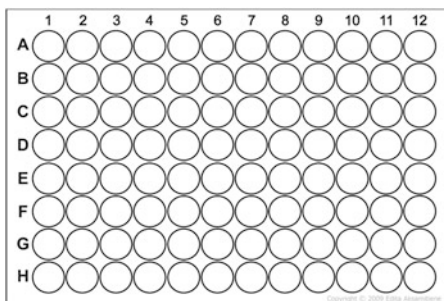
The given BIBD also permitted the investigation of dilutional linearity. Dilutional linearity relates to the extent of recovery across a range of known concentrations. For example, known samples of 25 %, 50 %, and so on could be plated out and tested to see if the measured concentrations are reasonably close to their known concentrations following the scheme laid out in Table 15.1. This is an example of two analysts, each analyst performing five runs and each run consisting of three plates, each plate accommodating three samples (known concentrations) plus a standard. Each plate would consist of a set of dilutions comparing the test samples to a standard curve and calculating a relative potency. The design could be extended to include additional analysts to improve the estimation of analyst as a random

component and additional factors such as Laboratory by repeating the design across multiple laboratories.

Typically, samples (concentrations) would be plated out according to a serial dilution scheme across columns 1–12, with pairs of contiguous rows starting with A–B corresponding to replicates of a given sample. In this case, the BIBD design calls for three samples to be present on each plate plus a standard curve, where each plate can be thought of as a block in relation to samples. Samples are plated out as if they were at 100 % concentration. The measured concentrations across all plates according to the given BIBD are related to linearity, accuracy (closeness of % recovery to target) and precision. The allocation of samples to pairs of rows across the 20 plates is an important design question. Ideally, one would seek to balance row locations (pairs of rows) on the plate with sample concentration, to orthogonalize location within plates effects with plate and sample concentration effects. One can imagine a latin square design across plates, arranged in such a way as to provide at least proportionality across plates for row and sample effects. It is not known whether a general construction method exists for the interweaving of a latin square with a BIBD to provide such balance as we have in this design. In such a situation, one can construct an allocation that achieves near-orthogonality, by allocating each sample concentration to each pair of row locations (A–B, C–D, E–F, G–H) as close to an equal number of times as possible across plates. In this design, each sample is present in half of the plates. Therefore, the best one can do is six permutations of 2,2,3,3 corresponding to number of plates with a given sample in locations A–B, C–D, E–F, G–H. An example of such a design is given in Table 15.2 where 1,2,3,4,5,6 correspond to sample concentrations 25,50,67,100,150,200 % of target respectively, and S stands for Standard.

The overall statistical modeling and analysis would be carried out in two steps. First, potency estimates will be generated based on the constrained four-parameter logistic model, followed by a linear mixed model incorporating within and between plate variance terms to produce the final combined estimates. Other important aspects related to potency estimation in the context of bioassays is given by Davidian and Giltinan 1993, Giltinan 1998, Lansky 2002, O'Connell et al 1993 and Rodbard et al 1994.

15.2.9.1   Statistical Model for Potency Estimation

For potency estimation purposes, assume a heteroscedastic model $y_i = f(x_i, \beta) + \varepsilon_i$ where y_i denotes the independent responses (count, OD, etc) at concentration x_i, $f(x_i, \beta)$ is typically the four parameter logistic given by:

$$f(x_i, \beta) = \beta_2 + \frac{\beta_1 - \beta_2}{1 + \left(\frac{x_i}{\beta_3}\right)^{\beta_4}} = \beta_2 + \frac{\beta_1 - \beta_2}{1 + \exp\left(\beta_4 \left[\log(x_i) - \log(\beta_3)\right]\right)} \quad (15.5)$$

Table 15.2 Allocation of treatments to plate locations

Analyst	Run	Plate	Plate row A–B	C–D	E–F	G–H
1	1	1	S	1	2	3
		2	5	S	1	6
	2	3	3	4	S	5
		4	2	3	6	S
	3	5	S	2	4	1
		6	6	S	5	2
	4	7	3	6	S	4
		8	4	1	6	S
	5	9	S	5	2	4
		10	1	S	3	5
2	1	1	4	5	S	6
		2	2	3	4	S
	2	3	S	6	1	2
		4	5	S	1	4
	3	5	6	5	S	3
		6	4	1	3	S
	4	7	S	2	5	1
		8	3	S	2	5
	5	9	1	6	S	3
		10	2	4	6	S

where $\beta = (\beta_1, \beta_2, \beta_3, \beta_4)$ and $\beta_1 = $ asymptote as the concentration $x \to 0$ (for $\beta_4 > 0$), $\beta_2 = $ asymptote as $x \to \infty$, $\beta_3 = $ concentration corresponding to response halfway between the asymptotes, $\beta_4 = $ slope parameter. One could argue that these parameter values are fixed constants related to physical properties of the molecule but subject to variation due to known and unknown factors. The mean response is given by $E(y_i) = \mu_i = f(x_i, \beta)$. Then $Var(y_i) = \sigma^2 g^2\{f(x_i, \beta), \gamma\}$, where $g^2\{f(x_i, \beta), \gamma\}$ is referred to as the variance function with parameter γ and expresses the heteroscedasticity, σ^2 is a scale parameter. The least squares method used to estimate the parameters given in (15.5) requires weighting as a consequence of the heteroscedastic model. The weights are provided by the variance function and therefore its correct specification is critical not only to the estimation procedure itself, but also to the calculation of standard errors for the parameter estimates. For the purposes of this discussion, the power of the mean function will be used relating variance to mean as $g^2\{f(x_i, \beta), \gamma\} = \mu_i^\gamma$.

Generalized least squares (GLS) method is commonly used to estimate the parameters. The method is as follows: (i) estimate β by a preliminary unweighted fit using ordinary least squares, (ii) use the residuals from the preliminary fit to estimate γ and σ, (iii) based on the estimates from (ii), form new weights, and re-estimate β using weighted least squares chosen to minimize the objective function given by

$$O_{GLS} = \sum \left[\frac{(y_i - \widehat{\mu}_i)^2}{\sigma^2 \widehat{\mu}_i^{\gamma}} + \log \left(\sigma^2 \widehat{\mu}_i^{\gamma} \right) \right] \tag{15.6}$$

where $\widehat{\mu}_i = f(x_i, \widehat{\beta})$ and $\widehat{\beta}_i$ is the estimator of β_i from the previous step and return to step (ii) iterating until convergence. The use of (15.6) to estimate γ and σ leads to pseudo-likelihood estimates of the parameters (Giltinan and Ruppert 1989).

15.2.9.2 The Constrained Four Parameter Logistic Model

The term 'constrained' implies that p (p > 1) curves are being estimated, and that 1 or more parameters may be common across the curves. Suppose p = 2, such as when a test preparation is being compared to a standard. Extend Eq. (15.5) to accommodate both curves using a piecewise regression notation. Let s,t index standard and test respectively and assume $\beta_{s1} = \beta_{t1}$, $\beta_{s2} = \beta_{t2}$, $\beta_{s4} = \beta_{t4}$ letting only β_{s3} and β_{t3} vary as a consequence of the conditions of similarity (Finney 1978). Further, let $\beta^*_s = \log \beta_{s3}$ and $\beta^*_t = \log \beta_{t3}$ and since the log relative potency $\rho_t = \beta^*_s - \beta^*_t$, a relative potency parameter can be incorporated yielding the model:

$$y_i = \beta_2 + \frac{\beta_1 - \beta_2}{1 + \exp \beta_4 \{I_s \log x_i + I_t (\log x_i + \rho_t) - \beta^*_s\}} + e_i \tag{15.7}$$

where the concentrations $x_i's$, and responses $y_i's$ are arranged according to indicator variables I_s and I_t denoting standard and test, respectively. Estimates of the parameters given in (15.7) and their estimated covariance matrix are produced as part of the GLS algorithm. The model is easily extended to the case of more than a single test vs. a standard.

Statistical testing of similarity is a requirement as a consequence of validity considerations. In the case of the common parallel line linear model with fixed parameters, the classical approach (Finney 1978) relies on the principle of extra sums of squares. Extensions of this approach to the non-linear mixed model case have not been reported in the literature although Lansky (1999) mentions a test for parallelism as the essential test in relation to a split-block design. More modern approaches rely on an equivalence test to establish similarity. The equivalence testing is applied to the two asymptote parameters and the shape parameter. A good discussion on equivalence testing to establish similarity is found in the USP 38/NF 33 (2015) and Lansky (2014).

Mixed effects modeling can be applied to combine estimates across plates to produce an average potency across tests, variance components and address linearity. An appropriate mixed effects model would lead to potency estimates adjusted for between and within plate sources of variation (Altan et al. 2002). The variance component estimates of run, plate and residual variance coming out of the mixed model would be combined to report intermediate precision. Repeatability would be the residual error. These estimates can be combined in various ways to report an assay format uncertainty term corresponding to a reportable value. These ideas are discussed in greater detail in USP 38/NF 33 <1033> 2015.

15.3 Bayesian Applications: Probability of Conformance and ICH Q8 Design Space

Process optimization and reliability assessment are key issues for CMC statistics. Because Bayesian methods directly lead to predictive distributions, they are in turn useful for reliability assessment and process optimization by way of reliability assessment. This is true for both single and multiple response processes.

15.3.1 Probability of Conformance and Process Capability

An important issue in quality-by-design is being able to calculate the probability that a process will meet its specifications. This is important, because a process may have been "optimized" by means of its associated mean response surface, but this does not mean that the process is likely to meet its specifications. In fact, a traditional approach of using "overlapping mean responses" for optimizing processes with multiple quality attributes can result in misleading inferences for assessing the probability of conformance and ICH Q8 design space (Peterson and Lief 2010).

A natural and easy-to-understand approach to process capability, particularly for multiple response processes, is to simply compute the probability that the process will meet all of its specifications simultaneously. This can be written from a Bayesian perspective as

$$\Pr\left(Y \in S \,|\, data\right), \qquad (15.8)$$

where the probability in (15.8) is based upon the posterior predictive distribution (unconditional on the model parameters). This notion for process capability was first put forth by Bernardo and Irony (1996). This concept was extended to regression models by Peterson (2004), thereby creating the function

$$p\left(x\right) = \Pr\left(Y \in S \,|\, x, data\right), \qquad (15.9)$$

where x is a vector of process factors and Y is related to x by way of a stochastic response surface model. By optimizing the $p(x)$ function, one can optimize the process in a way that optimizes a measure of process capability. As it turns out, this Bayesian predictive approach to process optimization adapts itself quite easily to solve many process optimization problems that have challenging complexities.

For multivariate problems, one would typically need to fit different model forms to each response-type, so as to not over fit. This is because some response types may require simple models, while others may require more complex model forms. If (parametrically) linear models are being used, different model forms will result in what are called "seemingly unrelated regression" (SUR) models (Zellner 1962; Srivastava and Giles 1987). Fortunately, SUR models are easy to analyze using

Gibbs sampling (Percy 1992; Peterson et al. 2009b). The Bayesm R package (Rossi 2012) has a function, rsurGibbs, which can be used to sample from the posterior for a SUR model.

The process optimization approach in (15.9) can be easily modified to solve the multiple-response robust parameter design problem (Miró-Quesada et al. 2004). For the robust parameter design problem, some of the factors are noisy. The idea is to adjust the controllable factors so as to reduce the variation transmitted by the noisy factors. The Bayesian predictive approach is able to do this in a very natural way by averaging over the noise distribution to obtain

$$p\left(\boldsymbol{x}_c\right) = E_{\boldsymbol{x}_n}\left\{\Pr\left(\boldsymbol{Y} \in S \,|\, \boldsymbol{x}_c,\, \boldsymbol{x}_n,\, data\right)\right\} = \Pr\left(\boldsymbol{Y} \in S \,|\, \boldsymbol{x}_c,\, data\right) \qquad (15.10)$$

where $\boldsymbol{x}_n$ is a vector of noise variables and $\boldsymbol{x}_c$ is a vector of the controllable factors. Optimization of $p\left(\boldsymbol{x}_c\right)$ in (15.10) adjusts the controllable factor levels so that noise variable error transmission is dampened in just the right way as to optimize the probability of conformance to specifications.

The posterior predictive optimization approach can also be generalized to (univariate or multivariate) mixed-effect models. For example, one can optimize the probability of conformance for a (univariate or multivariate) split-plot experiment. The most technically difficult part here is sampling from the posterior of the model parameters. Once that has been accomplished, optimization of the $p(\boldsymbol{x})$ is fairly straightforward.

WinBUGS (Lunn et al. 2000) or OpenBUGS (Thomas et al. 2006) software can be used to sample from the posterior for SUR models or mixed-effect SUR models that also contain random effects. In addition, the R package, MCMCglmm (Hadfield 2010) can also be used to sample from the posterior for such models. It is even possible to use Bayesian model averaging (Press 2003, chapter 13) to account for uncertainty of model form in the $p(\boldsymbol{x})$ function. See articles by Rajagopal and del Castillo (2005), Rajagopal et al. (2005), and Ng (2010) for details.

The textbook, *Process Optimization – A Statistical Approach*, (del Castillo 2007) is the only one that gives an introduction to process optimization using the Bayesian predictive approach. Del Castillo (2007) lists seven advantages of using the Bayesian predictive approach to process optimization. These are:

1. It considers the uncertainty of the model parameters. (Many response surface techniques do not do this.)
2. It considers the correlation among the responses. (Classical desirability function approaches (e.g. Harrington 1965; Derringer and Suich 1980) do not do this. However, Chiao and Hamada (2001) and Peterson (2008) show that the joint probability of conformance to specifications can be strongly dependent upon the correlation structure of the regression residuals.)
3. Informative prior information can be used, as well as non-informative priors.
4. A well-calibrated (factor) region ("sweet spot") of acceptable probability of conformance to specifications can be constructed in a straightforward manner. (The classical "overlapping means" approach to constructing a sweet spot can result in a region with quite low probabilities of meeting specification (Peterson and Lief 2010).

5. It can be used for more general optimizations such as $p(x) = \Pr(D(y) \geq D*|x, data)$, where $D(y)$ is a desirability function (Derringer and Suich 1980; Harrington 1965). Note, however, that such desirability function approaches as typically used in classical response surface optimization do not provide a measure of how likely it is that the desirability will exceed a pre-specified amount, e.g. $D*$.

6. It allows one to perform a pre-posterior analysis if the optimal probability of conformance is too low. Here, by "pre-posterior analysis" we mean that one can simulate additional data from a fitted model, and then see how much this additional information changes the posterior probability of conformance to specifications. If the change goes from say, 0.83 to 0.96 and 0.96 is acceptable, then this suggests that our process is adequate, and that we may only need additional data for confirmation purposes. If, however, after the addition of substantially more "data", the change goes from 0.83 to 0.88 and 0.88 is not acceptable, then this indicates that we may need to improve the process itself.

7. It is easy to add noise variables, thereby providing a solution to the multivariate robust parameter design problem.

It is also useful to note that Bayesian inference has been successfully applied to a variety of CMC problems. For example, the Bayesian predictive approach has been applied to complex sampling algorithms that are sometimes used in the pharmaceutical industry. For example, it may be desired to compute the probability of a certain drug passing a multi-stage USP test, whereby several tablets are sampled and tested, with the possibility that more may need to be sampled and tested. LeBlond and Mockus (2014) provide an example of quantifying the probability of passing a compendial standard for content uniformity. The Bayesian approach of computing relevant probabilities has also been applied to some assay validation problems (Novick et al. 2011, 2012). Recently, the Bayesian approach has been applied to dissolution curve comparisons. See LeBlond et al. (2015) for a review and Novick et al. (2015) for an innovative application.

15.3.2 Probability of Conformance and ICH Q8 Design Space

The ICH Q8 Guidance on Pharmaceutical Development has proposed the notion of a "design space". This is defined as "the multidimensional combination and interaction of input variables (e.g., material attributes) and process parameters that have been demonstrated to provide assurance of quality". One can think of a design space as a collection of manufacturing recipes each of which should be likely to meet quality specifications. However, for actually constructing a design space, it is helpful to have a mathematical definition, that is the design space equation. From a Bayesian predictive perspective, one can define the ICH Q8 design space as

$$\{x : \Pr(Y \in S | x, data) \geq R\},$$

where, as before, x is a vector of process factors and Y is a vector of quality response-types (also referred to as quality attributes) (Peterson 2008). Here, R is a pre-specified reliability level. See also Peterson (2009).

It is curious to note that Appendix 2 of the ICH Q8 Guidance gives an example of a design space that appears to be of the form of overlapping mean response surfaces. However, as shown by Peterson and Lief (2010) such a design space may have very low probability of simultaneously meeting all specifications! It may be that this statistical issue was a blind spot for the developers of the ICH Q8 Guidance. Fortunately, further publications supporting a Bayesian approach to ICH Q8 design space have appeared in the literature. See for example Stockdale and Cheng (2009), Peterson et al. (2009a, b), Castagnoli et al. (2010), Peterson and Kenett (2011), LeBrun et al. (2012), Maeda et al. (2012), Crump et al. (2013), Mockus et al. (2014), Gong et al. (2015), Chavez et al. (2015), Chatzizaharia and Hatziavramidis (2015), and LeBrun et al. (2015).

ICH Q8 design space concepts have been applied not only to pharmaceutical manufacturing, but also to pharmaceutical assay development. Thus one could have a collection of assay conditions that have been demonstrated to be likely to have quality attributes that meet the assay specifications. A Bayesian approach to design space for assay robustness is described in Peterson and Yahyah (2009). Subsequent articles have appeared in the literature: Mbinze et al. (2012), LeBrun et al. (2013), Hubert et al. (2014), Dispas et al. (2014a, b), and Rozet et al. (2012, 2015).

References

Altan S, Shoung J (2008) Block designs in method transfer experiments. J Biopharm Stat 18: 996–1004

Altan S, Manola A, Davidian M, Raghavarao D (2002) The constrained four parameter logistic model. Dev Biol Standard 107:71–76

ASTM E2655-14 (2014) Standard guide for reporting uncertainty of test results and use of the term measurement uncertainty in ASTM test methods. ASTM International, West Conshohocken. www.astm.org

Barlow RE, Proschan F (1975) Statistical theory of reliability and life testing – probability models. Holt, Rinehart, and Winston, Inc., New York

Bernardo JM, Irony TZ (1996) A general multivariate Bayesian process capability index. Statistician 45:487–502

Box GEP, Behnken DW (1960) Some new three-level designs for the study of quantitative variables. Technometrics 2:455–475

Breyfogle FW (2003) Implementing six sigma, 2nd edn. Wiley, Hoboken

Castagnoli C, Yahyah M, Cimarosti Z, Peterson JJ (2010) Application of quality by design prinicpals for the definition of a robust crystallization process for casopitant mesylate. Org Process Res Dev 14(6):1407–1419

Chatzizaharia KA, Hatziavramidis TD (2015) Dissolution efficiency and design space for an oral pharmaceutical product in tablet form. Ind Eng Chem Res. doi:10.1021/ie5050567

Chavez P-F, Lebrun P, Sacré P-Y, De Bleye C, Netchacovitch L, Cuypers S, Mantanus J, Motte H, Schubert M, Evrard B, Hubert P, Ziemons E (2015) Optimization of a pharmaceutical tablet formulation based on a design space approach and using vibrational spectroscopy as PAT tool. Int J Pharm 486:13–20

Chiao C, Hamada M (2001) Analyzing experiments with correlated multiple responses. J Qual Technol 33:451–465

Cho BR, Shin S (2012) Quality improvement and robust design methods to a pharmaceutical research and development. Math Prob Eng. http://dx.doi.org/10.1155/2012/193246

Cornell JA (2002) Experiments with mixtures – designs, models, and the analysis of mixture data, 3rd edn. Wiley, New York

Crump BR, Goss C, Lovelace T, Lewis R, Peterson J (2013) Influence of reaction parameters on the first principles reaction rate modeling of a nitro reduction. Org Process Res Dev 17(10): 1277–1286

Davidian M, Giltinan D (1993) Some simple methods for estimating intra-individual variability in nonlinear mixed effects models. Biometrics 49:59–73

Del Castillo E (2007) Process optimization—a statistical approach. Springer, New York

Derringer G, Suich R (1980) Simultaneous optimization of several response variables. J Qual Technol 12:214–219

Dispas A, Lebrun P, Andri B, Rozet E, Hubert P (2014a) Robust method optimization strategy—a useful tool for method transfer: the case of SFC. J Pharm Biomed Anal 88:519–524

Dispas A, Lebrun P, Ziemons E, Marini R, Rozet E, Hubert PH (2014) Evaluation of the quantitative performances of supercritical fluidchromatography: from method development to validation. J Chromatogr A 1353:78–88

Draper NR (1963) Ridge analysis of response surfaces. Technometrics 5:469–479

Food and Drug Administration (2011) Center for drugs evaluation research. Guidance for industry process validation: general principles and practices. http://www.fda.gov/downloads/Drugs/Guidances/UCM070336.pdf

FDA (2015a) Pharmaceutical quality/manufacturing standards (CGMP) http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/ucm064971.htm

FDA (2015b) Pharmaceutical cGMPS for the 21st century — a risk-based approach: second progress report and implementation plan. http://www.fda.gov/Drugs/DevelopmentApprovalProcess/Manufacturing/QuestionsandAnswersonCurrentGoodManufacturingPracticescGMPforDrugs/UCM071836

FDA (2015c) CFR – code of federal regulations title 21. http://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/cfrsearch.cfm

Finney DJ (1978) Statistical method in biological assay, 3rd edn. Charles Griffin, New York

Garcìa A, Gilabert E (2011) Mapping FMEA into Bayesian networks. Int J Performability Eng 7(6):525–537

Giltinan D (1998) Statistical issues in assay development and use. Adv Exp Med Biol 445:173–190

Giltinan DM, Ruppert D (1989) Fitting heteroscedastic regression models to individual pharmacokinetic data using standard statistical software. J Pharmacokinet Biopharm 17:601–614

Gong X, Chen H, Pan J, Qu H (2015) Optimization of *Panax notoginseng* extraction process using a design space approach. Sep Purif Technol 141:197–206

Hadfield JD (2010) MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R package. J Stat Softw 33(2):1–22

Harrington EC (1965) The desirability function. Ind Qual Control 21:494–498

Hoerl A (1964) Ridge analysis. Chem Eng Symp Ser 60:67–77

Hubert C, LeBrun P, Houari S, Ziemons E, Rozet E, Hubert P (2014) Improvement of a stability-indicating method by quality-by-design versus quality-by-testing: a case of a learning process. J Pharm Biomed Anal 88:401–409

International Society of Professional Engineers (2015) Quality metrics summit: driving quality through data and knowledge. Baltimore. http://www.ispe.org/2015-quality-metrics-summit

Jones B, Nachtsheim CJ (2011) A class of three-level designs for definitive screening in the presence of second-order effects. J Qual Technol 43(1):1–15

Juran JM (1988) Juran on planning for quality. The Free Press, New York

Kenett RS, Zacks S (2014) Modern industrial statistics: with applications in R, MINITAB and JMP. Wiley, Chichester

Kotz S, Johnson NL (2002) Process capability indices – a review, 1992–2000. J Qual Technol 34(1):2–19

Lansky D (1999) Validation of bioassays for quality control. Brown F, Mire-Sluis AR (eds): biological characterization and assay of cytokines and growth factors. Dev Biol Stand Basel, Karger 97:157–168

Lansky D (2002) Strip plot designs, mixed models, and comparisons between linear and non-linear models in microtitre plate bioassays. Dev Biol Stand 107:11–23

Lansky D (2014) Near-universal similarity bounds for bioassays. In: Conference on statistical and data management approaches for biotechnology drug development, August 27–28, 2014, Rockville, MD

LeBlond D, Mockus L (2014) The posterior probability of passing a compendial standard, part 1: uniformity of dosage units. Stat Biopharm Res 6(3):270–286

LeBlond D, Altan S, Novick S, Peterson J, Shen Y, Yang H (2015) In-vitro dissolution curve comparisons: a critique of current practice. Dissolut Technol, To appear http://www.qualitydigest.com/inside/quality-insider-article/problems-risk-priority-numbers.html

Lebrun P, Krier F, Mantanus J, Grohganz H, Yang M, Rozet E, Boulanger B, Evrard B, Rantanen J, Hubert P (2012) Design space approach in the optimization of the spray-drying process. Eur J Pharm Biopharm 80:226–234

Lebrun P, Boulanger B, Debrus B, Lambert P, Hubert P (2013) A Bayesian design space for analytical methods based on multivariate models and predictions. J Biopharm Stat 23(6): 1330–1351

Lebrun P, Giacoletti K, Scherder T, Rozet E, Boulanger B (2015) A quality by design approach for longitudinal quality attributes. J Biopharm Stat 25:247–259

Lostritto R (2014) Clinically relevant specifications (CRS): a regulatory perspective. IFPAC Conference, Baltimore

Lunn DJ, Thomas A, Best N, Spiegelhalter D (2000) WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility. Stat Comput 10:325–337

Maeda J, Suzuki T, Takayama K (2012) Novel method for constructing a large-scale design space in lubrication process by using Bayesian estimation based on the reliability of a scale-up rule. Chem Pharm Bull 60(9):1155–1163

Marroum PJ (2012) Clinically relevant dissolution methods and specifications. http://www.americanpharmaceuticalreview.com/Featured-Articles/38389-Clinically-Relevant-Dissolution-Methods-and-Specifications/

Mbinze JK, Lebrun P, Debrus B, Dispas A, Kalenda N, Mavar Tayey Mbay J, Schofield T, Boulanger B, Rozet E, Huberta P, Marini RD (2012) Application of an innovative design space optimization strategy to the development of liquid chromatographic methods to combat potentially counterfeit nonsteroidal anti-inflammatory drugs. J Chromatogr A 1263:113–124

Miró-Quesada G, del Castillo E, Peterson JJ (2004) A Bayesian approach for multiple response surface optimization in the presence of noise variables. J Appl Stat 31:251–270

Mockus L, Peterson JJ, Lainez JM, Reklaitis GV (2014) Batch-to-batch variation: a key component for modeling chemical manufacturing processes. Org Process Res Dev. doi:10.1021/op500244f

Montgomery DC (2009) Introduction to statistical quality control, 6th edn. Wiley, Hoboken

Myers RH, Montgomery DC, Anderson-Cook CM (2009) Response surface methodology - product and process optimization using designed experiments, 3rd edn. Wiley, Hoboken

Nelson PR (1992) Editorial. J Qual Technol 24(4):175

Ng SH (2010) A Bayesian model averaging approach to multiple-response optimization. J Qual Technol 42(1):52–68

Novick SJ, Chiswell K, Peterson JJ (2011) A Bayesian approach to show assay equivalence with replicate responses over a specified concentration range. Stat Biopharm Res 4(2):102–117

Novick SJ, Yang H, Peterson JJ (2012) A Bayesian approach to parallelism in testing in bioassay. Stat Biopharm Res 4(4):357–374

Novick SJ, Shen Y, Yang H, Peterson JJ, LeBlond D, Altan S (2015) Dissolution curve comparisons through the F_2 parameter, a Bayesian extension of the f_2 statistic. J Biopharm Stat 25(2): 351–371

O'Connell MA, Belanger BA, Haaland PD (1993) Calibration and assay development using the four-parameter logistic model. Chemometr Intell Lab Syst 20:97–114

Percy DF (1992) Prediction for seemingly unrelated regressions. J R Stat Soc Ser B 54:243–252

Peterson JJ (1993) A general approach to ridge analysis with confidence intervals. Technometrics 35:204–214

Peterson JJ (2004) A posterior predictive approach to multiple response surface optimization. J Qual Technol 36:139–153

Peterson JJ (2008) A Bayesian approach to the ICH Q8 definition of design space. J Biopharm Stat 18:959–975

Peterson JJ (2009) What your ICH Q8 design space needs: a multivariate predictive distribution. Pharm Manuf 8(10):23–28

Peterson J, Kenett R (2011) Modeling opportunities for statisticians supporting quality by design efforts for pharmaceutical development & manufacturing. Biopharm Rep 8(2):6–16

Peterson JJ, Lief K (2010) The ICH Q8 definition of design space: a comparison of the overlapping means and the bayesian predictive approaches. Stat Biopharm Res 2:249–259

Peterson JJ, Yahyah M (2009) A Bayesian design space approach to robustness and system suitability for pharmaceutical assays and other processes. Stat Biopharm Res 1(4):441–449

Peterson JJ, Snee RD, McAllister PR, Schofield TL, Carella AJ (2009a) Statistics in the pharmaceutical development and manufacturing (with discussion). J Qual Technol 41:111–147

Peterson JJ, Miro-Quesada G, del Castillo E (2009b) A Bayesian reliability approach to multiple response optimization with seemingly unrelated regression models. J Qual Technol Quant Manag 6(4):353–369

Press SJ (2003) Subjective and objective Bayesian statistics, 2nd edn. Wiley, Hoboken

Q10 Pharmaceutical Quality System (2009) International conference on harmonizaiton, guidance for industry. http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/ucm073517.pdf

Q8(R2) Pharmaceutical Development (2009) International conference on harmonizaiton guidance for industry. http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Quality/Q8_R1/Step4/Q8_R2_Guideline.pdf

Q9 Quality Risk Management (2006) International conference on harmonizaiton, guidance for industry. http://www.fda.gov/downloads/Drugs/.\kern\fontdimen3\font.\kern\fontdimen3\font.\kern\fontdimen3\font/Guidances/ucm073511.pdf

Rajagopal R, del Castillo E (2005) Model-robust process optimization using Bayesian model averaging. Technometrics 47(2):152–163

Rajagopal R, del Castillo E, Peterson JJ (2005) Model and distribution-robust process optimization with noise factors. J Qual Technol 37:210–222; Corrigendum 38, p 83

Rodbard D, Lenox RH, Wray HL, Ramaith D (1994) Statistical quality control and routine data processing for radioimmunoassay. Clin Chem 20/10:1255–1270

Rossi P (2012) Bayesm: Bayesian inference for marketing/micro-econometrics. R package version 2.2-5. http://CRAN.R-project.org/package=bayesm

Rozet E, Lebrun P, Debrus B, Philippe Hubert P (2012) New methodology for the development of chromatographic methods with bioanalytical application. Bioanalysis 4(7):755–758

Rozet E, Lebrun P, Michiels J-F, Sondag P, Scherder T, Boulanger B (2015) Analytical procedure validation and the quality by design paradigm. J Biopharm Stat 25:260–268

Scranton R, Runger GC, Keats JB, Montgomery DC (1996) Efficient shift detection using exponentially weighted moving average control charts and principal components. Qual Reliab Eng Int 12(3):165–172

Sharp SS (2012) Establishing clinically relevant drug product specifications: FDA perspective FDA/ONDQA/biopharmaceutics 2012 AAPS annual meeting and exposition, Chicago, IL. http://www.fda.gov/downloads/AboutFDA/CentersOffices/OfficeofMedicalProductsandTobacco/CDER/UCM341185.pdf

Shin S, Truong NKV, Goethals PL, Cho BR, Jeong SH (2014) Robust design modeling and optimization of a multi-response time series for a pharmaceutical process. Int J Adv Manuf Technol 74:1017–1031

Srivastava VK, Giles DEA (1987) Seemingly unrelated regression equations models. Marcel Dekker, New York

Stockdale G, Cheng A (2009) Finding design space and reliable operating region using a multivariate Bayesian approach with experimental design. Qual Technol Quant Manag 6(4):391–408

Thomas A, O'Hara B, Ligges U, Sturtz S (2006) Making BUGS open. R News 6(1):12–17

USP 38/NF 33 (2015) Design and development of biological assays <1032>, vol 1. United States Pharmacopeial Convention, Rockville, pp 769–796

USP 38/NF 33 <1033> (2015) Biological assay validation, vol 1. United States Pharmacopeial Convention, Rockville, pp 796–807

Vander Heyden, Y., Nijhuis, A., Smeyers-Verbeke, J., Vandeginste, B. G. M., and Massart, D. L. (2001). "Guidance for Robustness/Ruggedness Tests in Methods Validation," *Journal of Pharmaceutical and Biomedical Analysis*, 24, 723–753. 441, 444, 445, 446

Wheeler DJ (2011) Problems with risk priority numbers. Qual Dig. http://www.qualitydigest.com/inside/quality-insider-article/problems-risk-priority-numbers.html

Zellner A (1962) An efficient method of estimating seemingly unrelated regression equations and tests for aggregation bias. J Am Stat Assoc 57:348–368