PMI SCIENCE
PHILIP MORRIS INTERNATIONAL

STUDY DESIGN PRINCIPLES IN

SYSTEMS TOXICOLOGY

Table of Contents

1 Introduction

Standard and systems toxicology are key to assess the potential of RRP aerosols in reducing toxicity compared to cigarette smoke. They both rely heavily on using rigorous scientific methodologies to achieve their goals. The scientific methodology (Kosso 2011) generates knowledge in a systematic, rigorous manner based on the following steps: observing systems and phenomena to identify questions and form hypotheses; gathering information and knowledge to formulate plausible and testable hypotheses under uncertainty; developing and employing appropriate measurement systems to generate data; and testing hypotheses to draw conclusions. To achieve these objectives, sound study design and appropriate statistical methods are required to draw accurate conclusions. These methods must meet the highest standards, be accepted by the scientific community, and align with regulatory requirements (e.g., United States Food and Drug Administration; FDA).

The typical steps in an *in vivo* study in systems toxicology are depicted in **Figure 1**. If studies are not appropriately designed and/or if the methodologies used are not suitable, then it can be difficult to analyse the data, test the hypothesis and draw conclusions.

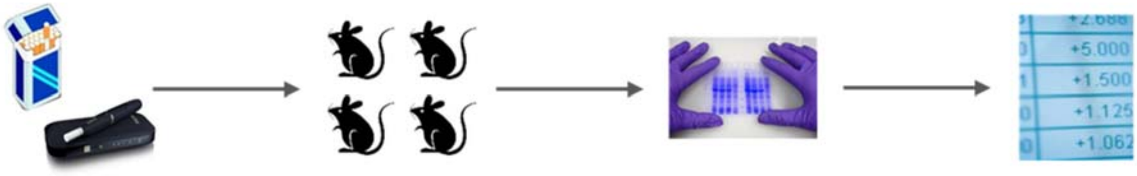

Figure 1. Typical steps in an *in vivo* toxicological study. Animals are exposed to different treatments, for example smoke or aerosol generated from different products. Biological samples are then analyzed using validated bioassays to quantify endpoints of interest and generate data for further statistical analyses.

This document provides a brief and concise introduction to study design principles used in Systems Toxicology studies conducted at the Research & Development site of Philip Morris International (PMI R&D) (Sections 2 and 3). Study design has a direct consequence on the statistical methods that will be used for data analysis and hypotheses testing. Basic principles behind statistical methods for data analysis are summarised in Section 4.

This is not a comprehensive document and is intended to be read in conjunction with relevant literature. As such, key references on study design principles and related statistical design methodologies are provided throughout as citations. Extensive presentation of statistical methodologies, description of good laboratory practices, exposure system characterization, assay development and method validation topics are considered out of scope.

2 Experimental design in Systems Toxicology

In the context of toxicological assessment biological systems (cells, tissues, animal models, etc.) are used to study the potential toxicity of chemicals. Such experiments seek to explore the impact of targeted system perturbations on the biological systems under study, as illustrated in **Figure 2** The resulting impact is generally quantified based on appropriate measured endpoints ($Y_{1,} Y_{2,} \dots, Y_{m,}$). System perturbations may result from controlled stimuli (experimental) or uncontrollable system parameters. Controllable experimental factors (commonly denoted by X) may directly impact the endpoints by either acting alone or in combination with a simultaneous change in one or more other factors. The latter are known as interactions and are key to understanding the system being studied. Factorial designs allow estimation of such interactions using experimental design techniques. Uncontrollable factors (commonly denoted by Z) represent system variations and can be partially controlled using experimental design techniques, such as blocking. Combination of the two factor groups (X and Z) constitute the design space of the systems being evaluated.

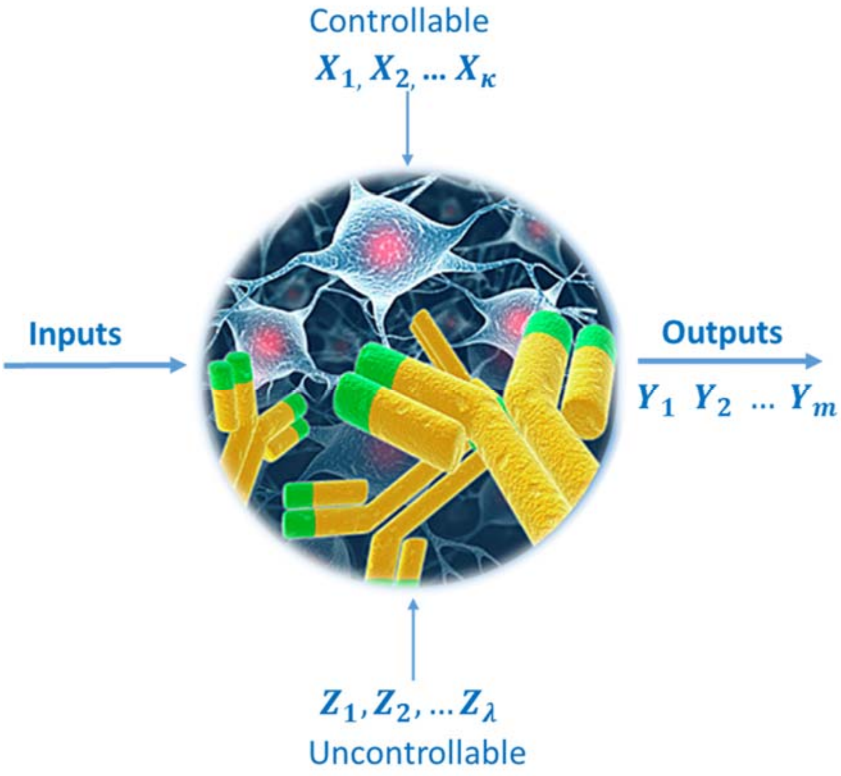
Controllable
$X_{1,} X_{2,} \dots X_{\kappa}$
Inputs
Outputs
$Y_1 \; Y_2 \; \dots \; Y_m$
$Z_1, Z_2, \dots Z_\lambda$
Uncontrollable

Figure 2. General experimental design in Systems Toxicology. Stimuli (controllable experimental factors) and uncontrollable factors result in variations in the system outputs. Experimental design methodologies allow for the impact of design factors on the output to be explored and quantified using statistical modelling, including both X and Z and how they relate to Y.

The ultimate goal of an experiment is to assess the impact of selected factors or treatments (factorial combinations within the design space) on the endpoints under study, to test hypotheses and to predict risks.

In systems toxicology research, changes in human organotypic tissue culture[1] induced by cigarette smoke exposure are evaluated and compared to those induced by RRP aerosol exposure. In *in vivo* studies, animals are exposed to either cigarette smoke or RRP aerosol, and exposure effects on various biological endpoints are compared. Hence, selected factors or treatments (T) are assigned to experimental units (U) in order to assess the impact on resulting endpoints (Y), as illustrated in **Figure 3**.

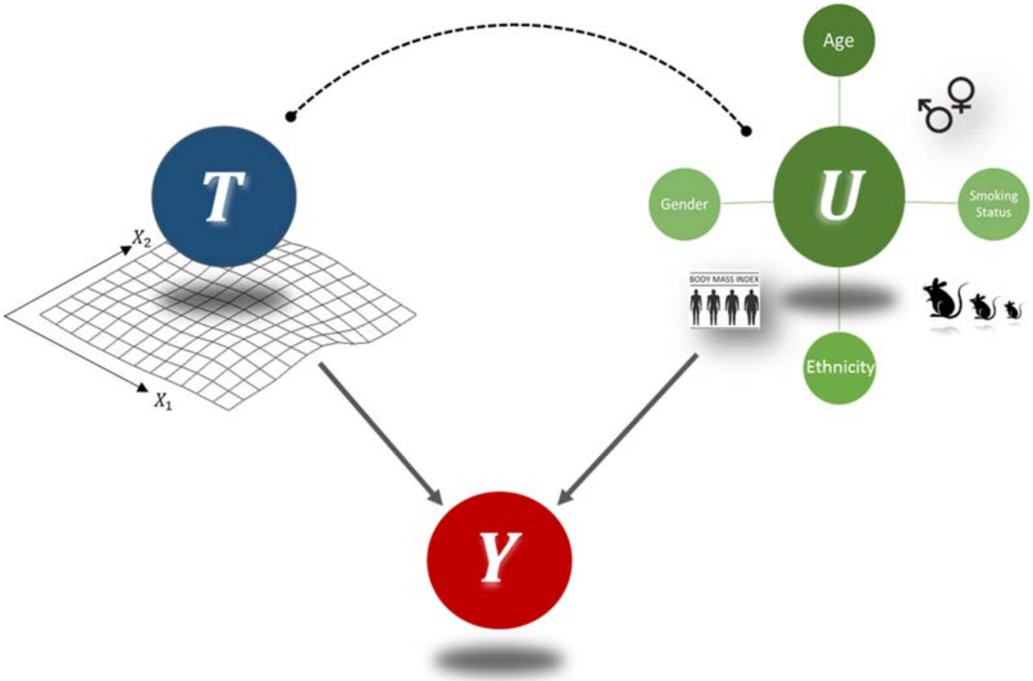

Figure 3. Experimental design schematic overview: Treatments (T) are assigned to experimental units (U) and endpoints (Y) are measured to assess the impact of T on Y. Experimental units are assigned treatments at random (dotted black line) to eliminate potential dependence between treatment and the units. This ensures that any observed link between Y and T is not due to joint dependence or a confounding influence of U.

[1] *in vitro* organotypic tissue cultures are grown in a three-dimensional (3-D) environment versus standard two-dimensional (2-D) dishes. 3-D culture systems are biochemically and physiologically more similar to *in vivo* tissues.

2.1 Test systems and endpoints

The success of an experiment relies on accurate, precise and repeatable test systems that are fit for a specific purpose. Availability or development of such test systems, methods and assays is critical in the lifecycle of a study and may influence the feasibility and objectives of an experiment.

Test systems quantify the targeted outcomes, often called endpoints. These should be biologically relevant and suitable to address the study objectives. They should be furthermore measured with high precision in order to respond the study hypotheses with an acceptable level of confidence. To do so appropriate measurement and test systems have to be used and, if needed, assays developed and optimized. Design of experiment (DOE) methodologies, including screening, optimization and robustness designs, are useful in developing and validating assays and test systems. Standard experimental methodologies, such as (fractional) factorial designs, response surface methods and repeatability and reproducibility (Montgomery 2009), are often used.

Endpoints are typically recorded in numeric formats[2] and may be either qualitative or quantitative. The majority of endpoints recorded in systems toxicology studies are quantitative as continuous and count data. Sometimes qualitative data may also result from animal studies. The statistical analysis methods are specific to the type of endpoints being analysed.

Numerous endpoints relating to cell viability, inflammation as well as differential gene expression in microarray analyses, are recorded during *in vitro* studies. Adenylate kinase (AK)-based cytotoxicity, cytochrome P450 (CYP) 1A1/1B1 activity, tissue histology, concentrations of secreted mediators in the basolateral media and transcriptomes are evaluated following exposure to air/aerosol/smoke (Iskandar et al. 2015). In *in vivo* studies, the registered endpoints include gene expression and proteomics as well as a series of biological endpoints that are specified in (OECD 2009a). For more information, refer to (Kogel et al. 2016).

2.2 Treatments, products and controls

Treatments in systems toxicology research are commonly time-varied exposure to smoke and aerosol resulting from 3R4F[3] (reference) and RRP (test) products. In the *in vitro* studies smoke and aerosol exposure is typically set to different levels (exposure durations) during the study design preparation while multiple post-exposure sample collection time points are defined. The same holds also for the *in vivo* studies, where additional treatment groups are designed. These include cessation and switching groups in which animals are removed from the exposure (cessation) or are switched from one treatment category to another (switching). In dose-ranging finding studies, different doses of the test item are used. The number, range and spacing of dose levels are critical considerations in the study design.

Given the comparative nature of systems toxicology studies, appropriate reference and control groups should be used. The control group can be either an untreated or a vehicle control group. The vehicle control is treated in exactly the same manner as the treatment group (the test item) except that the

[2] May include photos or videos of tissues in histopathology cases, which can be digitized and quantified. Visual inspection can lead to a single subjective judgment, which is in turn treated as a numeric index.
[3] The Kentucky Reference Cigarette. For more information see: https://www.coresta.org/sites/default/files/pages/tji0213-p150-154-refproducts.pdf.

treatment is not applied. Untreated and vehicle controls may be identical. In such cases, they can both be used to assess the degree of variability in the negative control and provide a better basis to address the biological importance of any observed effects. Without a control group, it would be difficult to determine treatment effects.

The reference group is a treatment group in which a reference treatment is applied and compared to other treatments. A typical example of a reference treatment is exposure to smoke from a 3R4F cigarette. A typical example of an untreated control group is the sham, which corresponds to fresh air exposure. For an illustration of treatment and control groups in a standard *in vivo* inhalation/cessation study, see **Figure 4**.

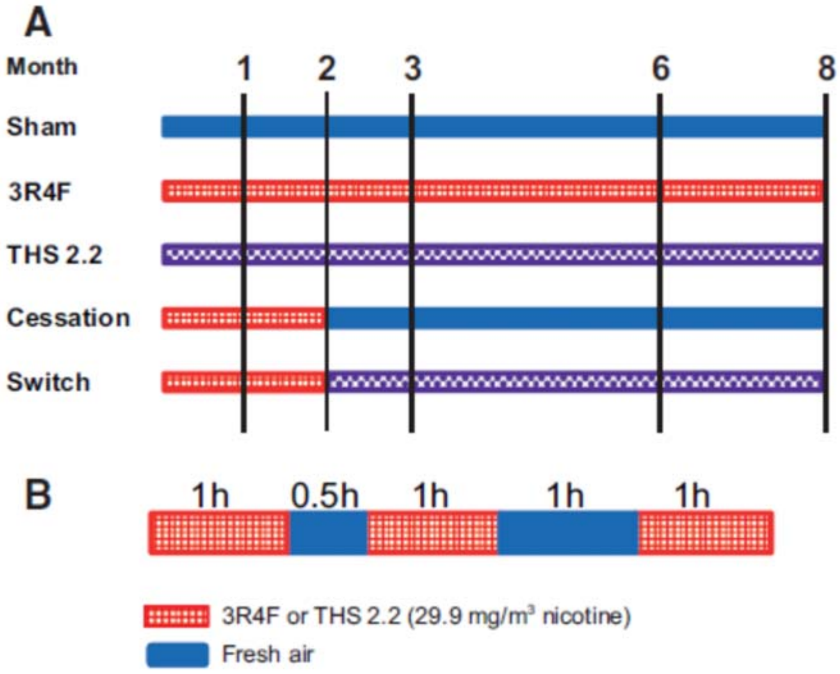

Figure 4. Treatment design in an *in vivo* inhalation/cessation study. (A) Treatment groups and controls are aligned across study duration (in months). Samples are collected at months 1, 2, 3, 6 and 8. (B) Daily exposure schedule. *Color codes:* Red indicates exposure to 3R4F, blue indicates exposure to fresh air and violet illustrates exposure to the tobacco heating system THS 2.2.

2.2.1 Other design factors

Treatments and controls may be administered in different sequences. Observation and analysis may also occur in a time-dependent manner. For instance, exposure duration of cells or tissues to smoke/aerosol and/or post-exposure duration are design factors commonly used in organotypic studies, resulting in a grid of observed points for each treatment group; see the grid spanned by X_1 and X_2 in **Figure 3**. Other design factors may also be used, such as exposure systems (e.g., modular whole body exposure chambers versus nose-only exposure) in *in vivo* studies. Given that inclusion of additional design factors will increase the complexity of the study, factors should be carefully selected. Pilot and/or screening studies using fractional-factorial designs and experience is helpful at this stage.

2.3 Experimental units

Experimental units are defined as the units that are randomly assigned to treatment groups. These units (e.g., animals, tissue cultures) may have properties, such as age or body weight, that are known or expected to impact the endpoints investigated; this is illustrated by the arrow linking U to Y in **Figure 3**. In such cases, experimental units can be stratified into homogeneous subgroups defined by gender, weight or cohort, for example. In an organotypic study, the experimental units are tissue cultures that are exposed to different treatments. They may be grouped according to tissue type, cell donor or even tissue batch. Experimental units are not always identical to sampling units, which are the objects measured in the experiment and are commonly subunits of experimental units. Sampling units are sometimes referred to as observational units. For example, in an animal study, the experimental units are the animals exposed to smoke, while the sampling units may be BALF that is sampled and analysed to measure a specific endpoint. The distinction between experimental and sampling units is critical and relates to replication, which will be discussed in section 3.4. Experimental units may have similar features with respect to other characteristics observed and measured during the experiment. These features may be used once the experiment is over to post-stratify the experimental units or adjust the observed endpoints. This becomes beneficial when the observed characteristics correlate well with the targeted endpoints. This is further discussed in section 3.3.

3 Study Design principles

3.1 Study types and objectives

Setting specific, measurable and achievable study objectives is important for proper study design and takes place prior to any study design specification. Based on the primary study objectives, two main study types, exploratory and assessment, may be established.

Exploratory studies include pilot, screening and optimization studies aimed at evaluating the performance of the test system, suitability and precision of the measured endpoints and validity of the developed assays. Exploratory studies also include dose-response experiments that determine suitable dose and time exposure levels to achieve the targeted toxicity effects and experiments to select appropriate negative and positive assay controls.

Assessment studies investigate the impact of the treatments on the selected endpoints. These investigations can be either exploratory or confirmatory[4]. In both cases, estimation of the effect of each treatment on the outcomes and comparisons between test, reference and control parameters are of primary interest.

In sections 3.2, 3.3 and 3.4 we present three fundamental study design principles.

[4] Confirmatory studies are well-designed trials towards the final phase of experimental research in which the goal is to confirm specific hypotheses.

3.2 Randomization

Randomization allows experimental units to be randomly assigned to study treatments. Each experimental unit is given a known (usually equal) probability to receive each treatment, although the treatment to be administered cannot be predicted with certainty. Randomization is important because it removes any selection or investigator bias in the process of treatment allocation. Randomization ensures that observed and significant differences between study groups are due to treatment effects and not to uncontrolled covariates or confounders[5]. In an *in vivo* study, randomization would prevent all animals in a single cohort from receiving the same treatment. Then, any observed treatment effect cannot be dissociated from that cohort. Randomization can be visualized by breaking the link (the dotted line in **Figure 3**) between experimental units (U) and treatments (T). Randomization can be carried out in a number of different ways, and different randomization schemes result in different statistical designs. Randomization can be completely at random (Completely Randomized Design) or can be applied within smaller units called strata or blocks (Completely Randomized Block Design).

3.3 Blocking and stratification

Blocking and stratification allow for smaller and more homogeneous experimental unit groups to be formed, thereby correcting for baseline differences in known covariates (e.g. animal body weight or age). The main purpose of blocking techniques is to better control for random variations that are not under control of the investigator. Blocks or strata restrict the randomization process, but create homogeneous experimental subgroups in which treatment effects may be easier to detect and estimate with higher precision.

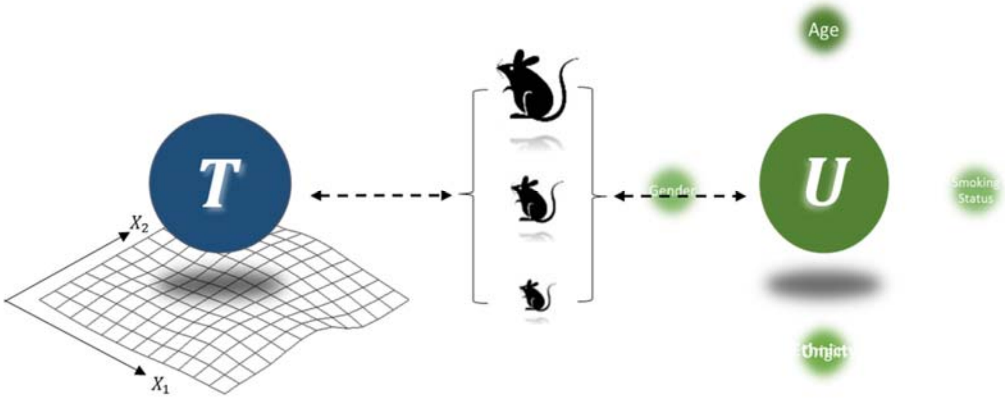

Figure 5. Randomization in blocks. Treatments are randomly assigned to animal groups with similar body weights.

There are several ways of blocking in systems toxicology studies. In *in vivo* studies, treatments are randomly assigned to animal subgroups with similar body weights, as illustrated in **Figure 5**. Animals are typically of a single gender and were raised in a single cohort. For *in vitro* experiments, tissues

[5] A confounding variable is one that is so closely related to another factor in the design that the individual contributions to an effect cannot be separated.

originating from the same batch are usually exposed to smoke/aerosol. Many different blocking factors may be used in the study design phase, and different blocking sources can simultaneously be controlled using existing design techniques, such as Latin squares and related designs (Cochran 1957). Given the modern instrumentation used during quantification of targeted endpoints (e.g., 96-well plates), blocking at the plate level can efficiently control for random plate-to-plate variations (Casella 2008).

3.4 Replication and sample size determination

All observed experimental phenomena are stochastic, and the resulting observations are variable. Variations may be due to (i) differences in how the investigator applies treatments, (ii) uncontrollable factors, (iii) errors[6] resulting from randomization, or (iv) other sources of technical error, such as measurement error and instrumental noise. Sampling and statistical errors result from randomization and sampling techniques. All of these variations can attenuate treatment effects and invalidate statistical testing and estimation. To overcome these issues, it is important that experiments are replicated.

Replicating an experiment increases confidence in the results and reduces the influence of confounding variables, while allowing estimation of other random variations (e.g., batch-to-batch variability) with improved precision. Studies should be adequately replicated and properly controlled for different sources of error. The organotypic studies in Systems Toxicology group, for instance, are run in three main phases, which are considered true experimental repeats. Each main phase includes new sample preparation using fresh tissues and new smoke/aerosol generation. The ultimate goal of replication is to quantify the experimental error and define reproducibility of the study results. Replication increases the sample size, which in turn reduces the standard error[7] of the estimated parameters. The final sample size should be sufficiently large to obtain statistical estimates with a given level of precision[8] and to detect a true treatment effect of a given size (magnitude) with a given probability when treatment groups are compared. Therefore, increasing the sample size strengthens the statistical power associated with the tested hypotheses; see section 4.3.

Experimental replication results from an increase in the number of experimental units and, consequently, sample size. This is not necessarily identical to an increase in sampling or measurement units, and *experimental replication* should be distinguished from *technical replication*. To increase the sample size and replicate the experiment, the number of experimental units should be increased, but not necessarily the number of sampling units.

4 Statistical data analysis considerations

Statistical methods are used to analyse experimental data generated from designed experiments. The statistical methods to be used in the data analysis phase are largely dependent on the study design (Casella 2008). Analysing the data under a complete randomized design will not be identical to data analysed under a randomized block design. Moreover, the statistical methodologies employed are often

[6] 'Error' is used in a statistical context as the deviation of an observed result from its true value.
[7] Estimates from small and highly variable populations may be associated with high standard error values. High standard error values may also be due to incorrect and/or insufficient sampling and study design, as well as inefficient estimation (see paragraph 5.3).
[8] Increased precision of the resulting estimate leads to reduced standard error (uncertainty).

different based on study type. Key statistical analysis considerations commonly used across all study types and designs are described below.

4.1 Sample versus population

Statistics can be used to generate a better understanding of populations of interest (e.g., human tissue or cells) by analysing samples from population units. In systems toxicology, samples and data are studied in order to understand populations. The bridge between these two is the sampling process. The sampling process defines the error related to sample selection and should be taken into account during data analysis.

4.2 Data transformations and scaling

The raw data collected are not necessarily the data used for statistical analyses. Raw data may be recorded per sampling unit (e.g., instrument reads) rather than per experimental unit (e.g., per animal). Moreover, based on the stochastic phenomena under study, raw data may not fit the right scale; data are then transformed and rescaled. It is quite common in Systems Toxicology to measure cytotoxicity in a relative rather than in an absolute scale. Cytotoxicity can then be expressed as a percentage computed based on the negative and positive controls[9]. The use of a relative scale to study and interpret most of the toxicological endpoints implies the use of the logarithms; this is very often the case in all biosciences and consequently in toxicology (Limpert 2001) (Limpert 2017). Last but not least, data transformation and scaling may be implied by distributional and error homoscedasticity assumptions related to statistical modelling techniques. In such cases, the investigator may conclude that it would be beneficial to transform the data prior to statistical data analysis and estimation.

4.3 Statistical estimation and testing

Statistics are data functions computed from samples. These functions capture information and estimate data characteristics at the population level under several optimality properties (Casella 2002).

Estimates should be *unbiased* and *precise*. In real life applications, the data generated often include outlying observations. These outliers can inflate variances and bias the final estimate. *Robust* estimates should then be used. Statistical estimates are used to infer population characteristics and to provide a measure of confidence. Confidence intervals are established and statistical hypotheses are tested. A

[9]

$$Cytotoxicity\ (\%)\ = \frac{AK_{tissue} -\ AK_{NegCTRL}}{AK_{PosCTRL} -\ AK_{NegCTRL}}$$

for $AK_{PosCTRL},\ AK_{NegCTRL}$ denoting the AK outcome for positive and negative controls, respectively.

hypothesis allows for answering questions such as: *Is the observed difference between toxicity resulting from exposure to RRP aerosol compared with exposure to 3R4F smoke statistically significant?*

A statistical hypothesis is therefore a statement about a targeted parameter or effect at the population level. Testing hypotheses involves two complementary hypotheses, the null and the alternative, denoted as H_0 and H_a, respectively. The null hypothesis reflects a state of ignorance on the parameter or the effect under study. For example, a null hypothesis of the form: *'The exposure effect on human organotypic cultures of RRP aerosol does not differ from that of a similar exposure to 3R4F cigarette smoke'* could be used in assessment studies. Rejection of the null hypothesis leads to accepting the alternative hypothesis, which may be a two-sided (not equal) or one-sided (less or greater) alternative. To decide if the null hypothesis should be rejected, a properly defined test-statistic is used. Statistical significance is declared whenever the probability of the observed test-statistic under the null hypothesis[10] is below a predefined threshold, called the *significance level*. In systems toxicology studies, this level is commonly set at equal to or less than 5% and is known as the *type-I error rate*. This rate reflects the probability of erroneously rejecting a valid hypothesis. The probability of not erroneously rejecting a false hypothesis is known as a *type-II error rate*. The *statistical power* of a test is determined when this rate is subtracted from 1. Hence, statistical power is a measure of correctly rejecting a non-valid hypothesis, or the chance to correctly identify a true effect. In order to increase the power, one needs to increase the sample size (see section 3.4) or appropriately block and/or stratify experimental units to gain precision (see section 3.3). The two error types and related statistical power are illustrated in **Figure 6**. Further insights on statistical testing and confidence intervals are described in (Casella 2002).

High-throughput data in systems toxicology raise many issues related to hypothesis testing and estimation. Unbiased estimates often do not exist, and when they do exist, they are highly variable. In such cases, minimum mean squared error estimates are preferred as a trade-off bias for variance reduction. In hypothesis testing, multiplicity testing should always be taken into account, and the family-wise error rate or the false discovery rate should be controlled. Sample sizes then offer limited statistical power given the correction for multiple testing. Numerous solutions to this issue have been proposed. For a discussion, see (Mehta 2004) and associated references therein.

[10] The well known p-value.

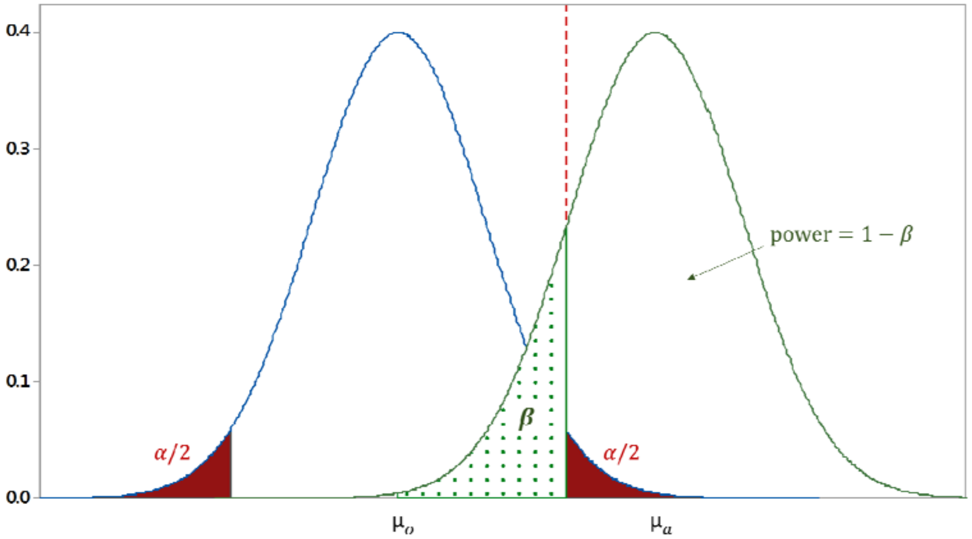

Figure 6. Statistical testing. Statistical power and error types in a simple two-sided alternative hypothesis testing setting. μ_0 and μ_a denote the mean under the null and alternative hypotheses H_0 and H_a, respectively.

4.4 Statistical models

Statistical data analysis in standard toxicology focusses on estimating treatment effects and testing their significance. In the past, toxicological studies reported the size of observed effects when comparing treatment to control groups and reported p-values. Statistically significant effects are traditionally identified by asterisks when p-values are below the 5% significance level. Nowadays, statistical models add a new dimension in systems toxicology research. They provide a generalized framework for dealing with standard toxicology objectives, such as the estimation and testing of the treatment effects, while allowing for prediction and classification. Statistical models are probabilistic models that incorporate biological knowledge of the mechanisms being studied, while accounting for uncertainties regarding mechanisms related to uncontrolled factors and other sources of variation. Given recent developments and improvements in computational tools for statistical modelling and simulation, the resulting models are more relevant and can be used in broader experimental settings. Nevertheless, the complexity of such models is constantly increasing, and study design principles remain relevant and critical for building realistic, trustworthy models (Mehta 2004); (Johnstone 2009).

5 Other Study Design Considerations

5.1 Standards and ethical considerations

PMI's goal is to develop a portfolio of new product alternatives to cigarettes. To this end, a comprehensive, rigorous assessment strategy inspired by standard methods used in the pharmaceutical industry has been implemented in alignment with the U.S. FDA draft guidance for Modified Risk Tobacco Product (MRTP) applications (FDA 2012), Good Laboratory Practice (GLP), and Good Clinical Practice

(GCP) guidelines. The Organization for Economic Co-operation and Development (OECD) series on the principles of GLPs can be viewed at: http://www.oecd.org/chemicalsafety/testing/oecdseriesonprinciplesofgoodlaboratorypracticeglpandcompliancemonitoring.htm, in addition to OECD guidelines related to *in vivo* testing (OECD 2009a, b).

In the *in vivo* studies small rodents are used, and the facility where we conduct these studies is accredited by the Association for Assessment and Accreditation of Laboratory Animal Care International (AAALAC), which ensures the highest level of care and humane treatment. The PMI Animal Welfare Committee reviews all proposed animal studies to evaluate whether study objectives are achievable or whether the objectives can be achieved through other testing means. We also have an Institutional Animal Care and Use Committee, which further evaluates experimental procedures in detail and makes sure that any opportunity to apply the *Replace, Reduce and Refine* strategy (Russell 1959) is implemented.

6 Summary

Standard and systems toxicology approaches are used to assess the potential of RRP aerosols in reducing toxicity compared to cigarette smoke. Both methods rely heavily on conducting properly designed experiments using advanced instrumentation and test systems while using rigorous scientific methodologies. Experimental goals in systems toxicology are two-fold: to explore the impact of exposure to RRP aerosol on human and animal cells and tissues and to assess differences in resulting toxicity between exposure to RRP aerosols and to 3R4F cigarette smoke.

Experiments designed in systems toxicology have clear, specific study objectives reported in study plans. System perturbations from exposure are investigated by measuring a set of reliable, suitable endpoints using advanced test systems and bioassays. Exploratory studies (e.g., Dose Ranging Finding studies) are very useful at this stage to optimize test systems, explore the toxicity effect under different experimental conditions and dose levels and to initially evaluate the precision of final measurements.

Assessment studies are conducted to evaluate the potential of RRP aerosols to reduce toxicity compared to cigarette smoke. Designing an assessment study consists of properly defining treatment, reference and control groups and allocating experimental units to different study groups. Treatment groups are commonly RRPs (test item) and 3R4F (reference item), while control groups are commonly the sham (control) and/or the vehicle. In *in vivo* studies, switching and cessation groups can also be included. Randomization techniques are used to assign treatments to experimental units to avoid bias and to ensure that any observed significant differences between study groups are due only to treatment effect. Randomization is generally performed inside small, homogeneous groups of experimental units and under homogeneous experimental conditions, or blocks.

If the study focuses on estimation of a treatment effect (e.g., toxicity during a Dose Ranging Finding study), then the precision of the final estimate is anticipated and the sample size is defined in order to achieve desired precision levels. If the assessment study is testing hypotheses on observed treatment effects, then the expected effect sizes should be known (or at least partially known) in advance from previous studies or from existing prior knowledge. Designing the study then includes proper selection of sample size (replication of the experiment) to achieve sufficient statistical power for testing hypotheses. When multiple tests are considered, proper corrections for multiple testing are anticipated to control for the final error rate. In toxicological assessment studies where the goal is still exploratory, adjustments for multiplicity are not made unless the analysis is related to high-throughput data, such as gene expression.

Statistical estimates, tests and models depend on randomization and blocking techniques used during the study design phase. Therefore, statistical design and analysis are intimately connected. Data are often rescaled and transformed. The need to do so is often implied either by the endpoint under consideration (e.g., % cytotoxicity resulting from AK measurements) or by the measurement system and the process of data generation (e.g., quantification based on relative areas are ratios and are usually log-normally distributed, therefore data are transformed to a logarithmic scale). Use of different data scales may be due to how data are going to be interpreted at the end of the study (e.g., ratio scales are used for interpretation rather than interval scales) and the way statistical models are parametrized.

Statistical methods in systems toxicology are not only limited to estimating and testing hypotheses. They are also used to predict and classify multi-parameter and multivariate phenomena related to the investigated systems. High-throughput data applications generate a large amount of information, which is then used to further investigate the biological mechanisms of the systems under study. In all cases, statistical models should avoid bias, attain high levels of precision and ensure robustness. In high-throughput data applications, biased estimates are sometimes less variable and are therefore preferred to unbiased estimates. Statistical models reflect the existing biological knowledge of the mechanisms under study to the extent possible. Statistical design is maintained and used to generate experimental data and is parametrized such that it leads to straightforward, interpretable conclusions.

7 References

Casella, G. 2008. Statistical Design. Springer.

Casella, G. and Berger, R.L. 2002. *Statistical Inference*. 2nd ed: Duxbury.

Cochran, W.G. and Cox, G.M. . 1957. *Experimental Designs*: Wiley.

FDA. 2012. Modified Risk Tobacco Product Applications.

Iskandar, Anita R., Yang Xiang, Stefan Frentzel, Marja Talikka, Patrice Leroy, Diana Kuehn, Emmanuel Guedj, Florian Martin, Carole Mathis, Nikolai V. Ivanov, Manuel C. Peitsch, and Julia Hoeng. 2015. "Impact Assessment of Cigarette Smoke Exposure on Organotypic Bronchial Epithelial Tissue Cultures: A Comparison of Mono-Culture and Coculture Model Containing Fibroblasts." *Toxicological Sciences* no. 147 (1):207-221. doi: 10.1093/toxsci/kfv122.

Johnstone, I.M. and Titterington, D.M. 2009. "Statistical challenges of high-dimensional data." *Phil. Trans. R. Soc. A* no. 367:4237–4253. doi: 10.1098/rsta.2009.0159.

Kogel, U., B. Titz, W. K. Schlage, C. Nury, F. Martin, A. Oviedo, S. Lebrun, A. Elamin, E. Guedj, K. Trivedi, N. V. Ivanov, P. Vanscheeuwijck, M. C. Peitsch, and J. Hoeng. 2016. "Evaluation of the Tobacco Heating System 2.2. Part 7: Systems toxicological assessment of a mentholated version revealed reduced cellular and molecular exposure effects compared with mentholated and non-mentholated cigarette smoke." *Regulatory Toxicology and Pharmacology* no. 81:S123-S138. doi: 10.1016/j.yrtph.2016.11.001.

Kosso, P. 2011. *A Summary of Scientific Method*. Vol. 1, *SpringerBriefs in Philosophy*: Springer.

Limpert, E. and Stahel, W. A. 2017. "The log-normal distribution." *Significance* no. 14 (1):8-9. doi: 10.1111/j.1740-9713.2017.00993.x.

Limpert, E. and Stahel, W.A. and Abbt, M. 2001. "Log-normal Distributions across the Sciences: Keys and Clues." *BioScience* no. 51 (5):341-352.

Mehta, T. and Tanik, M. and Allison, D.B. 2004. "Towards sound epistemological foundations of statistical methods for high-dimensional biology." *NATURE GENETICS* no. 36 (9):953-947. doi: 10.1038/ng1422.

Montgomery, Douglas C. 2009. *Design and Analysis of Experiments*. 9th ed: Wiley.

OECD. 2009a. *Test No. 412: Subacute Inhalation Toxicity: 28-Day Study*: OECD Publishing.

OECD. 2009b. *Test No. 413: Subchronic Inhalation Toxicity: 90-day Study*: OECD Publishing.

Russell, W.M.S. and Burch, R.L. and Hume, C.W. 1959. "The principles of humane experimental technique." In. http://altweb.jhsph.edu/pubs/books/humane_exp/het-toc.