

Chapter 2

Regulatory Nonclinical Statistics

Mohammad Atiar Rahman, Meiyu Shen, Xiaoyu (Cassie) Dong, Karl K. Lin, and Yi Tsong

Abstract The nonclinical statistics teams in the Center of Drug Review and Research of the Food and Drug Administration (FDA) conduct regulatory reviews, statistical consultation, and statistical methodology development in nonclinical regulations. In this chapter, we provide a brief description of the two teams and provide two examples in statistical research development. In the first example, we describe the historical background and evolution of statistical methodology development in the last 20 years for the acceptance sampling and lot evaluation procedures on dose content uniformity involved with FDA Chemistry Manufacturing, and Control (CMC) Statistics Team. In the second example, we illustrate the research activities of Pharmacological/Toxicological (Pharm-Tox) Statistics Team at FDA with the background and evaluation of multiple pairwise comparisons in animal carcinogenetic studies.

Keywords Chemistry manufacturing, and control • Acceptance sampling • Content uniformity • Pharmacological/toxicological studies

2.1 Background

Regulatory Nonclinical Statistics in the Center for Drug Evaluation and Research (CDER) of the Food and Drug Administration (FDA) consists of two teams: Chemistry, Manufacturing, and Control (CMC) and Pharmacological/Toxicological study teams. The two nonclinical statistics teams are located within the Division of Biometrics VI of the Office of Biostatistics.

This presentation reflects the views of the authors and should not be construed to represent FDA's views or policies.

M.A. Rahman • M. Shen • X. Dong • K.K. Lin • Y. Tsong (✉)
Division of Biometrics VI, Office of Biostatistics, Office of Translational Science, Center for Drug Evaluation and Research, Food and Drug Administration, 10903 New Hampshire Ave., Silver Spring, MD 20993, USA
e-mail: yi.tsong@fda.hhs.gov

2.1.1 CMC Statistics Team

The CMC Statistics team provides statistical expertise to FDA to ensure product quality through review, regulation, and research for new drugs, biological products, biosimilar products, and generic drugs. Currently, this team consists of seven Ph.D. level statisticians including one team leader and one technical leader.

There are two types of reviews: consultation review and regulatory review. For the consultation review, the CMC Statistics team responds to the consultation requests from CDER/FDA on a broad range of CMC issues, including new drug CMC reviews, scale-up and post-approval changes (SUPAC), new drug bioequivalence including in-vivo and in-vitro, generic drug in-vitro bioequivalence product stability, product specification, botanical drug product consistency, post-marketing quality surveillance, and other quality issues. These consultation reviews provide important statistical support to many CMC related offices in CDER/FDA. These consults may be requested by Office of Policy for Pharmaceutical Quality (OPPQ), Office of New Drugs (OND), Office of Generic Drugs (OGD), Office of Product Quality (OPQ), Office of Biotechnology Products (OBP), Office of New Drug Products (ONDP), Office of Lifecycle Drug Products (OLDP), Office of Testing and Research (OTR), Office of Process and Facilities (OPF), and Office of Surveillance (OS).

The aim of a statistical CMC consultation review is to provide statistical expertise to address specific issues in a regulatory submission, e.g., the shelf life determination for a product. It may also provide evaluation on statistical approaches proposed by sponsors, e.g., to evaluate the suitability of a stability model used in a submission. Under the consultation setup, review comments from the CMC Statistics team may not be conveyed directly to the sponsor nor is a part of the regulatory decision making.

The typical process of a CMC statistical consultation is outlined as follows. Chemist reviewers or biologist reviewers send a CMC statistical consultation request to the CMC Statistics team through the project manager. Upon receiving the request, the CMC Statistics team leader assigns the work to one CMC team member. The assigned primary statistical reviewer and the team leader or technical leader will meet with the chemist or biologist to discuss the review issues. Once the work is completed, the statistical consultation review will be put into the FDA's filing system and then will be signed off by the reviewer, the team leader, and the division director.

In addition to the consultation review, the CMC Statistics team conducts regulatory reviews for biosimilar submissions. As part of the review team since 2013, the CMC statistical review team has been playing a key role in reviewing the statistical assessment of analytical similarity for Investigational New Drug (IND) and Biological License Application (BLA) submissions for biosimilar biological products. Our review comments will be conveyed to the sponsor directly and are critical to the regulatory decision making. Our statistical findings and evaluation may be presented at Advisory Committee meetings.

Besides the consultation and biosimilar reviews introduced earlier, the CMC Statistics team also develops CMC statistical methodologies for product quality assurance. We will illustrate this with one example in Sect. 2.2.

2.1.2 Pharmacological/Toxicological Review Team

The risk assessment of a new drug exposure in humans usually begins with an assessment of the risk of the drug in animals. It is required by law that the sponsor of a new drug conducts nonclinical studies in animals to assess the pharmacological actions, the toxicological effects, and the pharmacokinetic properties of the drug in relation to its proposed therapeutic indications or clinical uses. Studies in animals, designed for assessment of toxicological effects of the drug, include acute, subacute, subchronic, chronic toxicity studies, carcinogenicity studies, reprotoxicology studies, and pharmacokinetic studies.

The statistical reviews and evaluations of toxicology studies of new drugs is an integrated part of FDA drug review and approval process. The Pharm/Tox Statistics Team in the Office of Biostatistics in the Center for Drug Evaluation and Research of FDA is responsible for this area of the review and approval process. An assessment of the risk for carcinogenicity includes life-time tests in mice and rats. The primary purpose of a long-term animal carcinogenicity experiment is to determine the oncogenic potential of a new drug when it is administered to animals for the majority of their normal lifespan.

Regular long-term (chronic) carcinogenicity studies of a new drug are usually planned for 2 years (104 weeks) in rats and mice. However, in the 1990s ICH started allowing a drug sponsor to conduct a 26-week transgenic mouse study to replace the regular 2-year mouse study in its new drug application submission. At least three dose groups and a negative control and a positive control (treated with a known carcinogen, e.g., *p*-cresidine, *N*-methyl-*N*-nitrosourea, benzene, or 12-*O*-tetradecanoyl-phorbol-13-acetate (TPA)) are used in the transgenic mouse study with 25–30 mice sex/group. The histopathology endpoints of the transgenic mouse study are the same as those used in the 2-year studies except the 26-week study using Tg.AC transgenic mice. The interest in detecting oncogenic potential of a new drug is to test if there are statistically significant positive linear trends (or dose–response relationships) induced by the new drug. Based on the interest, a typical statistical review and evaluation of the carcinogenicity studies of a new drug performed by FDA nonclinical statisticians includes the essential parts described below. The statistical methods used in the FDA review and evaluation are based on results of FDA internal research and guidances or guidelines of regulatory agencies and research institutions such as WHO, ICH, NIH inside and outside U.S.

In the analysis of tumor data, it is essential to identify and adjust for the possible differences in intercurrent mortality (or longevity) among treatment groups to eliminate or reduce biases caused by these differences. Intercurrent mortality refers to all deaths not related to the development of a particular type or class of

tumors to be analyzed for evidence of carcinogenicity. Like human beings, older animals have many times higher probability of developing or dying of tumors than those of younger age. The Cox's Test, the generalized Wilcoxon or Kruskal–Wallis test, and the Tarone trend tests are routinely used to test the heterogeneity in survival distributions and significant dose–response relationship (linear trend) in mortality. The choice of a survival-adjusted method to analyze tumor data depends on the role which a tumor plays in causing the animal's death. Tumors can be classified as “incidental”, “fatal”, and “mortality-independent (or observable)” according to the contexts of observation described in the WHO monograph by Peto et al. (1980). Tumors which are directly or indirectly responsible for the animal's death, but are merely observed at the autopsy of the animal after it has died of some unrelated causes, are said to have been observed in an incidental context. Tumors which kill the animal either directly or indirectly are said to have been observed in a fatal context. Tumors, such as skin tumors, whose times of criterion attainment (that is, detection of the tumor at a standard point of their development, other than the times or causes of death, are of primary interest in analyses, are said to have been observed in a mortality-independent (or observable) context. To apply a survival-adjusted method correctly, it is essential that the context of observation of a tumor be determined as accurately as possible.

Different statistical techniques have been proposed for analyzing data of tumors observed in different contexts of observation. For example, the prevalence method, the death-rate method, and the onset-rate method are recommended for analyzing data of tumors observed in incidental, fatal and mortality-independent contexts of observation, respectively, in Peto et al. (1980). Misclassifications of incidental tumors as fatal tumors, or of fatal tumors as incidental tumors, will produce biased results. When a tumor is observed in a fatal context for a set of animals and is observed in an incidental context for the other animals in the experiment, data should be analyzed separately by the death-rate and the prevalence methods. Results from the different methods can then be combined to yield an overall result. The combined overall result can be obtained by simply adding together the separate observed frequencies, the expected frequencies, and the variances, or the separate T statistics and their variances.

The prevalence method, the death-rate method, and the onset-rate method use a normal approximation in the test for the positive linear trend or difference in tumor incidence rates. It is also well known that the approximation results will not be stable and reliable, and mostly tends to underestimate the exact p-values when the total numbers of tumor occurrence across treatment groups are small. In this situation, it is advisable to use the exact permutation trend test to test for the positive linear trend. The exact permutation trend test is a generalization of the Fisher's exact test to a sequence of $2 \times (r + 1)$ tables. The widely used prevalence method, the death rate method, and the onset rate methods for analyzing incidental, fatal, and mortality independent tumors, respectively, described in previous sections rely on good information about cause of death of tumors. There are situations in which investigators have not included cause of death information in their statistical analyses and electronic data sets.

To avoid the use of the cause-of-death information needed in the above Peto methods described in the WHO monograph, the Bailer–Portier poly-3 (in general poly-k) tests have been proposed for testing linear trends in tumor rates. These tests are basically modifications of the survival unadjusted Cochran–Armitage test for linear trend in tumor rate. The Cochran–Armitage linear trend test is based on a binomial assumption that all animals in the same treatment group have the same risk of developing the tumor over the duration of the study. This assumption is thus no longer valid if some animals die earlier than others.

The Bailer–Portier poly-3 test adjusts for differences in mortality among treatment groups by modifying the number of animals at risk in the denominators in the calculations of overall tumor rates in the Cochran–Armitage test to reflect “less-than-whole-animal contributions for decreased survival”. The modification is made by defining a new number of animals at risk for each treatment group. After weighting the pros and cons of the Peto methods and of the poly-k method, the FDA nonclinical statisticians have recently switched from the Peto methods to the poly-k method in their statistical reviews and evaluations of carcinogenicity studies of new drugs.

Interpreting results of carcinogenicity experiments is a complex process. Because of inherent limitations, such as the small number of animals used, low tumor incidence rates, and biological variation, a carcinogenic drug may not be detected (i.e. a false negative error is committed). Also, because of a large number of statistical tests performed on the data (usually 2 species, 2 sexes, 20–30 tissues examined, and 4 dose levels), there is a large probability that statistically significant positive linear trends or differences in some tumor types are purely due to chance alone (i.e. a false positive error is committed). Therefore, it is important that an overall evaluation of the carcinogenic potential of a drug should be made based on the knowledge of multiplicity of statistical significance of positive linear trends and differences, historical information, and other information of biological relevance.

In order to reduce the false positive rate, statistical reviewers in CDER use data of the concurrent control group(s) and historical control data to classify common and rare tumors, and adopt the following decision rule in their evaluation: A positive linear trend (dose–response relationship) is considered not to occur by chance of variation alone if the p-value is less than 0.005 for a common tumor, and 0.025 for a rare tumor. For the test of a pairwise increase in incidence rate, the significance levels of 0.01 and 0.05 are used, respectively.

To ensure that the committed false negative rate is not excessive, statistical reviewers collaborate with the reviewing pharmacologists, pathologists, and medical officers to evaluate the adequacy of the gross and histological examination of both control and treated groups, the adequacy of the dose selection, and the durations of the experiment in relation to the normal life span of the tested animals.

In negative studies, the statistical reviewers will perform a further evaluation of the validity of the design of the experiment, to see if there were sufficient numbers of animals living long enough to get adequate exposure to the chemical, and to be at risk of forming late-developing tumors, and to see if the doses used were adequate to present a reasonable tumor challenge to the tested animals. In Sect. 2.3, we

provide one example in animal study design to illustration the regulatory nonclinical statisticians' contribution in evaluating and developing more advance design of animal studies. A summary of the chapter is given in Sect. 2.4.

2.2 Example of CMC Methodology Development

As an example of regulatory methodology development and evaluation conducted by the CMC Statistics team at CDER/FDA, we will describe the statistical methodology development of dose content uniformity assessment for both small and large sample sizes during the last 10 years.

As one of the most important quality attributes for drugs, dose content measures the amount of the active ingredient of the product relative to the label claim (LC). For a therapeutic product, most of the dose contents should be within (85,115)%LC to ensure the homogeneity of the product. We can evaluate the content uniformity through the acceptance sampling, which is required by FDA to meet the quality standards for ensuring the consistency of the dose content with the label claim. The U.S. Pharmacopoeia (USP) publishes the dose content uniformity (DCU) sampling acceptance procedure and its revision for applicable to products seeking licensure in the US market every 5 years. The Europe Pharmacopoeia (EP) and Japan Pharmacopoeia (JP) frequently publish the DCU testing procedure used in Europe and Japan, respectively. The DCU procedure used in these three regions may be different due to the differences in quality requirements and statistical considerations. In the following, we outline the USP, EU, and JP testing procedures.

USPXXIV that was published by USP in 2005 recommended a two-tier sampling acceptance procedure as follows.

1st tier: A sample of 10 units is collected. The lot complies with the USP DCU requirement if the dose content of each unit is within (85 %, 115 %) LC and RSD (i.e. the sample standard deviation divided by the sample mean) is less than 6 %. If it fails to comply, we move to the 2nd tier.

2nd tier: Additional 20 units are randomly sampled and measured. The lot complies with the DCU requirement if the dose content of each of 30 units is within (75 %, 125 %) LC, no more than 1 unit has the dose content outside (85 %, 115 %) LC, and $RSD \leq 7.8$ %. It fails the DCU requirement otherwise. The requirement of none of the content of the 30 units is allowed to be outside (75 %, 125 %) LC is often referred to as zero tolerance condition.

EPIII defines the DCU test as a two-tier procedure similar to USPXXIV procedure but without requirement on RSD.

1st tier: 10 units are sampled. The lot complies with the DCU requirement if the dose content of each of all 10 units is within (85 %, 115 %) LC. It fails to comply if more than 1 unit has dose content outside (85 %, 115 %) LC or at least 1 unit has dose content outside (75 %, 125 %) LC. Otherwise, we move to the 2nd tier.

2nd tier: Additional 20 units are randomly sampled and measured. The lot complies with the DCU requirement if dose content of each of 30 units is within (75 %, 125 %) LC, and no more than 1 unit has dose content outside (85 %, 115 %) LC. It fails the DCU test otherwise.

JPXIV testing procedure is based on the tolerance limit and also has a zero tolerance requirement. JPXIV consists of two tiers as follows. 1st tier: 10 units are randomly sampled and tested. The lot complies if all 10 units' dose contents are within (75 %, 125 %) of LC and $85 \% < (\bar{x} - 2.2*s) < 115 \%$, where \bar{x} and s respectively are the sample mean and the sample standard deviation of the 10 units. It fails to comply if at least 1 unit has the dose content outside (75 %, 125 %) of LC. Otherwise, randomly sample 20 units more tested and move on to the 2nd tier.

2nd tier: The lot complies if all 30 units' dose contents are within (75 %, 125 %) of LC and $85 \% < (\bar{x} - 1.9s) < 115 \%$, where \bar{x} and s are the sample mean and the sample standard deviation of the 30 units, respectively. Otherwise, the lot does not comply with the DCU requirement.

In 2006, FDA CMC Statistics team evaluated the statistical properties of the USP, EU, and JP procedures outlined above with the operating characteristics curves. We proposed that a lot is accepted if only a small proportion of units in the lot have the dose contents below a lower specification or above an upper specification limit. Our proposed procedure is a two-tier sampling plan based on the two one-sided tolerance intervals. We further adapt Pocock's group sequential boundaries to control the confidence levels at the two tiers. Our proposed DCU testing procedure (Tsong and Shen 2007) is described as follows.

1st Tier: Sample 10 units, the lot is accepted if

$$\bar{x}_{11} - 115\% < A^{U_1} \text{ and } \bar{x}_{11} - 85\% > A^{L_1},$$

where $A^{U_1} = -C(\alpha_1)(s_1/\sqrt{10}) - s_1 Z_{0.9375}$, $A^{L_1} = C(\alpha_1)(s_1/\sqrt{10}) - s_1 Z_{0.9375}$; \bar{x}_1 and s_1 are the sample mean and the sample standard deviation of the 10 units, respectively. Otherwise, move to the 2nd tier and sample an additional 20 units.

2nd Tier: We accept the lot if

$$\bar{x}_2 - 115\% < A^{U_2} \text{ and } \bar{x}_2 - 85\% > A^{L_2},$$

where $A^{U_2} = -C(\alpha_2)(s_2/\sqrt{30}) - s_2 Z_{0.9375}$, $A^{L_2} = C(\alpha_2)(s_2/\sqrt{30}) - s_2 Z_{0.9375}$; \bar{x}_2 and s_2 are the sample mean and the sample standard deviation of the 30 units, respectively. Otherwise, we conclude that the lot fails to comply with the DCU requirement. We remove the zero tolerance requirement in the procedure by allowing a small probability of any sample falling outside of the zero tolerance limits (75 %, 125 %) LC under normality assumption.

In order to harmonize the acceptance sampling plans across the United States, Europe, and Japan regions, a harmonization procedure (U.S. Pharmacopoeia XXV

2010) was developed to replace USP XXIV and EP III plans. The two-stage harmonized procedure is derived based on a sequential procedure using the two-sided tolerance interval combined with an indifference zone for the sample mean and zero tolerance criteria for the observed dose content of each unit.

The two-stage harmonized procedure for DCU is described in (U.S. Pharmacopeia XXV 2010).

For the first stage, the sample mean (\bar{x}) and the sample standard deviation (s) of the 10 units are calculated. The indifference zone (M) at the first stage is defined as $M = 98.5\%$ if $\bar{x} < 98.5\%$; $M = 101.5\%$ if $\bar{x} > 101.5\%$; and $M = \bar{x}$ if $98.5\% \leq \bar{x} \leq 101.5\%$. The two-sided tolerance interval is calculated as $(\bar{x} - 2.4s, \bar{x} + 2.4s)$, where the constant 2.4 can be interpreted as the tolerance coefficient with approximately 87.5 % coverage and a confidence level of 90.85 % for a sample size of 10. The dose content uniformity is accepted if all 10 samples are within $(75\%, 125\%)M$ and $(\bar{x} - 2.4s, \bar{x} + 2.4s)$ is covered by $(M - 15\%, M + 15\%)$. If the lot fails to be accepted, go to the second stage.

In Stage 2, additional 20 samples are randomly collected. With a total of 30 samples, a tolerance coefficient of 2.0 is used for calculation. The constant 2.0 can be interpreted as the tolerance coefficient with 87.5 % coverage and 95.14 % confidence level for a sample size of 30. The lot is accepted if all 10 samples are within $(75\%, 125\%)M$ and $(\bar{x} - 2s, \bar{x} + 2s)$ is covered by $(M - 15\%, M + 15\%)$. Otherwise, the lot fails the DCU test.

FDA CMC Statistics team evaluated this harmonized USP procedure. Based on our evaluation, it is biased toward the lot with the true mean deviating from 100 % label claim. In other words, the probability of passing the lot with an off-target mean is higher than that of the lot with an on-target mean (100 % LC) based on simulations (Shen and Tsong 2011).

Since 2007, the pharmaceutical industry has expressed an interest in conducting large sample testing for dose content uniformity due to the availability of near-infrared spectroscopy in the manufacturing process. With this NIRS technology, continuously testing of the dose content without destroying the units becomes possible. Thus, the pharmacopeia acceptance sampling procedure for small samples should be extended to large samples. Many statistical testing procedures have been proposed for this purpose. The approach proposed by the FDA CMC Statistics team is a large sample DCU testing procedure based on the two one-sided tolerance intervals (TOSTI). Our proposed approach maintains a high probability to pass the USP compendia by restricting the TOSTI OC curves for any given sample size to intersect with the USP OC curve for a sample size of 30 at the acceptance probability of 90 % when the individual unit is assumed to be normally distributed with an on-target mean of 100 %LC (Shen et al. 2014). The derivation of the tolerance coefficient K for any large sample size n was provided in Shen et al. (2014), Dong et al. (2015), and Tsong et al. (2015). We denote this extension as PTIT_matchUSP90 method in the remaining section.

The large sample dose uniformity tests with two options were published in European Pharmacopeia (Council of Europe 2012). EU option 1 is a parametric approach based on a two-sided tolerance interval approach with an indifference zone and a

counting limit for the number of dosage units outside $75 \%M - 125 \%M$, with M defined the same way as that in the USP Harmonized procedure. EU option 2 is a non-parametric approach developed for non-normally distributed contents of dosage units. The EU option 2 is actually a counting procedure with two acceptance criteria. The number of dosage units outside of $[1 - L1, 1 + L1] M$ and $[1 - L2, 1 + L2] M$ are required to be no more than $C1$ and $C2$, respectively, where $L1, L2, C1$, and $C2$ are defined in Table 2.9.47.-2 of European Pharmacopeia (Council of Europe 2012).

FDA CMC Statistics team compared PTIT_matchUSP90 method with the two options of European Compendia under normality and mixture of two normal variables (Shen et al. 2014). In this work, we found that the two options of European Compendia give very different acceptance probabilities. In addition, the acceptance probabilities of both parametric and non-parametric options are higher than that obtained from our proposed PTIT_matchUSP90 procedure. Furthermore, these two EU options in European Pharmacopoeia 7.7 still have the same bias property as in the harmonized procedure recommended in USP XXV.

Here FDA CMC Statistics team compare the acceptance probability of the EU option 1 in European Pharmacopoeia Supplement 8.1 with the PTIT_matchUSP90 against the coverage within 85–115 %LC under the normality assumption. The results are shown in Fig. 2.1. As can be seen, the acceptance probability of EU

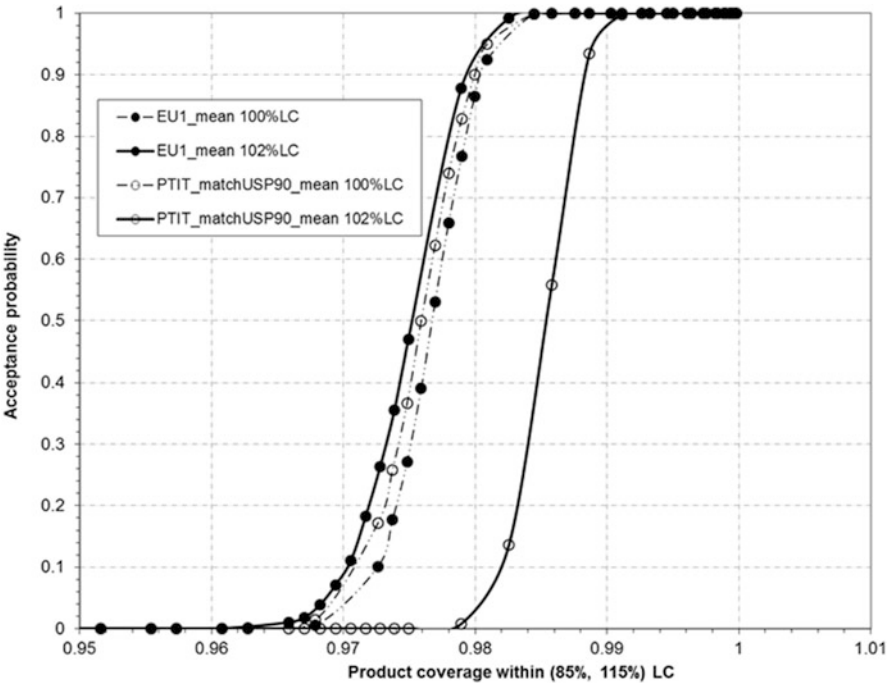


Fig. 2.1 Bias of European Union option 1 for individual dose content distributed as independent and identical normal variable with sample size $n = 1000$

Table 2.1 Comparison of the acceptance probability between the PTIT_USPmatch90 and EU option 2 method

Sample size, <i>n</i>	Acceptance probability	
	EU option 2	PTIT_USPmatch90
100	0.6458	0
150	0.5276	0
200	0.6047	0
300	0.455	0
500	0.3509	0
1000	0.2075	0

option 1 approach with a mean content of 102 %LC is higher than that with a mean of 100 %LC. We also compare the acceptance probability of the EU option 2 (nonparametric method) with those of the PTIT_matchUSP90 procedure when the individual unit dose content follows a uniform distribution in the range from 85 % to 115 % with 97 % probability and a value 84 % with 3 % probability in Table 2.1. For this particular distribution, the acceptance probability of USP harmonized DCU procedure is 3.72 % for a sample size of 30 units. Table 2.1 shows that the EU option 2 has acceptance probabilities higher than 50 % for sample sizes up to 200. The acceptance probability only reduces when the sample size is significantly larger than 1000. On the other hand, such a lot would have almost 0 % probability of passing USP DCU harmonized procedure and the PTIT_macthUSP90. The EU option 2 does not take the content variability into consideration and misses the purpose of dose content uniformity test. Further research on appropriate comparison is in progress.

2.3 Examples of Carcinogenicity Study Methods Development

One of the responsibilities of the Pharmacological—Toxicological Statistical Review Team is to keep track of new statistical methodologies developed in the area of animal carcinogenicity studies and works on the development of new or modified methodologies better suited for carcinogenicity data analysis. Following are two examples of the team’s research efforts.

Pairwise Comparisons of Treated and Control Group In the carcinogenicity data analysis routinely the treated groups are compared to the control group as primary or additional tests. For these pairwise comparisons, by convention only data from the selected two groups are used by ignoring data from the other dose groups. The test is termed as the unconditional test. Members of the Pharm–Tox Statistics team proposed two modifications to this conventional test. In the first modification, we proposed to use the data from all dose groups in variance calculation, in the spirit of test for contrasts of the general ANOVA analysis. The test is termed as the conditional test. It is shown that the asymptotic relative efficiency (ARE) of

conditional modification versus unconditional test is greater than 1. The second modification is to use the variance estimation method proposed by Hothorn and Bretz (2000). It is shown through simulation study that the second modification provides more power in both exact and asymptotic situations and has higher power than the unconditional test. Furthermore, the simulation results showed that asymptotically the conditional test always has more power than the second modified test. The detailed results were published in Rahman and Tiwari (2012).

Multiple Contrast Type Tests A typical animal carcinogenicity study involves the comparison of several dose groups to a negative control group for dose response relationship. The typical shape of the outcome (proportion of tumor bearing animals vs. dose levels) is assumed to be linear or approximately linear. However, in practice the shape of the outcome may turn out to be concave, convex or some other non-linear curve. The Cochran–Armitage (CA) test is most frequently used to test the positive dose response relationship. This test is based on a weighted linear regression on proportions. It is well known that the CA test lacks power for the nonlinear shape of the outcomes. For general shape of outcomes, Hothorn and Bretz (2000) proposed the multiple contrasts test (MC). This test suggests the use of the maximum over several single contrasts, where each of them is chosen appropriately to cover a specific dose response shape. In mathematical form the MC test statistic T^{MC} is defined as

$$T^{MC} = \max \{T_1^{SC}, T_2^{SC}, T_3^{SC}\},$$

where, MC = Multiple Contrast, SC = Single Contrast,

T^{MC} = Multiple Contrast test proposed by Hothorn and Bretz,

T_1^{SC} = Test with Helmert contrast $\tilde{c}^{(1)} = (-1, -1, \dots, -1, k)$, powerful for convex profiles,

T_2^{SC} = Test with Linear contrast $\tilde{c}^{(2)} = (-k, -k + 2, \dots, k - 2, k)$, powerful for linear profiles,

T_3^{SC} = Test with step contrast $\tilde{c}^{(3)} = (-1, -1, \dots, -1, 1, \dots, 1)$ for k odd, $\tilde{c}^{(3)} = (-1, -1, \dots, -1, 0, 1, \dots, 1)$ for k even, powerful for sub-linear profiles, and,

$$T_a^{SC} = \frac{\sum_i \frac{r_i}{n_i} c_i^{(a)}}{\sqrt{p(1-p) \sum_i \frac{(c_i^{(a)})^2}{n_i}}}, \text{ for the } i^{\text{th}} \text{ dose group, } r_i \text{ is the number of tumor}$$

bearing animals, n_i is the number of animals at risk, $c_i^{(a)}$ are the elements of $\tilde{c}^{(a)}$, p_i is the proportion of tumor bearing animals and p is the common value of p_1, p_2, \dots, p_k under null hypothesis.

Hothorn and Bretz compared their test with the CA test and concluded that the MC test on the average is more powerful than the CA test. A team member took interest in investigating this topic and found that the MC method performs well for

convex outcome, but not as good for concave outcome. He also proposed a new test method based on the maximum of sequential Cochran–Armitage (SCA) test over dose groups. In mathematical form SCA test statistic T^{SCA} is defined as

$$T^{SCA} = \max \{T_1^{CA}, T_2^{CA}, T_3^{CA}\},$$

where, T_1^{CA} = CA test with dose groups $\tilde{d}^{(1)} = (0, 1, 2, 3)$; T_2^{CA} = CA test with dose groups $\tilde{d}^{(2)} = (0, 1, 2)$; and T_3^{CA} = CA test with dose groups $\tilde{d}^{(3)} = (0, 1)$,

$$\text{and } T_a^{CA} = \sqrt{\frac{N}{r(N-r)}} \frac{\sum_i \left(r_i - \frac{n_i}{N}r\right) d_i^{(a)}}{\sqrt{\sum_i \frac{n_i}{N} (d_i^{(a)})^2 - \left(\sum_i \frac{n_i}{N} d_i^{(a)}\right)^2}}.$$

This new test has similar power as CA and MC tests for linear dose response, and has higher power for concave outcome. The MC test still has higher power for convex dose response.

In 2014, members of Pharm–Tox Statistics team evaluated the approaches with their interpretation of the results of a carcinogenicity experiment. If a significant linear dose response is found, the interpretation is straight forward. However, they found that the interpretation of other shapes of the dose response is difficult and somewhat subjective. Finding a highly statistically significant U-shaped or sign-curved dose response may not have any practical value. Therefore, finding a simple dose response like linear or at most quadratic is very important for practical purposes. The MC method has relatively simple models. However, it is a combination of several linear and non-linear models. If a significant dose response is found using the MC model, it is difficult to know what kind of dose response it is. On the other hand since the SCA methods are based on linear model, the interpretation of any findings from these tests is easy. Also, since the SCA test is based on the maximum of all possible CA tests, one additional advantage of SCA test is that it can capture a positive dose response, which may be present in a part of the data. For example in a concave data like 6, 10, 15, and 6 for control, low, medium and high groups, where a clear positive linear dose response is evident in the first three dose levels, the CA test shows a non-significant dose response ($p = 0.346$); however, SCA test still captures such dose response ($p = 0.035$).

The major criticism of both the HB and SCA methods is that the dose response shape is not known prior to the completion of the experiment. This is of serious concern which hinders the practical use of these methods. For this reason the agency has not yet adopted these methods as a part of their regular tumor data analysis. However, a post hoc analysis can still be performed, in relation to the observed the outcome pattern after the completion of the experiment.

A manuscript of the evaluation of SCA method has been submitted to the Journal of Biopharmaceutical Statistics for publication.

2.4 Summary

In summary, FDA regulatory nonclinical statisticians played many critical roles in review, regulation, and research for drug development and evaluation. Each of the two teams consists of five to seven members. They make significant contribution to the public health through assessing the carcinogenicity potential, product quality and product manufacturing control of each biopharmaceutical product seeking licensure in the US market.

References

- Council of Europe (2012) Uniformity of dose units (UDU) using large sample sizes, 7th edn. Chapter 2.9.47. of European Pharmacopeia 7.7. Report Pub Co Ltd, pp 5142–5245
- Dong X, Tsong Y, Shen M (2015) Equivalence tests for interchangeability based on two one-sided probabilities. *J Biopharm Stat* 24(6):1332–1348
- Hothorn LA, Bretz F (2000) Evaluation of animal carcinogenicity studies: Cochran–Armitage trend test vs. multiple contrast test. *Biometrical J* 42:553–567
- U.S. Pharmacopoeia XXV (2010) USP 905 uniformity of dosage units, Mack Printing Company, Easton
- Peto R, Pike MC, Day NE, Gray RG, Lee PN, Parish S, Peto J, Richards S, Wahrendorf J (1980) Guidelines for sample sensitive significance test for carcinogenic effects in long-term animal experiments. In: Long term and short term screening assays for carcinogens: a critical appraisal, international agency for research against cancer monographs, Annex to supplement. World Health Organization, Geneva, pp 311–426
- Rahman MA, Tiwari RC (2012) Pairwise comparisons in the analysis of carcinogenicity data. *Health* 4(10):910–918
- Shen M, Tsong Y (2011) Bias of the USP Harmonized test for dose content uniformity, Stimuli to the revision process, vol 37. Mack Printing Company, Easton
- Shen M, Tsong Y, Dong X (2014) Statistical properties of large sample tests for dose content uniformity. *Therapeutic Innovation Regulatory Sci* 48(5):613–622
- Tsong Y, Shen M (2007) Parametric two-stage sequential quality assurance test of dose content uniformity. *J Biopharm Stat* 17:143–157
- Tsong Y, Dong X, Shen M, Lostritto RT (2015) Quality assurance test of delivery dose uniformity of multiple-dose inhaler and dry powder inhaler drug products. *J Biopharm Stat*. doi:[10.1080/10543406.2014.972510](https://doi.org/10.1080/10543406.2014.972510)
- Wald A, Wolfowitz J (1946) Tolerance limits for a normal distribution. *Ann Math Stat* 19:208–215

Nonclinical Statistics for Pharmaceutical and
Biotechnology Industries

Zhang, L. (Ed.)

2016, XXII, 698 p. 144 illus., 113 illus. in color.,

Hardcover

ISBN: 978-3-319-23557-8