Automating Selection for Incurred Sample Reanalysis: A Meta-Analytical-Based Algorithm for Error Reduction

Jack Rogers, Jessie Allen Ph.D., Yoka Thomas, Cheikh Kane Ph.D. – KCAS Bio, Olathe, Kansas 66061 USA

INTRODUCTION

Recommendations for incurred sample reanalysis (ISR) were initially formalized at the Third AAPS/FDA Bioanalytical Workshop in 2006 to address concerns and challenges related to reproducibility of results for quantitative bioanalytical methods. These recommendations have evolved over nearly two decades, culminating most recently in the ICH M10 guidelines published in 2022. The M10 guidelines provide recommendations for high-level strategy in selecting and performing ISR batches, whereas the clarity on which method to use for the ISR selection is left to the bioanalytical laboratory decision. Specifically, the ICH makes three overarching recommendations: that ISRs be selected 1) to be "representative" of the whole study, 2) to represent "adequate coverage of the concentration profile," including around Cmax and in the elimination phase, and 3) "as randomly as possible". Considering the current regulatory landscape on ISRs, it is required for the bioanalytical laboratory to formalize this selection process in their internal guiding documents. In light of this and to alleviate a potential bias in the manual selection of ISRs, an algorithm with predetermined criteria for each of these three primary considerations would ensure the quality and consistency of ISR selections, thus increasing the likelihood of the detection of anomalous results. This project analyzed a set of pre-existing ISR selections as a baseline for potential errors or biases, then developed a program to dynamically adjust its selection process to account for the three overarching recommendations listed above, and finally used the program to retroactively re-select ISRs for a number of previously completed studies to recalculate and compare quality metrics.

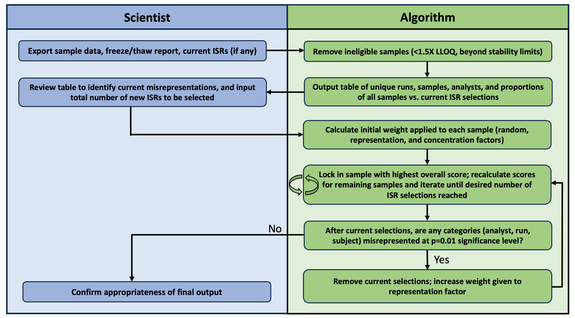

Figure 1. Distribution of activities between the scientist and ISR selection algorithm.

NOVELTY

The Global Bioanalysis Consortium Harmonization Team identified selection of ISR samples (whether chosen manually, entirely randomly, or via a pre-defined algorithm) as an approach needing refinement, and while various facets of incurred sample reanalysis continue to be an active area of discussion in WRIB meetings and their annual white papers, there is still a lack of general consensus on the more refined details of the mode for selection. To the best of the authors' knowledge, there does not exist in the literature either an in-depth examination of manual ISR selections compared directly to those generated by a pre-defined algorithm, or data-driven discussion of parameter selection intended to best balance the various recommendations for ISR selection included in the 2022 ICH M10 guidance. To address this need, the current project developed and characterized a novel algorithm for automated ISR selection building off both the ICH M10 guidance and a meta-analysis of ISR selections from previously performed bioanalytical studies.

METHODS

Analysis was divided broadly into two phases – a retrospective analysis of existing ISR data, spanning 30 studies with a combined total of approximately 47,000 samples, followed by reselection of ISRs for a subset of these studies to demonstrate the efficacy of the generated program. To maximize accessibility for intended users, the chosen language and application for all portions of the project were Visual Basic for Applications (VBA) and Excel, respectively. To accomplish the first phase of analysis, a program to evaluate pre-existing ISR selections was written. It takes four inputs exported directly from Watson LIMS: an inventory of all samples, a freeze/thaw report, a run summary report, and an ISR report. These are compiled and then analyzed by the program to generate tables for expected vs. actual representation. Here, "representation" requires that both the pre-clinical or clinical study (i.e., subject distribution, with corresponding categories like gender and cohort) and the bioanalytical study (i.e., original sample analysis runs and original analysts) occur proportionally amongst selected ISRs compared to the overall ratios. The threshold for significance was established at the p=0.01 level.

Due to the small number of samples in some instances, particularly subjects where (n) is sometimes 10 or less, a standard approach assuming normality is inappropriate; rather, an exact null distribution is generated for each individual run, analyst, and subject according to:

$$p = \sum_{j=0}^{k} \binom{n}{j} {p_i}^j (1-p_i)^{n-j}$$

where p is the calculated test statistic, pi is the fraction of the individual observed among all samples in the study, n is the total number of ISR samples, and k is the number of ISR samples currently observed for the individual.
Two metrics are then calculated to quantify degrees of misrepresentation: a "breadth error," which is penalized for the total relative amounts of subjects, runs, and analysts which are misrepresented at the p=0.01 level, and a "depth error," which scales with the most extreme p- value. These are represented by:

$$E_B = 1 - \sum_{i=1}^{3} \min(\max(e_i - 0.01, 0), \frac{1}{3}) \qquad E_D = -\log_{10}(\max(p_1, p_2, p_3))$$

where E_B and E_D are breadth error and depth error, respectively, $\boldsymbol{e}_i$ is the percentage of individuals in each of the three considered categories (subjects, runs, and analysts) which have p-values smaller than 0.01, and pi is the smallest (or in other words, most extreme) p-value in the three categories. As an example, a study which has 5% of subjects, 10% of runs, and 0% of analysts misrepresented at the p=0.01 significance level, with the most sever p-value being a run with a p-value of 0.005 would have an E_B of 1–0.04–0.09–0=0.87, or 87%, and a E_D of 2.30. Descriptively, a study which has a large percentage of subjects misrepresented, but not necessarily at an extreme level, would have a larger breadth error, while one which has one or more isolated instances of egregious misrepresentation would have a larger depth error score, even if overall most subjects, runs, and analysts were represented appropriately.

For the second phase of analysis, involving reselection of ISRs for existing studies, a second program was created, taking the same four inputs exported from Watson LIMS. After removing ineligible samples, such as those with a reportable result below the limit of quantitation or samples excluded due to other analytical reasons, the program assigns an overall weighting to each sample based on three factors: 1) a representative factor, combining in equal parts the degree to which the study subjects, original analysis runs, and original analysts are currently under- or over-represented among the existing ISR selections, 2) a concentration factor, which gives a slight prioritization to samples near C_{max} and elimination phase (where the top and bottom quintiles of concentrations observed for each individual subject are used as surrogates), and 3) a random factor. These factors are combined to create an overall prioritization score for the first sample selection, according to: Each score is subsequently recalculated, and the script iterates until the pre-determined number of ISRs have been selected. Afterwards, the program evaluates the entire set to ensure that study subjects, runs, and analysts are equally represented, at a threshold of p=0.01. If a misrepresentation in any category is detected, the original selections are rejected, and the script is run again, increasing the weighting of the "representative" score while decreasing the weight of the "concentration" and "random" scores. Total iterations were set at five, for which cycle the "representative factor" accounted for 90% of the overall score for each sample.

Ten studies were identified for reselection and subsequent characterization: five small (defined any less than 1000 total samples) and five large (defined as any greater than 1000 samples). Within each set of five, four were chosen which displayed subjectivity in one or more category of subjects, runs, and analysts, with the fifth being a well-represented study in all 3 categories to act as a control. ISRs were reselected in a simulated fashion (e.g., piecewise throughout the study, rather than at one large batch at the end) to demonstrate comparability of results had the program been employed from study outset. Afterwards, one study was chosen as a case study for further characterization of program output and suitability. This includes reproducibility upon repeated execution of the program, and the ability of the program to correct imbalances in ISR selection had the program been introduced only partway through sample analysis.

RESULTS

During retrospective analysis, ISR selections from 30 existing studies were characterized for accurate representation. These studies ranged in size from less than 200 to over 8000 total samples, and in sample analysis length from less than two weeks to multiple years. Study length, study size, and number of ISR selection dates were correlated with overall ISR passing percentage, breadth error score, and depth error score (as defined in Methods). ISR passing percentage did not correlate at a significant level (all p>0.05) with these metrics. However, the breadth error score correlated positively with the number of ISR selection dates (p=0.033), and the depth error score correlated positively with study length, study size, and number of ISR selection dates (p=0.013, p=$5.5x10^{-3}$, and p=$8.6x10^{-4}$, respectively; see Table 1). These results indicate that the potential for having increasingly imbalanced choices as studies increase in length and/or number of samples, as well as studies which have ISRs picked more frequently. These correlations demonstrate the opportunity for a more methodical and automated approach to ensure more representative choices.

	ISR Passing %	E_B (Breadth Error)	E_D (Depth Error)
Study Length	3.1E-01	1.4E-01	1.3E-02*
Study Size	4.1E-01	5.8E-02	5.5E-03**
# of ISR Selection Dates	3.7E-01	3.3E-02*	8.6E-04**

Table 1. Correlations between performance metrics and study characteristics. * = significant at p=0.05 level, ** = significant at p = 0.01 level.

Of the ten studies where ISR selections were simulated from study outset, each resulted in appropriately proportional selections. Two of the ten studies displayed a small number of subjects (less than 2% for each) which were over-represented (again at a significance level of p<0.01) in the final selection. This likely resulted from the boost applied for the representation factor described in the Methods above, which gives a prioritization to subjects, runs, or analysts which are currently under-represented but does not inherently penalize those which are over-represented. This design choice preferred to correct under-representation, which was deemed to be more problematic than over-representation. For the eight studies which had misrepresentations in one or more categories, both error scores were reduced significantly (p=$5.52x10^{-3}$ and p=$1.11x10^{-2}$, respectively), while the control studies remained relatively equal.

One of these ten studies was chosen for further in-depth characterization. The chosen study had approximately 3,000 samples, with relatively large breadth and depth error scores. Different parameter options for the equation used for determining weight based off calculated t- scores all produced acceptable ISR selections over ten independent selections for each combination, though each parameter combination produced different distributions for skewing towards the upper and lower concentrations favored by the "concentration score." These results provide a basis for final decision of parameters based off the desired average concentration distribution. Additionally, the program simulated the introduction of automated selection partway into a bioanalytical study by inputting the first half of ISR selections chosen manually while allowing the program to determine the latter half (see Figure 2).

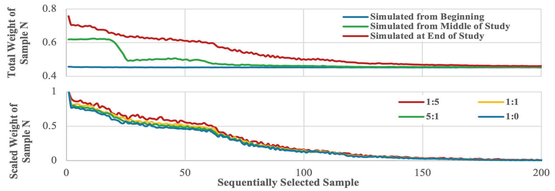

Figure 2. Diagnosing study misrepresentations by tracking total weight over course of sample selection. Top: Total weight assigned at each sample simulated at three different points in the same study. Bottom: Scaled weight assigned at each sample for different ratios of Proportional Factor:Random Factor; 1:0 signifies weight score entirely determined by the proportional factor. Each line is the average of 10 simulations.

In these iterations, the program was able to correct for under-representations appropriately, though runs and subjects which were greatly over-represented in the first half of ISRs selected in the actual study could only be statistically corrected by choosing a much larger set of ISRs. These results demonstrate the reproducibility of acceptable selections, as well as the utility for introducing an automated selection method into ongoing studies.

To test the ability of F_{conc} (the percent weight given to the concentration factor which boosts samples in the upper and lower quintile of concentrations observed within each individual subject) to modulate the concentration distribution of samples selected, the case study was reselected from scratch with six different weights values ranging from 0 to 0.25, with ten iterations performed at each assigned weight. Representative and random factors were equal to each other at each value of F_{conc} and scaled for total weight to always be equal to 1. Figure 2 demonstrates the scaling influence of this weighting factor; as F_{conc} increases towards 0.25, nearly all of the samples selected fall in the upper and lower quintiles. Note that because some subjects have less than 10 samples total, the 1st and 10th quintile are artificially inflated at each level (e.g., if a subject only has 4 samples and the lowest concentration was selected, this was categorized as the lowest decile). These tests demonstrate a high degree of control permitted to the bioanalytical laboratory to decide a distribution which they believe to satisfy the spirit of the ICH M10 recommendations towards prioritizing samples near Cmax and the elimination phase with balancing the need for adequate coverage of the entire concentration profile. Based off these results, a value of F_{conc} equal to 0.10 was chosen for current and future testing.

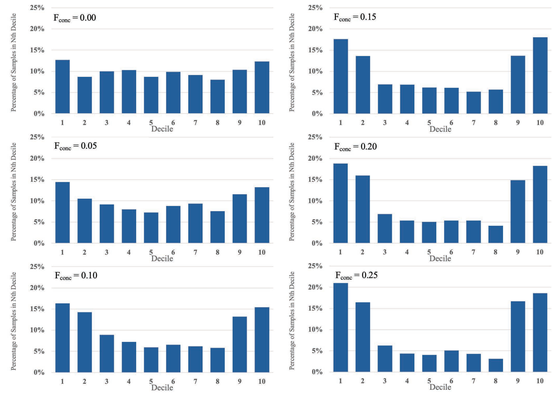

Figure 3. Modulation of sample concentration distribution via F_{conc}; weight applied for samples in upper and lower quintiles.

CONCLUSIONS

While high-level recommendations exist in the literature for selection of ISRs, there are nevertheless different practices on how to implement these recommendations and which method to use. During this retrospective analysis of 30 studies, we found that there was a spectrum of potential misrepresentations across subjects, analysts, and runs. This demonstrated a need for standardization of the selection process with a rigorously defined method or algorithm, which would consider each of the overarching ICH M10 recommendations and industry best practice. Upon simulating ISR reselection for these studies, we have observed major improvements in proportionality. Furthermore, the potential for inappropriate ISR selection (e.g., in selecting samples which were BQL, or inadvertently repeated in ISR) as well as human bias can be greatly reduced. By utilizing a deductive approach to ensure quality of ISR selections, the intent to demonstrate reproducibility and consistency of sample results can be better supported.