Error Analysis and Propagation in Metabolomics Data Analysis

Hunter N.B. Moseley [a,*]

Abstract: Error analysis plays a fundamental role in describing the uncertainty in experimental results. It has several fundamental uses in metabolomics including experimental design, quality control of experiments, the selection of appropriate statistical methods, and the determination of uncertainty in results. Furthermore, the importance of error analysis has grown with the increasing number, complexity, and heterogeneity of measurements characteristic of 'omics research. The increase in data complexity is particularly problematic for metabolomics, which has more heterogeneity than other omics technologies due to the much wider range of molecular entities detected and measured. This review introduces the fundamental concepts of error analysis as they apply to a wide range of metabolomics experimental designs and it discusses current methodologies for determining the propagation of uncertainty in appropriate metabolomics data analysis. These methodologies include analytical derivation and approximation techniques, Monte Carlo error analysis, and error analysis in metabolic inverse problems. Current limitations of each methodology with respect to metabolomics data analysis are also discussed.

R E V I E W   A R T I C L E

Introduction

Error analysis is the detection, identification, and quantification of different types of uncertainty present in measurements and the propagation of this uncertainty through mathematical calculations and procedures. This definition associates the term *error* more with precision and less with mistake or accuracy. As such, error analysis plays a fundamental role in describing the degree of confidence in results derived from experiments across a variety of disciplines. Naturally, the importance of error analysis has grown with the increase in the number and heterogeneity of measurements obtained from newer, high-throughput, omics-level technologies.

The increase in heterogeneity is particularly problematic for metabolomics, which has more heterogeneity than other omics technologies due to the much wider range of molecular entities detected and measured: thousands of distinct metabolites versus single classes of repetitive linear polymers like DNA, RNA, and proteins. Also within the context of metabolomics, error analysis has several fundamental uses: i) the improvement of experimental design, ii) the quality control of experiments, iii) the selection of appropriate statistical methods, and iv) the determination of uncertainty in results. This review will introduce the fundamental concepts of error analysis as they apply to a wide range of metabolomics experimental designs and will discuss current methodologies for determining the propagation of uncertainty in appropriate metabolomics data analysis, with increasing level of statistical detail as the review progresses.

[a]Department of Chemistry, Center for Regulatory and Environmental Analytical Metabolomics, University of Louisville, Louisville, Kentucky, USA

* Corresponding author.
E-mail address: hunter.moseley@louisville.edu

Basic statistical terminology

Owing to the confusion and misconceptions about statistical definitions within published metabolomics literature, I begin by concisely defining the main statistical terminology and concepts used throughout the rest of this review, but this paper is no substitute for more in-depth reading and study [1-3]. As shown in Table 1, the estimate of an expected value is usually represented as the mean or average of repeated measured values ($\bar{x}$) of a given measured variable x. The median, defined as the middle value in a sorted list of repeated measured values, is another common estimate of an expected value, often preferred when the distribution of the measured variable is significantly skewed (non-symmetric). The variance σ^2_x represents the spread of these repeated measured values around this mean. The standard deviation σ_x is just the square root of the variance. Another useful term for describing uncertainty is the standard error $\sigma_{\bar{x}}$ (also known as the standard deviation of the mean), which is a probabilistic description of how close the mean is to the expected value.

A related term is a confidence interval, which identifies a range that includes the expected value at some level of confidence (typically 95% or 99%). A multidimensional generalization of a confidence interval is a confidence region, typically approximated with an elliptical shape. Both confidence intervals and confidence regions are especially useful descriptions of uncertainty when the probability distribution of the measurement(s) is unknown and clearly non-normal. However, the term confidence interval should not be confused with a tolerance interval which is more analogous to standard deviation and describes a range that includes a certain proportion of the population. Next, covariance σ^2_{xy} describes how two measured variables vary together. And correlation r_{xy} describes the dependence between two measured variables. Both of these terms are useful in describing the relationship between measured variables.

Finally, the term statistical power is the probability that a statistical test will properly reject the null hypothesis and not make a false negative decision (Type II error). A null hypothesis is a falsifiable assertion. A statistical test is a method for making

decisions from data by deciding whether to reject a null hypothesis at a certain level of significance. Statistical methods can be divided into parametric methods that assume an underlying probability distribution for the data and non-parametric methods that make no such assumptions and treats data from a categorical or ordinal (having an order or ranking) perspective. A p-value is the probability that a true null hypothesis would have an observed value at a certain extreme or worse. And a Type I error is the improper rejection of a null hypothesis, also known as a false positive.

Table 1. Common statistical terms and their mathematical definitions.

Term	Equation
Mean (estimate of the expected value)	$\bar{x} = \dfrac{\sum_i^N x_i}{N}$
Variance	$\sigma_x^2 = \dfrac{\sum_i^N (x_i - \bar{x})^2}{N - 1}$
Standard Error	SE_x or $\sigma_{\bar{x}} = \dfrac{\sigma_x}{\sqrt{N}}$
Covariance	$\sigma_{xy}^2 = \dfrac{\sum_i^N (x_i - \bar{x})(y_i - \bar{y})}{N - 1}$
[a](Pearson's) Correlation	$r_{xy} = \dfrac{\sum_i^N (x_i - \bar{x})(y_i - \bar{y})}{(N - 1)\sigma_x \sigma_y}$

[a]Pearson's correlation coefficient [4]

Types of variance and error

The major divisions of variance in bioanalytical experiments are biological versus analytical variance, which categorize the source of the variance (Figure 1). Biological variance arises from the spread of measured values observed from multiple biological samples due to differences in individuals. Analytical variance arises from the spread of measured values observed from multiple measurements made from the same biological sample, including all technical steps from sample acquisition to primary analytical data collection. More often, the biological variance is significantly larger than the analytical variance. The major divisions of error are systematic versus nonsystematic (random) error, which describe the type of error. Systematic error is a type of uncertainty not revealed by repeated measurements and represents biases in measurements that must be tested for separately in order to address or correct. While systematic error does not appreciably affect variance, it can affect covariances and correlations between measured variables. Nonsystematic error, also known as error variance, is the experimental uncertainty revealed by repeated measurements and can be reliably estimated by statistical methods.

A third kind of variance is systematic variance, which represents the variance between groups of related samples in the sample set. Depending on the specific experiment and the applied statistical analysis, specific systematic variances can be part of the detectable signal or part of the uncertainty in the measurements due to confounding factors. In other words, one scientist's uncertainty is another scientist's usable systematic variance. For example, in matched case-control measurements, the intergroup variance is the actual desired signal to detect. In other experiments, differences in group composition between case-control measurements like subject sex (male/female) can be an additional source of confounding systematic variance. In the special case when the sample set is unintentionally uniform (i.e. homogeneous), the group effect on the measurements will be systematic error and will not contribute to systematic variance. The classification of systematic error is further complicated by interpretive errors that often have the appearance of systematic error. The classic example is the error introduced by a "standard" measurement used in the correction of other measurements. Also, these divisions in variance and error are not mutually exclusive (Figure 1). Both biological and analytical variance can be divided into systematic variance and nonsystematic error components.

Typical biases in metabolomics

Most forms of error and confounding variance in measured variables originate from some type of bias. This term *bias* refers to any factor that distorts the design, execution, analysis, and interpretation of a measurement [5] and is an ever-present problem for metabolomics experiments [6,7]. A bias typically causes: i) a systematic error that distorts the measured values but does not change the variance; ii) a systematic variance arising from a confounding factor that is either unknown or inadequately addressed; or iii) an interpretive error due to an inadequate or improper statistical method of analysis. One can also categorize these biases as biological, analytical, and interpretive according to whether their effects manifest in the biological experiment, the analytical measurements, or subsequent data analysis (Figure 1).

Of the biological biases, selection bias has historically been one of the most problematic for case-control experiments [5], which are very common in metabolomics observational studies. There are very many types of selection bias [5], but usually they represent an unbalanced selection of subjects that differ genetically or epigenetically. Also, temporal selection biases (when samples are taken) are especially hard to control for, given the natural cycles that exist in organisms. Biological conditions bias is another significant type of biological bias where some experimental condition endured by the subjects is not properly controlled for or considered.

Of the analytical biases, sample preparation bias usually presents the largest set of problems for the experimentalist. Any deviation in how and how long a sample is extracted, quenched, and stored can greatly impact the measured analytes [8,9]. Standards bias represents the effects that different standards have on deriving some absolute or even relative quantification from a measurement. Sample complexity bias represents the effects on measurement due to physical interaction between a mixture of analytes present in a sample. Analytical conditions bias is due to some change in analytical conditions either during or between measurements.

Various interpretive biases and errors can also cause serious problems. However, many of the biases are simply born out of ignorance or laziness. Researchers tend to use the methods that they know and are easy to use, whether or not such methods are appropriate for the analysis. As scientists, it is a tendency that we should resist. Of the methodological biases, statistical assumptions about metabolomics data are most rampant. The two worst assumptions are that: i) nonsystematic error is completely Gaussian distributed and ii) measured variables are uncorrelated. The first assumption should never be made with analytical techniques like mass spectrometry, which by their very nature should include Poisson distributed error [10]. For NMR data, the nonsystematic error is often complicated and may include Lorentzian distributed observables, for which the variance is not analytically defined, but can be numerically estimated [11]. The second assumption of
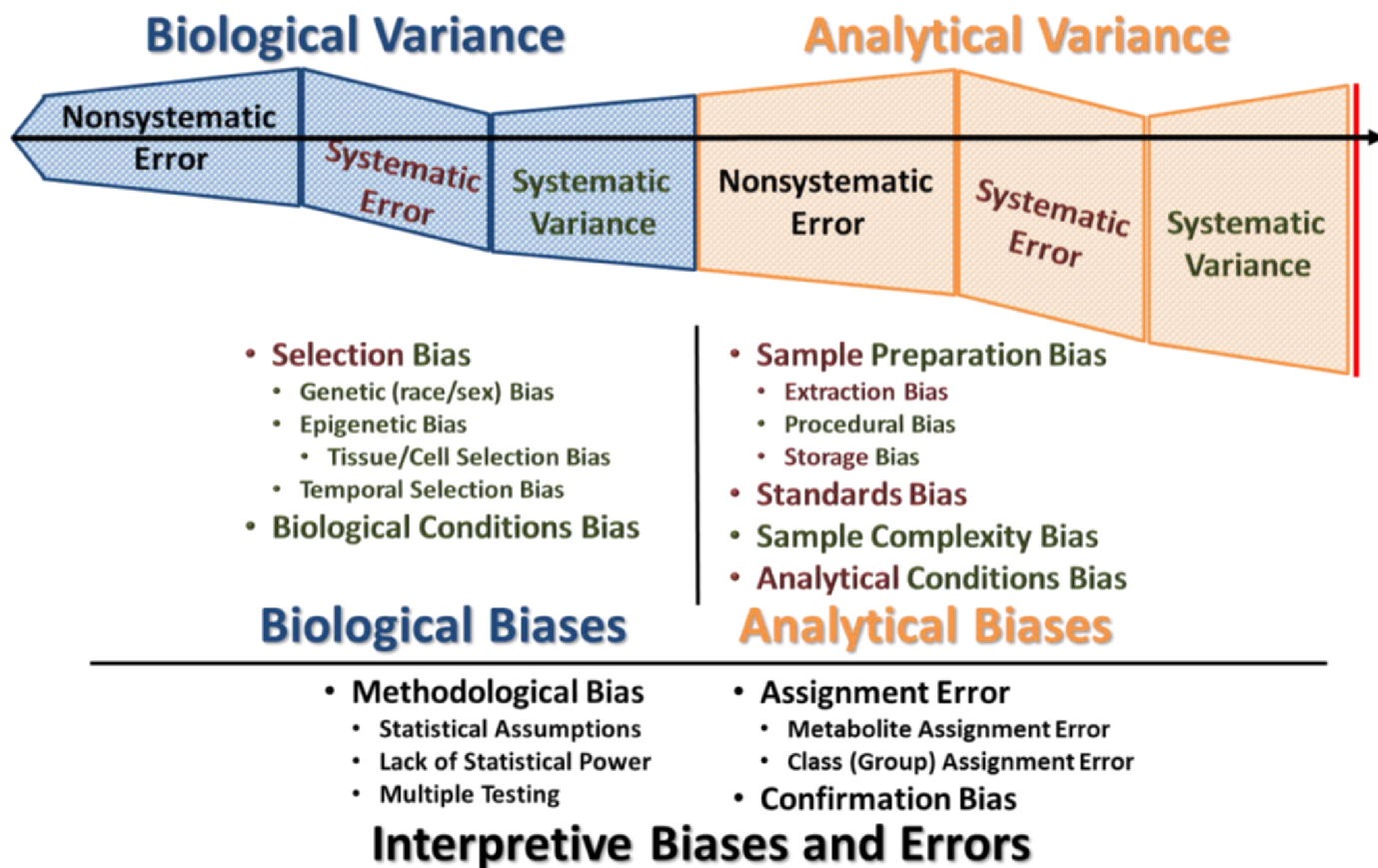
2

Figure 1. The major divisions of variance, error, and bias in bioanalytical experiments. The expanding cone represents the effects of major types of variance and error on the spread (red line) and distribution of measured values. The major divisions of variance by source are biological versus analytical variance. The major divisions of error are systematic versus nonsystematic error, with a third related entity, systematic variance, representing the variance between groups of related samples in the sample set. The major divisions of bias are biological, analytical, and interpretive, based upon when their effects manifest in the bioanalytical experiment and data analysis.

independence is unfounded given the correlation of metabolites in cellular metabolic networks. Another common interpretative bias is a lack of statistical power due to poor design and insufficient subject numbers. This situation is often confounded by a lack of correction for multiple testing. Also, assignment errors like the misassignment of a subject to a specific group (i.e. class assignment error or class noise) or the misidentification of an analyte can lead to serious misinterpretations. Another common interpretive bias is the use of preconceptions in interpretation, which can lead to a confirmation bias. For example, use of unverified metabolic models (i.e. a model of part or all of the metabolism of one or more specific organisms) is a common confirmation bias in metabolic data analysis. Also, the careless use of priors and limitation to expected metabolic models can lead to confirmation bias within a Bayesian statistical formalism [12].

Dealing with biases

Constant vigilance is required to mitigate the many biases (Figure 1) that may add both significant systematic error and confounding systematic variance to a metabolomics dataset [6,7,13]. While there is no simple checklist of biases to look for and procedures to follow, there are several straight-forward strategies for dealing with many of these biases. The first is to use reasonably consistent experimental designs that exclude: i) partial consistencies for specific groups of samples, which may lead to a systematic variance from biological or analytical biases, and ii) trivial consistencies that may limit the

generalization of results, due to systematic errors from biological or analytical biases. Building on the first strategy, the second strategy is to use effective experimental designs like matched-pair case-control experiments to limit the effects of confounding factors, especially from large biological biases. Also, the samples should be balanced between case-control groups for possible confounding factors (sex, age, related biological condition), in order to prevent systematic variance. The optimal design is equally balanced sampling (i.e. blocking) between the most likely confounding factors within case-control groups. Blocking allows the use of more sophisticated statistical methods [13]: i.e. ANOVA instead of a t-test or Welch ANOVA [14] instead of a Welch's t test [15].

The third strategy is to directly test how well a set of measured values for a given measured variable fits an expected/assumed analytical nonsystematic error distribution. Various implemented algorithms exist for testing how well a set of measured values fits different common distributions. For example, the nortest R package [16] includes several normality (normal distribution) tests including the Shapiro-Wilk test [17] and the Anderson-Darling test [18].

The fourth strategy is to validate results with temporally-separated datasets (i.e. analytical cross-validation) as a method for detecting the presence of biases. However, failure to detect bias by this method does not guarantee a bias-free approach.

The fifth strategy is to use blinded metabolomics experiments to reduce bias [6]. The double-blind randomized control trial is

3

considered the gold standard for reducing researcher-introduced performance bias that can affect both biological and analytical conditions [19]. However, even this very effective method has known masking biases due to the psychological effects of the trial itself, especially when human subjects are involved [20]. Also, many observational studies cannot be double-blinded. Despite this, the blinding of analytical and/or statistical researchers (i.e. analytical and/or interpretive single-blinding) can reduce performance biases that affect both the analytical conditions and statistical interpretation [21].

The sixth strategy is to use analytical controls to correct for or to prevent analytical biases. However, application of this strategy is often specific to the analytical technique and instrumentation. With each sample in its own tube or vial, periodic controls or time-stamped near-random controls (random except to prevent neighbor effects) can be used to track analytical conditions during measurement [9]. For samples handled on plates, Latin square or 2D near-random patterns can be used [9]. An easier alternative representing a combination of strategies five and six involves the use of blind controls to detect and correct for several analytical biases including performance bias [22]. For full thoroughness, often a series of controls composed of complex mixtures of representative or chemically similar metabolites must be used to determine systematic error arising from sample extraction methods [23] and even from interactions between the metabolites themselves with respect to the specific analytical techniques employed. The latter is particularly problematic for electrospray mass spectrometry.

The seventh and final strategy is to fully document: i) the biological and analytical experimental procedures that produced a given metabolomics dataset; ii) the statistical procedures used in the analysis of the dataset; iii) a detailed list of all known or potential biases and assumptions, along with results of any analysis and testing of these bias and assumptions; and iv) results with adequate measures of uncertainty and confidence or at least a good explanation for why uncertainty and confidence measures are not provided. The two most common explanations are that not enough replicates were collected for error analysis or methods to determine the propagation of uncertainty do not exist for a specific statistical method or procedure. Such documentation enables thorough evaluation and peer-review by others and it facilitates future meta-analyses, especially when the datasets and documentation are deposited into public repositories like MetaboLights [24]. Minimum reporting standards for metabolomics experiments already exist for biological (plant specific) experimental procedures [25], analytical experimental procedures [26], and statistical procedures for analysis [27]. However, it would be better to augment these standards, especially for reporting known and potential sources of bias by borrowing from well-documented clinical standards [6] like STARD [28] and CONSORT [29,30].

The overall purpose of these seven strategies is to identify which possible biases (Figure 1) may affect the proper interpretation of a metabolic experiment and then to limit their effect. However, in the end, a judgement call must be made on whether these strategies were adequate for the proper interpretation of the obtained metabolomics dataset(s).

Methodology

Standard error analysis methods

One central question that error analysis must address is whether nonsystematic error, variances, or covariances from either an analytical or biological source will prevent the desired detection and interpretation of biological systematic variance in a given metabolomics dataset. From this perspective, any nonsystematic error, variance, and covariance that could interfere with biological interpretation represent uncertainty that should be determined/estimated and properly addressed. To be most effective, error analysis needs to be part of the experimental design before the experiment even begins. Decisions must be made on the number of replicates at each stage of the experimental protocol in order to have the necessary dataset for thorough error analysis [31]. In some instances, it will be practically impossible to obtain the necessary replicates for a thorough error analysis and other approaches for estimating sources of variance and error will be necessary. Also, issues of statistical power must be considered for the statistical methods employed [34]. However, even when issues of statistical power are satisfied, typically when statistical power $\geq$ 0.8 at α=0.05 with "reasonable" statistical assumptions, it is advisable to test these "reasonable" assumptions by increasing analytical replicates for a subset of the samples. If these assumptions do not hold true or there is a lack of statistical power due to large analytical variance, then additional steps may be taken to address these issues, including: i) increasing the number of analytical replicates to deal with analytical nonsystematic error; ii) correcting for factors that cause analytical systematic variance; iii) switching to appropriate statistical methods that do not rely on these failed assumption(s); and/or iv) incorporating estimates of analytical variance and covariance into more sophisticated statistical methods.

So standard error analysis can be broken down into two major steps: i) error estimation and probability distribution testing and ii) error (uncertainty) propagation analysis. The first step involves the distribution testing, calculation/estimation, modeling, and comparison of nonsystematic error, variance, and covariance arising from biological and analytical sources. Testing for a normal distribution (normality test) is by far the most important probability distribution test. For normality testing, 8 to 10 replicates are considered the minimum needed with the Shapiro-Wilks test [35]. However, 20 to 30 replicates are typically desired for significant power, even though the Shapiro-Wilks test is considered the most statistically powerful normality test [17,36]. When a measured variable fails normality tests, then other non-parametric or more fault-tolerant statistical methods should be employed to compensate for the lack of normality. For example, the non-parametric Wilcoxon-Mann-Whitney test is preferred to a t-test when the data are significantly non-normal [37-39]; but neither test works well if the data are highly skewed [40]. Sometimes, it is advantageous to interpret continuous measured variables categorically as discreet ranges and use binomial or multinomial statistical tests.

While the use of 3 analytical replicates is the minimum needed for quality control that includes the ability to detect outliers [9], 13 replicates (12 + 1) are considered the minimum for calculating variances with ~half-width confidence intervals at least at the 90% confidence level when approximately normally distributed; 30 replicates are required to calculate variances with ~half-width confidence intervals at the 99% confidence level [41]. In addition, these analytical replicates will naturally lower the standard error of the measured variable; however, there is a practical limit due to the availability of analytical instrumentation or of usable sample. Also, some analytical techniques have rather complex nonsystematic error structures that may pose additional difficulties [10,42]. In these instances, a bootstrap procedure can be used to calculate confidence intervals [43,44]. If the appropriate replicates across an experimental protocol are obtainable, then advanced mixed effect modeling methods can be employed to tease apart different sources of variance,

4

$$y = f(x_1, \ldots, x_n) \approx \underbrace{f(\bar{x}_1, \ldots, \bar{x}_n)}_{0th\ order} + \underbrace{\sum \frac{\partial f(\bar{x}_1, \ldots, \bar{x}_n)}{\partial x_i}(x_i - \bar{x}_i)}_{1st\ order} \left[+ \underbrace{\frac{1}{2!}\sum\sum \frac{\partial^2 f(\bar{x}_1, \ldots, \bar{x}_n)}{\partial x_i \partial x_j}(x_i - \bar{x}_i)(x_j - \bar{x}_j)}_{2nd\ order} + \cdots \right.$$

Equation 1

$$\bar{y} = \frac{1}{N}\sum_a^N f(x_{1,a}, \ldots, x_{n,a}) \approx \frac{1}{N}\sum_a^N \left(f(\bar{x}_1, \ldots, \bar{x}_n) + \sum \frac{\partial f(\bar{x}_1, \ldots, \bar{x}_n)}{\partial x_i}(x_{i,a} - \bar{x}_i) \right)$$

$$\approx f(\bar{x}_1, \ldots, \bar{x}_n) + \sum \frac{\partial f(\bar{x}_1, \ldots, \bar{x}_n)}{\partial x_i}\frac{1}{N}\sum_a^N (x_{i,a} - \bar{x}_i) \approx f(\bar{x}_1, \ldots, \bar{x}_n) + \sum \frac{\partial f(\bar{x}_1, \ldots, \bar{x}_n)}{\partial x_i}(0)$$

$$\approx f(\bar{x}_1, \ldots, \bar{x}_n)$$

Equation 2

$$\sigma_y^2 = \frac{1}{N-1}\sum_a^N (y_a - \bar{y})^2$$

$$= \frac{1}{N-1}\sum_a^N \left(f(x_{1,a}, \ldots, x_{n,a}) - f(\bar{x}_1, \ldots, \bar{x}_n) \right)^2$$

$$\approx \frac{1}{N-1}\sum_a^N \left(f(\bar{x}_1, \ldots, \bar{x}_n) + \sum \frac{\partial f(\bar{x}_1, \ldots, \bar{x}_n)}{\partial x_i}(x_{i,a} - \bar{x}_i) - f(\bar{x}_1, \ldots, \bar{x}_n) \right)^2$$

$$\approx \frac{1}{N-1}\sum_a^N \left(\sum \frac{\partial f(\bar{x}_1, \ldots, \bar{x}_n)}{\partial x_i}(x_{i,a} - \bar{x}_i) \right)^2$$

$$\approx \frac{1}{N-1}\sum_a^N \left(\sum \left(\frac{\partial f(\bar{x}_1, \ldots, \bar{x}_n)}{\partial x_i}(x_{i,a} - \bar{x}_i) \right)^2 + \sum\sum_{j\neq i} \frac{\partial f(\bar{x}_1, \ldots, \bar{x}_n)}{\partial x_i}\frac{\partial f(\bar{x}_1, \ldots, \bar{x}_n)}{\partial x_j}(x_{i,a} - \bar{x}_i)(x_{j,a} - \bar{x}_j) \right)$$

$$\approx \underbrace{\sum \left(\frac{\partial f(\bar{x}_1, \ldots, \bar{x}_n)}{\partial x_i} \right)^2 \sigma_{x_i}^2}_{variance\ sum} + \underbrace{\sum\sum_{j\neq i} \frac{\partial f(\bar{x}_1, \ldots, \bar{x}_n)}{\partial x_i}\frac{\partial f(\bar{x}_1, \ldots, \bar{x}_n)}{\partial x_j}r_{x_i x_j}\sigma_{x_i}^2 \sigma_{x_j}^2}_{covariance\ sum}$$

Equation 3

$$\sigma_y^2 \approx \mathbf{j}(\bar{\mathbf{x}})^{\mathrm{T}}\mathbf{C_x}\,\mathbf{j}(\bar{\mathbf{x}})$$

Equation 4

5

$$\mathbf{C_y} \approx \mathbf{J}_F(\bar{\mathbf{x}})\mathbf{C_x}\,\mathbf{J}_F(\bar{\mathbf{x}})^{\mathrm{T}}$$

Equation 5

including from random sources [32,33,45]. However, inadequate number of replicates requires the use of other approaches for estimating variance and its sources: i.e., estimation of variance indirectly from similar experiments, assuming biological variance is any variance not directly attributable to an analytical source, and/or modeling variance from time series measurements [46].

Calculating or estimating covariance is typically problematic for metabolomics datasets due to the small number of replicates n versus the number of measured variables p ($n<<p$). Also, reasonable guidelines for calculating/estimating covariances are not straight-forward and the number of analytical replicates needed often depends on the number of measured variables and the amount of correlation to accurately detect and estimate. One rule of thumb based on single correlation estimation is $n \approx 16/(\ln((1+r)/(1-r)))^2$ for detecting a correlation r at an α (Type I error) of 0.05 and power of 80%, but this rule ignores the fact that $p(p-1)/2$ correlations are being estimated [41]. Despite these problems, there are ways to deal with the $n<<p$ condition in metabolomics datasets: i) by using high variance of covariance to weight towards an expected covariance matrix structure [47]; ii) by averaging analytical covariance across biological replicates [31]; and iii) by using known variance-covariance relationships to estimate an analytical covariance matrix from calculated analytical variances [48].

Error (uncertainty) propagation analysis

The propagation of uncertainty (error) through functions and algorithms is analyzed by two fundamentally different types of methods: i) mathematical (analytical) derivation and approximation and ii) numerical analysis. Several software platforms exist to facilitate the use of these methods [49]; however, custom implementation of these methods is often necessary to handle specific issues of a given metabolomics dataset and its analysis. With few exceptions, almost all analyses of error propagation via mathematical derivation and approximation are performed from a linear perspective [50]. This linear assumption is used, whether the functions and algorithms being analyzed are linear or nonlinear [1]. To fully understand the effects of this linear perspective, we start with Equation 1, which describes the multivariate Taylor series approximation for a given function f. Most error propagation analyses use only the 0th and 1st order terms, due to the exponential growth of higher order terms, as highlighted by the sum of sums of 2nd order terms versus the simple sum of 1st order terms in Equation 1. However for linear equations, this approach simplifies to an exact solution, since all 2nd order and higher terms are zero.

In Equation 2, the approximation of $\bar{y}$ (mean of f), using only the 0th and 1st order terms of the series, simplifies to f applied to $\bar{\mathbf{x}}$ (mean of vector $\mathbf{x}$). This sets up for the approximation of σ^2_y (variance of f), again using only the 0th and 1st order terms of the series (Equation 3). The rearrangement of terms simplifies to sums of variance and covariance terms that approximate the effects of the variances of $\mathbf{x}$ on f using only the tangent of f at $\bar{\mathbf{x}}$ ($\partial f/\partial x_i(\bar{x}_1,\ldots,\bar{x}_n)$). Thus, the sums of variance and covariance terms are just linear approximations for the propagation of uncertainty through f. These sums of variance and covariance terms are often represented with a simpler vector/matrix notation (Equation 4), where $\mathbf{C_x}$ is the covariance matrix for $\mathbf{x}$ (sometimes called a variance-covariance matrix and represented as Σ_x). And $\mathbf{j}(\bar{\mathbf{x}})$ is the vector of partial derivatives at $\bar{\mathbf{x}}$, i.e. $[\partial f/\partial x_1(\bar{\mathbf{x}}), \ldots, \partial f/\partial x_n(\bar{\mathbf{x}})]$ [51,52]. This concept can be further generalized for a system of equations $\mathbf{y} = F(\mathbf{x})$ as shown in Equation 5, allowing the calculation of $\mathbf{C_y}$ using the Jacobian matrix of all first-

order partial derivatives of F at $\bar{\mathbf{x}}$, i.e. $\mathbf{J}_F(\bar{\mathbf{x}})$. But when no correlation exists between x_i variables, Equation 3 simplifies to just a sum of variance terms (or standard error terms) [2,53], which is often referred to as Gaussian error propagation (GEP). It is this "no correlation" GEP version that underlies almost all standard error propagation rules [1]. However, this GEP approximation can be very poor when significant correlation exists and/or for nonlinear functions when $\bar{\mathbf{x}}$ is near a critical point or an inflection point [50].

Numerical error propagation analysis

Owing to these approximation problems with nonlinear functions and complex algorithms or significant deviations from normality, a second approach to error propagation using numerical analysis is often more accurate and much easier to implement in these situations, which are typical for metabolomics data analysis. The most common numerical approach is known as the Monte Carlo method, which samples a given function or algorithm applied to random input values [54]. However, from this broad definition, the Monte Carlo method is really a large collection of methods with a wide variety of applications beyond the scope of this review [55]. Therefore, we will focus on a simple Monte Carlo method (Equation 6) where a set of pseudo-random input vectors of values ($\mathbf{x}_i \in X$) with specific probability distributions ($X_j \sim D_j$) are used to generate a set of vectors of values ($\mathbf{y}_i \in Y$) from a given function or algorithm (f) for analyzing the propagation of uncertainty through f. However, there are more advanced Monte Carlo approaches that can analyze the propagation of uncertainty from both repeated measurements and other information within a Bayesian framework [56]. As previously mentioned, it is important to characterize the probability distributions for the vector of input variables arising from analytical procedures via testing of common probability distributions and the calculation or estimation of expected values, variances, and correlations from experimental data.

$$\mathbf{y}_i = f(\mathbf{x}_i) \text{ where } \mathbf{x}_i \in X \text{ and } X_j \sim D_j$$

Equation 6

With these statistical characteristics, there are several ways to generate a pseudo-random sampling for the input variables for common probability distributions. Both Matlab [57] and R (a free, open source, statistical programming language that is robustly supported) [58] have built-in functions for generating pseudo-random values for most of the common probability distributions There are also straight-forward algorithms to calculate a set of pseudo-random values that fit several of the common probability distributions using a set of uniformly distributed pseudo-random values (i.e. U[0,1]) [59,60]. In particular, the Box Muller method is a popular (easy to implement) algorithm for calculating normally distributed pseudo-random pairs of values from pairs which are U[0,1] distributed [60]. Also by definition, the inverse of a cumulative distribution function can be used to calculated pseudo-random values from U[0,1] distributed values [61]. But care should be taken in selecting a good uniformly distributed random number generator for use in generating a given distribution. An enhanced Wichmann-Hill algorithm is recommended by the current draft of the Guide to the Expression of Uncertainty in Measurements [62], but other algorithms have passed rigorous testing as well [63].

Even pseudo-random values of complex or unknown distributions can be estimated using a two-sample Kolmogorov–Smirnov test (K-S test), which is a non-parametric test that compares the distributions of

6

two samples and determines whether they deviate significantly or not [64,65]. In this simple approach, sets of pseudo-random values are generated based on bootstrap-derived statistical parameters and tested against an experimentally derived set of measured values using the two-sample K-S test. But there are also various approaches to simulate significantly non-normal multivariate random variables that even include correlation [66,67] or approaches that take random variables with given distributions and introduce correlation [68]. Finally, with the appropriate set of pseudo-random input vectors of values ($\mathbf{x}_i \in X$), application of the given function or algorithm (Equation 6) produces a set of vectors of values ($\mathbf{y}_i \in Y$) that can be directly analyzed in an analogous manner as experimental data, i.e. probability distribution testing and calculation of expected values, variances, standard errors, and correlations when common probability distributions are present.

If the function f is linear and X variables are (reasonably) independent and identically distributed ($X_i \sim X_j$) with finite variance, then Y often reflects the distribution of X ($Y_i \sim X_j$) or even approximates a normal distribution for Y variables (Y_i) that depend on many X variables (X_j). However, if f is nonlinear, then drastically non-normal distributions are common for Y and quite distinct from X, even if X is normally distributed. And nonlinearity is very common for metabolic models with exchange and bidirectional fluxes, which sometimes can be solved by a linearization of the model [69]. However, in non-linearizable situations or simply when the Y variables are very non-normal (a multimodal distribution for example), a median with a confidence interval is preferable to a mean with a standard error, and can be estimated by a straight-forward approach with samples sizes of 1000 or 10000, depending on the desired level of confidence in these confidence intervals [70]. The method simply orders the sampling for each Y_i and takes the interval ($y_{(n+1)(1-c)}$, $y_{(n+1)c}$) where c is the level of confidence as a fraction (for example 0.95). Likewise, a Spearman's rank correlation coefficient can be used to calculate correlation in a non-parametric way in these situations [71].

Error propagation in inverse problems

Another common numerical approach used to analyze metabolomics data is the optimization of parameters in an inverse problem [72]. Often a model (g in Equation 7) with parameters ($\mathbf{y}_i$) for calculating values ($\mathbf{x}_i$) that are directly comparable to experimental data ($\mathbf{x}_{exp}$) is much easier to construct than an analytical function (f in Equation 6) that can calculate desired parameters from experimental data, especially when experimental data is collected in a time series [73]. For example, a model of the relevant chemical reactions for a "known" cellular metabolic network (to the limited degree of our current scientific knowledge) is more easily constructed and used to calculate specific metabolite fluxes and pools (mass-related characteristics) that can be compared to experimental values [74-76]. These models are becoming more available and easier to use and modify thanks to databases like Biomodels.net and associated modeling tools [77,78]. However, care must be taken in the interpretation of the term "model": sometimes model refers to just the framework of equations (g), sometimes it refers to g and fixed input parameters ($y_{i,j} = c_j$), and sometimes it refers to g and optimized parameters ($\mathbf{y}_{opt}$). For the rest of this discourse, model will simply refer to g.

$$\mathbf{x}_i = g(\mathbf{y}_i) \text{ where } g \approx f^{-1}$$

Equation 7

Equation 8 describes an objective function (O_s - simple), also known as a target function or energy function depending on context, which compares the results from this model $\mathbf{x}_i$ with experimental data $\mathbf{x}_{exp}$ through some norm function, like an ℓ_2-norm in this instance. This objective function is minimized while model parameters are optimized to $\mathbf{y}_{opt}$ using an optimization method of choice, which is often some type of Monte Carlo method by definition (cf. simulated annealing).

$$O_s(\mathbf{y}_i) = \left\| g(\mathbf{y}_i) - \mathbf{x}_{exp} \right\|^2 = \left\| \mathbf{x}_i - \mathbf{x}_{exp} \right\|^2$$
$$= \sum_j \left(x_{i,j} - x_{exp,j} \right)^2$$

Equation 8

However, almost all metabolomics inverse problems are ill-posed due to model complexity and non-linearity, limitations in the number and variety of measurements, and significant amounts of uncertainty in the measurements. These characteristics give rise to common properties of ill-posed problems: i) preclusion of a unique solution $\mathbf{y}_{opt}$ to a given set of experimental measurements $\mathbf{x}_{exp}$; ii) the existence of multiple solutions $\mathbf{y}_{opt,l}$; iii) the presence of discontinuities in the objective function; and iv) high conditioning (i.e. large variation) in model parameters with respect to small changes in experimental measurements. Regularization of the ill-posed objective function as a better-posed objective function is required to mitigate these properties of ill-posedness and to minimize the amplification of analytical uncertainty in the derived model parameters $\mathbf{y}_{opt}$ during optimization. Without such regularization, the optimization of the objective function in Equation 8 quickly leads to overfitting of model parameters to the error in the experimental data. However, an important question with regularization is how much information is present in the available data, and to what extent prior knowledge can be used to overcome limitations in the data without introducing undue bias?

Tikhonov regularization is a common regularization method used to prevent overfitting in inverse problems, due to its effectiveness and simplicity of implementation [79]. Tikhonov regularization uses *a priori* knowledge of expected model parameters ($\mathbf{y}_E$) to allow stable optimization with the experimental data ($\mathbf{x}_{exp}$) while minimizing fitting to its error [80]. Equation 9 incorporates a Tikhonov regularization $R(\mathbf{y}_i, \mathbf{y}_E)$ as a *p*-norm ($||\mathbf{w}||_p$) into Equation 8 with a certain weighting α, which must be kept small enough to prevent any significant bias to $\mathbf{y}_E$ but large enough to prevent overfitting [79], and a *p* that properly balances between the norms $||\mathbf{x}_i||$ and $||\mathbf{y}_i||$ [81]. Once the issues of picking a proper α and p are overcome, even a confidence region around $\mathbf{y}_E$ that includes $\mathbf{y}_{exact}$ can be estimated with respect to $\|\mathbf{y}_i - \mathbf{y}_E\|_p^2$ based on a Fisher distribution, if normality conditions hold and the dimensionality of $\mathbf{x}_{exp}$ is significantly greater than the dimensionality of $\mathbf{y}_E$ [72,82].

$$O_T(\mathbf{y}_i) = \left\| g(\mathbf{y}_i) - \mathbf{x}_{exp} \right\|^2 + \alpha R(\mathbf{y}_i, \mathbf{y}_E)$$
$$= \left\| \mathbf{x}_i - \mathbf{x}_{exp} \right\|^2 + \alpha \left\| \mathbf{y}_i - \mathbf{y}_E \right\|_p^2$$
$$= \sum_j \left(x_{i,j} - x_{exp,j} \right)^2 + \alpha \left(\sum_k \left| y_{i,k} - y_{E,k} \right|^p \right)^{\frac{2}{p}}$$

Equation 9

But when the analytical error for $\mathbf{x}_{exp}$ can be determined, a more general approach for regularization can be employed, where the optimization is stopped when the objective function is below the observed analytical error [72]. One error-bounded generalized least squares implementation, shown in Equation 10, stops optimization when the objective function (O_g) is below an error threshold $\delta_{\mathbf{x}}$. This threshold can be approximated by a χ^2 statistic $\chi^2_{n-m}(1-\beta)$, with n-m degrees of freedom and a p-value of 1-β, where n is the number of measured experimental variables, m is the number of parameters in the model, and β is the desired level of confidence. However, this implementation assumes that $O_g(\mathbf{y}_{opt}) < \chi^2_{n-m}(1-\beta)$, where $\mathbf{y}_{opt}$ is determined by the lowest $O_g(\mathbf{y}_i)$. With this assumption holding true, it is straight-forward to estimate the confidence region $CR_{1-\beta}(\mathbf{y}_{opt})$ and the individual confidence interval $CI_{yj,1-\beta}(\mathbf{y}_{opt})$ with a large set of optimizations in Equations 11 and 12 respectively [83].

$$O_g(\mathbf{y}_i) = \left(g(\mathbf{y}_i) - \mathbf{x}_{exp}\right)^{\mathrm{T}} \mathbf{C_x}^{-1}\left(g(\mathbf{y}_i) - \mathbf{x}_{exp}\right) \leq \delta_{\mathbf{x}}$$

Equation 10

$$CR_{1-\beta}\left(\mathbf{y}_{opt}\right)$$
$$\approx \left\{\mathbf{y}_i \middle| O_g(\mathbf{y}_i) \leq \delta_{\mathbf{x}} \approx O_g(\mathbf{y}_{opt}) + \chi^2_{n-m}(1-\beta)\right\}$$

Equation 11

$$CI_{y_j,1-\beta}\left(\mathbf{y}_{opt}\right)$$
$$\approx \left\{y_{j0} \middle| \min O_g(\mathbf{y}_i)\middle|_{y_j=y_{j0}} \leq \delta_{\mathbf{x}} \approx O_g(\mathbf{y}_{opt}) + \chi^2_1(1-\beta)\right\}$$

Equation 12

However, this general approach in Equations 10, 11, and 12 only holds if all of the measured variables are normally distributed and the analytical covariance matrix $\mathbf{C_x}$ is known or well-estimated. Also for Equation 11, the residuals normalized by the square root of the covariance matrix (i.e. normalization matrix $\mathbf{C_x}^{-\frac{1}{2}}$) should be tested for normality. Plus, the optimization needs a large number of repetitions; however, the extra computational requirements can sometimes be mitigated by optimization methods that take advantage of first and second order partial derivatives, typically in the form of Jacobian and Hessian matrices [83]. If the measured experimental variables are not normally distributed, then it is extremely challenging to devise an objective function that will properly weight the residuals to a single error threshold and preserve the underlying distributions and covariance structure. There is another Monte Carlo method that estimates the confidence region and confidence interval via an analytical estimation of the covariance matrix of $\mathbf{y}$ ($\mathbf{C_y}$) from the inverse of the Jacobian matrix of g normalized by the square root of the analytical covariance matrix, $[\mathbf{C_x}^{-\frac{1}{2}}\mathbf{J}_g]^{-1}$ [69]. However, it suffers from both the need for normally distributed measured variables and the linear approximations made in the estimation.

Metabolic model verification and selection

Equations 8 through 12 are based on the grand assumption that model g is "reasonably" accurate, which has the potential of being a very large interpretive bias. Moreover, the faith in certain metabolic models is quite troubling, given the lack of verified details and errors in metabolic databases used in the construction of models of metabolic networks, especially models of eukaryotic metabolic networks [84]. Also, many metabolic models include more parameters than measured variables, which greatly limits the ability to verify such models. Given these caveats, there are two general approaches for improving model verification: a) pare down the metabolic model to what is relevant to the observables; and b) design experiments where there are enough observables to perform model selection (i.e. $n >> m$).

For the first approach, there are three main ways to pare down a metabolic model: i) gross model paring, ii) specific variable pairing by independence, and iii) specific variable paring by sensitivity. Gross model paring is simply limiting the model to relevant pathways and modules of a metabolic network [85], assuming that they are known. Specific variable pairing by independence limits the model parameters to the smallest set of independent or "free" model parameters from which other intermediate model parameters are derived [86]. An optimal set of free model parameters can be directly calculated via the determination of a basic set for the null space of the stoichiometry matrix [83]. Specific variable paring by sensitivity removes and/or simplifies parts of a model that include insensitive model parameters with respect to measured experimental variables. Determining which model parameters are insensitive to measured experimental variables can be done in a directed manner by carefully removing parts of a model and seeing if this model change does not appreciably worsens the objective function nor appreciably change the optimized values of other model parameters [87]. Also, the relative size of model parameter confidence intervals can help direct this paring. But more sophisticated analyses require calculating a sensitivity matrix (i.e. Jacobian matrix of f, $\mathbf{J}_f$), which is only straight-forward when the direct model f (from Equation 6) can be derived [88]. However, an estimation of the model parameter sensitivity matrix can be calculated, starting with an inverse of the Jacobian matrix of g normalized by the square root of the analytical covariance matrix, $[\mathbf{C_x}^{-\frac{1}{2}}\mathbf{J}_g]^{-1}$ in a few steps [69,83]. Both of these analytical approaches represent a linear interpretation of sensitivity with limitations similar to analytical error propagation as described above.

The second approach for improving metabolic model verification is to design experiments where there are significantly more measurables than model parameters. For metabolomics experiments, this has been greatly aided by the combined use of stable isotopes along with analytical techniques like NMR and ultra-high resolution mass spectrometry that can detect the specific incorporation of these labeling isotopes in metabolites [89-92]. The ability to detect specific isotopomers by NMR and isotopologues (mass-equivalent sets of isotopomers) by ultra-high resolution MS, greatly increases the number of possible measured experimental variables and provides a rigorous internal normalization that obviates the need for external controls, when absolute quantification is not necessary. In addition, there are several potential multiplicative factors on the number of measurable like time series measurements, the use of multiple stable isotopes (^{13}C, ^{15}N, ^{2}H), and the use of multiple isotope labeling source metabolites. In fact, the design of metabolomics is coming full circle, where metabolic models are being used to design optimal stable isotope labeling experiments [93-96]. However, this approach for metabolomics experimental design appears more robust for high quality metabolic models from model prokaryotic organisms.

Once appropriate metabolic models are constructed, model parameters pared, and enough experimental data collected (n >> m), models should be verified to limit significant interpretive confirmation bias. Often this verification process starts with gross measurement error detection via analysis of elemental and heat balances through the metabolic stoichiometry matrix, if the appropriate measurements are available [97,98]. If analytical error of

the experimental data can be adequately determined or estimated and the measured variables are approximately normally distributed, then the model can be further verified by a goodness-of-fit test based on a χ^2 statistic that reflects the number of degrees of freedom in the optimization and a desired level of confidence. When $O_g(\mathbf{y}_{opt}) > \chi^2_{n-m}(1 - \beta)$ (see Equation 10), the model should be rejected, especially if gross measurement errors can be ruled out [83]. However, rejection of a model is just a starting point for its improvement [99] and parts of the model involving parameters which are highly sensitive to small changes in measured variables are often a good place to look [83]. Eventually selection of models using standard methods like Akaike information criterion [100] should be a goal of the field, but these methods require approaches that limit the effects of overfitting during model parameter optimization, like the use of independent sets of measurements [101].

Discussion

Metabolomics has a range of applications including: the discovery or detection of biomarkers related to a cellular, physiological, or disease state of interest; the generation and verification of biochemical mechanism-based hypotheses for biological processes and phenomena; and the improvement of industrial fermentation processes via simulation. In all of these applications, the accurate determination of how uncertainty propagates through data analysis is required for the proper evaluation of results. Towards this end, experiments should be designed to minimize both biological and analytical error, to have the appropriate statistical power and controls, and to contain enough replicates to derive the analytical error and covariance, for at least a portion of the biological samples. Also, data and error analyses should include the testing of statistical assumptions which drive the use of appropriate statistical methods. One excellent example that embodies these principles involves the use of an analytical correlation matrix estimated from analytical replicates and used in a maximum likelihood principal component analysis instead of a standard principal component analysis [31,48].

In addition, the reporting of results in journals and data repositories must accurately include the derived and propagated uncertainty in the results. All p-values must be properly or at least conservatively adjusted under multiple testing conditions to prevent over-optimistic interpretation, especially in biomedical settings [7]. Also, p-values should be included with correlations (R^2) derived from regression analysis. Standard errors and confidence intervals should be reported for all measured and derived variables, when reasonably possible. The lack of reporting confidence region/intervals for published metabolic model parameters is due to a host of factors including: the requirement for approximately normally distributed measured variables, a reasonable metabolic model, requirement for well-estimated analytical covariances, requirement for more measurements than model parameters, error analysis expertise, and computational expense. But a few examples of reporting do exist ([102] for example), which will hopefully become the norm in the future. Moreover, a focus on error analysis in the development of and documentation of data analysis methods is critical to changing reporting habits.

Acknowledgements:

References

1. Taylor JR (1997) An introduction to error analysis: the study of uncertainties in physical measurements: Univ Science Books.
2. Bevington P, Robinson DK (2002) Data Reduction and Error Analysis for the Physical Sciences. McGraw Hill.
3. Hughes I, Hase T (2010) Measurements and their uncertainties: a practical guide to modern error analysis: Oxford University Press.
4. Pearson K (1920) Notes on the history of correlation. Biometrika 13: 25-45.
5. Sackett DL (1979) Bias in analytic research. Journal of chronic diseases 32: 51-63.
6. Ransohoff DF (2005) Bias as a threat to the validity of cancer molecular-marker research. Nature Reviews Cancer 5: 142-149.
7. Broadhurst DI, Kell DB (2006) Statistical strategies for avoiding false discoveries in metabolomics and related experiments. Metabolomics 2: 171-196.
8. Marcellin E, Nielsen LK, Abeydeera P, Krömer JO (2009) Quantitative analysis of intracellular sugar phosphates and sugar nucleotides in encapsulated streptococci using HPAEC-PAD. Biotechnology journal 4: 58-63.
9. Korman A, Oh A, Raskind A, Banks D (2012) Statistical methods in metabolomics. Methods in Molecular Biology (Clifton, NJ) 856: 381-413.
10. Du P, Stolovitzky G, Horvatovich P, Bischoff R, Lim J, et al. (2008) A noise model for mass spectrometry based proteomics. Bioinformatics 24: 1070-1077.
11. Laatikainen R, Niemitz M, Malaisse WJ, Biesemans M, Willem R (2005) A computational strategy for the deconvolution of NMR spectra with multiplet structures and constraints: Analysis of overlapping 13C-2H multiplets of 13C enriched metabolites from cell suspensions incubated in deuterated media. Magnetic resonance in medicine 36: 359-365.
12. Nickerson RS (1998) Confirmation bias: A ubiquitous phenomenon in many guises. Review of General Psychology; Review of General Psychology 2: 175.
13. Hinkelmann K, Kempthorne O (2007) Design and Analysis of Experiments, Introduction to Experimental Design: Wiley-Interscience.
14. Tamhane AC (1977) Multiple comparisons in model I one-way ANOVA with unequal variances. Communications in Statistics-Theory and Methods 6: 15-32.
15. Welch BL (1947) The generalization of student's' problem when several different population variances are involved. Biometrika: 28-35.
16. Gross J (2003) Nortest: tests for normality. R package version 1.0. University of Dortmund, Dortmund, Germany.
17. Shapiro SS, Wilk MB (1965) An analysis of variance test for normality (complete samples). Biometrika 52: 591-611.
18. Stephens MA (1974) EDF statistics for goodness of fit and some comparisons. Journal of the American Statistical Association: 730-737.

9

19. Jüni P, Altman DG, Egger M (2001) Assessing the quality of controlled clinical trials. Bmj 323: 42-46.

20. Kaptchuk TJ (2001) The double-blind, randomized, placebo-controlled trial: gold standard or golden calf? Journal of Clinical Epidemiology 54: 541-549.

21. Hansson L, Hedner T, Dahlöf BÖR (1992) Prospective randomized open blinded end-point (PROBE) study. A novel design for intervention trials. Blood Pressure 1: 113-119.

22. Allen JR, Earp R, Farrell EC, Gruemer H (1969) Analytical bias in a quality control scheme. Clinical Chemistry 15: 1039-1044.

23. Link H, Anselment B, Weuster-Botz D (2008) Leakage of adenylates during cold methanol/glycerol quenching of *Escherichia coli*. Metabolomics 4: 240-247.

24. Haug K, Salek RM, Conesa P, Hastings J, de Matos P, et al. (2013) MetaboLights—an open-access general-purpose repository for metabolomics studies and associated meta-data. Nucleic Acids Research 41: D781-D786.

25. Fiehn O, Sumner LW, Rhee SY, Ward J, Dickerson J, et al. (2007) Minimum reporting standards for plant biology context information in metabolomic studies. Metabolomics 3: 195-201.

26. Sumner LW, Amberg A, Barrett D, Beale MH, Beger R, et al. (2007) Proposed minimum reporting standards for chemical analysis. Metabolomics 3: 211-221.

27. Goodacre R, Broadhurst D, Smilde AK, Kristal BS, Baker JD, et al. (2007) Proposed minimum reporting standards for data analysis in metabolomics. Metabolomics 3: 231-241.

28. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, et al. (2003) The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. Clinical Chemistry 49: 7-18.

29. Altman DG, Schulz KF, Moher D, Egger M, Davidoff F, et al. (2001) The revised CONSORT statement for reporting randomized trials: explanation and elaboration. Annals of internal medicine 134: 663-694.

30. Schulz KF, Altman DG, Moher D (2010) CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. BMC medicine 8: 18.

31. Karakach TK, Wentzell PD, Walter JA (2009) Characterization of the measurement error structure in 1D 1H NMR data for metabolomics studies. Analytica Chimica Acta 636: 163-174.

32. Pinheiro JC, Bates DM (2000) Mixed-effects models in S and S-PLUS: Springer Verlag.

33. Pinheiro J, Bates D, DebRoy S (2007) Linear and nonlinear mixed effects models. R package version 3: 57.

34. Cohen J (1988) Statistical power analysis for the behavioral sciences: Lawrence Erlbaum.

35. Lavagnini I, Magno F (2006) A statistical overview on univariate calibration, inverse regression, and detection limits: Application to gas chromatography/mass spectrometry technique. Mass spectrometry reviews 26: 1-18.

36. Razali NM, Wah YB (2011) Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. Journal of Statistical Modeling and Analytics 2: 21-33.

37. Fay MP, Proschan MA (2010) Wilcoxon-Mann-Whitney or t-test? On assumptions for hypothesis tests and multiple interpretations of decision rules. Statistics surveys 4: 1.

38. Ruxton GD (2006) The unequal variance t-test is an underused alternative to Student's t-test and the Mann–Whitney U test. Behavioral Ecology 17: 688-690.

39. Siegel S (1957) Nonparametric statistics. The American Statistician 11: 13-19.

40. McElduff F, Cortina-Borja M, Chan SK, Wade A (2010) When t-tests or Wilcoxon-Mann-Whitney tests won't do. Advances in Physiology Education 34: 128-133.

41. Van Belle G (2011) Statistical rules of thumb: Wiley-Interscience.

42. Lee HN, Marshall AG (2000) Theoretical maximal precision for mass-to-charge ratio, amplitude, and width measurements in ion-counting mass analyzers. Analytical chemistry 72: 2256-2260.

43. Manly BFJ (2006) Randomization, bootstrap and Monte Carlo methods in biology: Chapman & Hall/CRC.

44. Efron B, Tibshirani R (1986) Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. Statistical science 1: 54-75.

45. Bolker BM, Brooks ME, Clark CJ, Geange SW, Poulsen JR, et al. (2009) Generalized linear mixed models: a practical guide for ecology and evolution. Trends in ecology & evolution 24: 127-135.

46. Zhao Z, Kuijvenhoven K, Ras C, van Gulik WM, Heijnen JJ, et al. (2008) Isotopic non-stationary 13C gluconate tracer method for accurate determination of the pentose phosphate pathway split-ratio in *Penicillium chrysogenum*. Metabolic Engineering 10: 178.

47. Schäfer J, Strimmer K (2005) A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Statistical applications in genetics and molecular biology 4: 32.

48. Karakach TK, Knight R, Lenz EM, Viant MR, Walter JA (2009) Analysis of time course 1H NMR metabolomics data by multivariate curve resolution. Magnetic Resonance in Chemistry 47: S105-S117.

49. Pavese F, Ber M, Forbes AB (2009) Advanced Mathematical and Computational Tools in Metrology and Testing Viii: World Scientific Publishing Company Incorporated.

50. Clifford AA (1973) Multivariate error analysis: A handbook of error propagation and calculation in many-parameter systems: Applied Science Publishers.

51. Hamilton WC (1964) Statistics in physical science. Estimation, hypothesis testing, and least squares. New York: Ronald Press, 1964 1.

52. Tellinghuisen J (2001) Statistical error propagation. The Journal of Physical Chemistry A 105: 3917-3921.

53. Draper NR, Smith H, Pownell E (1966) Applied regression analysis: Wiley New York.

54. Metropolis N, Ulam S (1949) The monte carlo method. Journal of the American Statistical Association 44: 335-341.

55. Liu JS (2008) Monte Carlo strategies in scientific computing: Springer.

56. Cox MG, Siebert BRL (2006) The use of a Monte Carlo method for evaluating uncertainty and expanded uncertainty. Metrologia 43: S178.

57. (2012) MATLAB Documentation. Natick, MA: The MathWorks Inc.

58. Ihaka R, Gentleman R (1996) R: a language for data analysis and graphics. Journal of computational and graphical statistics: 299-314.

59. Knuth DE (2007) Seminumerical algorithms.

60. Knuth DE (2006) The art of computer programming: addison-Wesley.

61. Bindel D, Goodman J (2009) Principles of Scientific Computing.

62. Cox M, Harris P (2008) Software specifications for uncertainty evaluation: National Physical Laboratory.

63. Wichmann B, Hill I (2006) Generating good pseudo-random numbers. Computational Statistics & Data Analysis 51: 1614-1622.

10

64. Massey Jr FJ (1951) The Kolmogorov-Smirnov test for goodness of fit. Journal of the American Statistical Association 46: 68-78.

65. Young IT (1977) Proof without prejudice: use of the Kolmogorov-Smirnov test for the analysis of histograms from flow systems and other sources. Journal of Histochemistry & Cytochemistry 25: 935-941.

66. Vale CD, Maurelli VA (1983) Simulating multivariate nonnormal distributions. Psychometrika 48: 465-471.

67. Headrick TC, Sawilowsky SS (1999) Simulating correlated multivariate nonnormal distributions: Extending the Fleishman power method. Psychometrika 64: 25-35.

68. Iman RL, Conover W (1982) A distribution-free approach to inducing rank correlation among input variables. Communications in Statistics-Simulation and Computation 11: 311-334.

69. Wiechert W, Siefke C, deGraaf AA, Marx A (1997) Bidirectional reaction steps in metabolic networks .2. Flux estimation and statistical analysis. Biotechnology and Bioengineering 55: 118-135.

70. Buckland ST (1984) Monte Carlo confidence intervals. Biometrics: 811-817.

71. Spearman C (1904) The proof and measurement of association between two things. The American journal of psychology 15: 72-101.

72. Engl HW, Flamm C, Kügler P, Lu J, Müller S, et al. (2009) Inverse problems in systems biology. Inverse Problems 25: 123014.

73. Tarantola A (2005) Inverse problem theory and methods for model parameter estimation: Society for Industrial Mathematics.

74. Lee JM, Gianchandani EP, Papin JA (2006) Flux balance analysis in the era of metabolomics. Briefings in Bioinformatics 7: 140.

75. Wahl S, Nöh K, Wiechert W (2008) 13C labeling experiments at metabolic nonstationary conditions: An exploratory study. Bmc Bioinformatics 9: 152.

76. Niklas J, Schneider K, Heinzle E (2010) Metabolic flux analysis in eukaryotes. Current opinion in biotechnology 21: 63.

77. Le Novere N, Bornstein B, Broicher A, Courtot M, Donizelli M, et al. (2006) BioModels Database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems. Nucleic Acids Research 34: D689-D691.

78. Li C, Courtot M, Le Novère N, Laibe C (2009) BioModels. net Web Services, a free and integrated toolkit for computational modelling software. Briefings in Bioinformatics 11: 270-277.

79. Engl HW, Hanke M, Neubauer A (1996) Regularization of inverse problems: Springer.

80. Groetsch CW (1984) The theory of Tikhonov regularization for Fredholm equations of the first kind: Pitman Boston.

81. Natterer F (1984) Error bounds for Tikhonov regularization in Hilbert scales. Applicable Analysis 18: 29-37.

82. Bates DM, Watts DG (2008) Nonlinear regression: iterative estimation and linear approximations: Wiley Online Library.

83. Antoniewicz MR, Kelleher JK, Stephanopoulos G (2006) Determination of confidence intervals of metabolic fluxes estimated from stable isotope measurements. Metabolic Engineering 8: 324-337.

84. Radrich K, Tsuruoka Y, Dobson P, Gevorgyan A, Swainston N, et al. (2010) Integration of metabolic databases for the reconstruction of genome-scale metabolic networks. BMC Systems Biology 4: 114.

85. Schilling CH, Letscher D, Palsson BO (2000) Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective. Journal of theoretical biology 203: 229.

86. Wiechert W, de Graaf AA (1997) Bidirectional reaction steps in metabolic networks: I. Modeling and simulation of carbon isotope labeling experiments. Biotechnology and bioengineering 55: 101-117.

87. Suthers PF, Burgard AP, Dasika MS, Nowroozi F, Van Dien S, et al. (2007) Metabolic flux elucidation for large-scale models using ^{13}C labeled isotopes. Metabolic Engineering 9: 387-405.

88. Goudar CT, Biener R, Konstantinov KB, Piret JM (2009) Error propagation from prime variables into specific rates and metabolic fluxes for mammalian cells in perfusion culture. Biotechnology progress 25: 986-998.

89. Fan TW-M, Lane AN, Higashi RM (2004) The Promise of Metabolomics in Cancer Molecular Therapeutics. Current Opnion in Molecular Therapeutics 6: 584-592.

90. Lane AN, Fan TW, Higashi RM (2008) Isotopomer-based metabolomic analysis by NMR and mass spectrometry. Methods Cell Biol 84: 541-588.
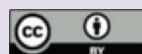
91. Fan TWM, Lorkiewicz P, Sellers K, Moseley HNB, Higashi RM, et al. (2012) Stable isotope-resolved metabolomics and applications for drug development. Pharmacology & Therapeutics 133: 366.

92. Szyperski T (1998) 13C-NMR, MS and metabolic flux balancing in biotechnology research. QUARTERLY REVIEWS OF BIOPHYSICS 31: 41-106.

93. Schellenberger J, Zielinski DC, Choi W, Madireddi S, Portnoy V, et al. (2012) Predicting outcomes of steady-state 13C isotope tracing experiments using Monte Carlo sampling. BMC Systems Biology 6: 9.

94. Nöh K, Wiechert W (2006) Experimental design principles for isotopically instationary 13C labeling experiments. Biotechnology and Bioengineering 94: 234 - 251.

95. Crown SB, Ahn WS, Antoniewicz MR (2012) Rational design of 13C-labeling experiments for metabolic flux analysis in mammalian cells. BMC Systems Biology 6: 43.

96. Metallo CM, Walther JL, Stephanopoulos G (2009) Evaluation of 13C isotopic tracers for metabolic flux analysis in mammalian cells. Journal of Biotechnology 144: 167-174.

97. Wang NS, Stephanopoulos G (1983) Application of macroscopic balances to the identification of gross measurement errors. Biotechnology and bioengineering 25: 2177-2208.

98. Van der Heijden R, Romein B, Heijnen J, Hellinga C, Luyben K (1994) Linear constraint relations in biochemical reaction systems: II. Diagnosis and estimation of gross errors. Biotechnology and bioengineering 43: 11-20.

99. Palsson B (2000) The challenges of in silico biology. Nature Biotechnology 18: 1147-1150.

100. Akaike H (1974) A new look at the statistical model identification. IEEE Transactions on Automatic Control 19: 716-723.
101. Moseley HNB, Lane A, Belshoff A, Higashi R, Fan T (2011) A novel deconvolution method for modeling UDP-N-acetyl-D-glucosamine biosynthetic pathways based on 13C mass isotopologue profiles under non-steady-state conditions. BMC Biology 9: 37.
102. Hiller K, Metallo CM, Kelleher JK, Stephanopoulos G (2010) Nontargeted elucidation of metabolic pathways using stable-isotope tracers and mass spectrometry. Analytical chemistry 82: 6621-6628.