Debrief : Cycles xx.xx — xx.xx

Prepared for:

Prepared by:

Proposal number: 000-001

Debrief -- Dedicated to you who asked me about infection control models.

Here are some brief considerations for life science and infection control problems in the context of the systems we worked on. I debated writing this debrief at all because, parts of this document appear extremely self-evident. Yet, we stumbled around with our own Data Collection and Analysis so I decided to write it anyways so that mistakes may be avoided in the future.

SUMMARY

Data Collection & Analysis Planning:

As of December 31st, 2018, FDA CGLP does not explicitly require a data analysis plan before the initiation of a GLP non clinical study. See https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/CFRSearch.cfm?CFRPart=58 for more information. As such, there is a temptation for smaller scale studies to generate skeletal data analysis plans. Only when a study produces unexpected results, does a data analysis plan suddenly become critical to project success. In those scenarios, a skeletal data analysis plan quickly starts to look like a critical error.

Controlling randomness:

It is usually impossible to generate results that encompass all combinations of independent variables in real world studies. However, there are study designs, structures and procedures which can reduce the number of combinations involved.

For instance, experiment outcomes are often linked to the quality of the study's Standard Operating Procedure (SOP) and the closeness by which the SOP is followed. These two factors directly relate to the specific researcher executing the procedure. Personnel schedules should be adjusted so that meaningful data can be extrapolated, regardless of whether the result is a true or false positive. It is easier to extrapolate the cause of a positive result if there are 3 researchers who perform a procedure, rather than 20. Not only is it significantly easier to train 3 people, but having such a large pool of people performing a procedure makes it to differentiate between operator error(false positives) and true positives.

This seems self-evident but all data analysis plans should keep randomness(variance) at a minimum where feasible. Other trivial examples include specifying the exact machine used for an analysis if multiple are in hand, or using different brands of a medical product if they vary significantly in design.

Feature Engineering:

Feature engineering after data collection can be subject to systematic bias. It is sometimes unavoidable that the people collecting the data are also exposed to the environment where the data is collected or are even the ones analyzing it. *(This is why clinical trials have blinded analysts and statisticians.)*

Perceived rankings of certain independent variables can be influenced by the collected results (confirmation bias). Making matters worse, the people most equipped to rank these variables are often exposed to the quantity or a derivative of the quantity that we are trying to estimate. How are physicians supposed to be able to rank the quality of practice, of other physicians, without observing them practice? There is a need for features to be created in the data analysis plan to prevent confirmation bias.

Examples include ranking batches of chemicals, batches of bacteria, binning environmental variables such as temperature and pH/kH, or transforming environmental variables into continuous, ranked variables. (For example, ATCC strain #999 grows best at ph7. Consider transforming pH into the absolute value of pH7 – actual pH)

Many features are the results of statistical processes and have high variance. We know from research and experience that colony counts from a single plate are not particularly reliable estimators. https://www.microbiol.org/wp-content/uploads/2010/07/Sutton.jvt_.2011.17_3.pdf Thus it often makes sense to aggregate or combine results to create a new metric. Metrics created purely from collected data without subjective human input are less likely to be biased, and are thus less critical to identify before the execution of the study.

Data Quality:

As datasets have grown, we have observed that collinearity between independent variables has on occasion, changed significantly -- without obvious explanation. This suggests that even in medium sized datasets, variance can be high. The pathology of acquiring an infection or a positive bacteria plating result is very diverse and may not always be fully reflected in a dataset. In controlled settings such as ours(GLP-ish), the same effect has been observed.

It also makes it hard to remove variables because data which seems to be noise, may not be in specific circumstances. This is especially concerning in the case of imbalanced classes. (Often the case with infection control data)

Data Interpretation:

Good study design can help to reduce collinearity. The example listed above of reducing and rotating researchers performing a procedure demonstrates this. Sometimes collinearity is unavoidable and can make interpreting results very difficult as it becomes necessary to interpret correlation vs causation. If bacteria counts are high and pH is correspondingly low, is low pH encouraging bacteria growth? Or is the fall in pH just a byproduct of bacteria growth? Is low pH actually inhibiting bacteria growth? Some examples are much harder to interpret. (e.g. Physician consulted vs Stage of Infection – One conclusion might suggest that a physician is an expert who is consulted in the most difficult cases, the other might suggest sub-par practice – Completely opposite conclusions!) Feature reduction or removal becomes much more subjective and difficult.

Keeping the data which appears to be the independent variable has served us well in practice. In the case of the example above where the focus is on patient outcomes, the independent variable would be the stage of infection. The rationale is that one can use standard risk management techniques or failure mode analysis techniques to ask the question: “What happened at that particular stage of infection that led to bad outcomes?” An investigation can then determine if dependent variables such as the physician consulted, played a significant role in affecting the outcome.

RESULTS

Raw Data

***

TIMELINE

***

OBSERVATIONS

***

IMPROVEMENTS

***

RECOMMENDATIONS

***

Recommendations: See xx.xx.xxxx—2017.01.31