

Splendida fugilente oculi, e.g. dummy randomization lists to more than designed patient profiles

Magnus Mengelbier, Limelogic project, London, United Kingdom

ABSTRACT

There are many different and popular programming techniques to protect the integrity of open label and blinded clinical trials. We consider common approaches from a program development point of view where both conscious and unconscious influence may originate from generating draft added-value data sets and report output. Dummy randomization lists, rescaling, recycling, designed subject response profiles and data scrambling all aim to hide and obscure any true treatment differences. We also consider experiences from clinical study teams to support trial integrity in order to offer sensible and applicable tactics.

INTRODUCTION

There are many different approaches employed to hide or make study results difficult to determine prior to unblinding, locking away a study database, or final analysis. The underlying reasoning is most often founded on business needs, ethical considerations, or both in an effort to protect the results from external and internal bias. We shall in brief consider five different techniques to support the development process of output deliverables prior to any form of unblinding.

The most frequent and commonly preferred approach is to produce and incorporate a false allocation. The allocation is designed to protect the subsets we are interested in. In some instances where subject responses may provide subtle to clear indications as to what treatment has been received, rescaling of sensitive results or a designed subject profile for a predefined data subset is usually the tool of choice.

Rare occasions may require more extensive attempts to hide results and study outcomes. Provided that disjoint development teams or development after database lock is not a viable option, a study team can go to extreme lengths in order to provide a sensible development environment. Rather than simulating part of or a complete study database, study teams can enact the role of re-use librarians by recycling data from previously unblinded or reported studies. In contrast, we also discuss data scrambling, a comprehensive technique based on the philosophy of designed subject profiles.

There are several additional considerations that will determine a preferable technique as each one has limitations. As the efforts increase, so do other aspects of study management, coordination, communication, and activities.

ASSUMPTIONS AND AIM

The central assumption is that the data represents a potential and significant difference between two or more subsets of interest along the aim of the study design as well as outside of the primary endpoints. At the same time, we assume that the same difference is sensitive to influence. Most often, the central assumption is in relation to the treatment arms and other endpoints of interest and may only be applicable to one or more analysis subsets.

Provided our main assumption that the study data contains inherent relationships, our aim will be to hide or mask any relationships, differences or at least the level of significance such that development output will not seem to directly reveal or allude to the study results or violate any other integrity constraints. To fully eliminate relationships can be extremely difficult as analysis rely on many key factors and study team members retain detailed knowledge of the product from previous trials as well as ongoing safety reviews.

DUMMY RANDOMIZATION LISTS

A classic, popular approach is to establish a false treatment allocation. The goal is to swap entire patient profiles across the subsets of interest where most often this includes the study treatment arm, study phases or, if possible, analysis populations.

A most convenient approach is to list subject numbers in a random, ascending or descending order while cyclically allocating group identities. We discuss identities instead of solely the treatment arm as we may wish to protect the study phase or analysis population.

For example, a treatment arm identity can be assigned as

$\text{treatment}_1, \text{treatment}_2, \dots, \text{treatment}_n$

for $\text{subject}_1, \text{subject}_2, \dots, \text{subject}_n, \text{subject}_{n+1}, \text{subject}_{n+2}, \dots, \text{subject}_{2n}, \text{subject}_{2n+1}, \dots$

Adapting the method of allocating subject identities can vary the number within the different subsets. Two common random based approaches utilize the uniform and exponential distributions. More advanced allocating techniques may have to be explored given more complex study designs and requirements.

The dummy randomization or subset list is a weak masking approach as you retain the whole subject profile. There are cases where the true treatment allocation was uncovered by simply performing simple statistical analysis or considering one or more on-study parameters. While reviewing subsequent events in these cases, it became clear that the study outcome had been influenced, even if the project members acted in good faith and under best intentions. In general, you will most likely be able to circumvent any false treatment allocation strategy as most studies may contain more than one or a group of on-study parameters, say laboratory test results, which are sensitive to the study treatment.

The false treatment allocation can also be undone if the method of generating the true treatment allocation is not taken into consideration. A programmer once designed a dummy randomization list by alternatively assigning the active treatment versus placebo to a sorted list of subject numbers. Unfortunately, this was how the true allocations were generated and posed a serious risk for the study credibility as the study was extremely sensitive to influence and crucial to the success of the product. A lesson learned is that any false treatment allocation is only as good as the method of the true treatment allocation scheme.

RESCALING AND DESIGNED SUBJECT PROFILES

Rescaling or designed subject profiles are effective when you wish to hide an entire response or lack response data altogether. It is entirely straightforward to create a profile depending on the distribution of values you wish to obtain, as the method is applicable for both continuous and categorical data.

A simple implementation of a designed subject profile would be rescaling results that are considered influential. The scalars and approach is most often more whim than extensively considered, as the values are not intended to be biologically plausible. Rescaling can be an effective contribution to dummy randomization lists as influential parameters are moved out of context. Unfortunately, moving a value out of scope may not be applicable if the values are members of associative relations, discreet values representing categories or used in any calculations other than reporting.

As with false allocations, rescaling can be circumvented. Consider a case where the rescaled value is summarized and the scalar is known. For a given value x , where x has been rescaled as $y = a \cdot x$, then the true mean and variance of x can most often be derived quite easily by using the mean and variance of y .

A more focused approach would be to design a specific data profile. As said profile can build on both pre-specified and randomly generated relationships, care can be taken to provide a semi-plausible model. The source of data can be fictive, but as companies amass larger volumes of clinical data, the re-use of already unblinded or reported data can be more effective as we discuss in the next section.

The approach of designed subject profiles has several positive and negative characteristics that can be easily controlled provided the data points are considered in context. The source for a specific profile can efficiently be used repetitively and manipulated accordingly to create a sought after treatment response.

As the data is generated and fit into the study design, it may reflect your expectations too closely and not reflect any issues with the true data. Unless taken into account, this can contribute its share of surprises when considering and locking away analysis models or when assembling the real analysis output into study reports. The latter most often occurs under tight timelines and resource constraints, so any disruption can have significant impact if the release of a database for a study has to be delayed. The unblinding process can also highlight unrealized and unknown data issues unless specific screening is in place for key factors and relationships.

RECYCLED CLINICAL DATA

Larger pharmaceutical and medical device companies amass significant volumes of both pre-clinical and clinical data as development programs progress. The granted re-use of already unblinded or reported data can become an effective and viable approach to replace or augment designed profiles, as the data itself does not have to be within the context of the study design. Although the data will most likely resemble the real thing, it is crucial that the data is understood to be fictive when reviewing results prior to unblinding.

The use of data banks within development programming is becoming more straightforward as standardization and the introductions of data capture and transfer templates proceed. Add the possibility to develop cases applicable to analysis model development and verification; the approach is an interesting option. The investment in developing data banks of recycled data is also resource efficient as the invested time can be recouped through pre-packaged, verified and simplified study programming setups.

Similar to designed subject profiles, any issues with the true data will be well hidden. As the data banks are developed, many organizations also focus to include standardized screening for key factors and relationships on the true data that is not reflected through the data bank.

DATA SCRAMBLING

Data scrambling was the answer to a sensitive pivotal study where the greater majority of study parameters would require a designed profile and other techniques were not a viable option. As a consequence of a growing reliance on designed profiles, the subsequent data structure tended to deviate significantly from the characteristics of the true data. Scrambling was designed to address the impact of deviation and continuous changes in the study database as it relies on the true data source as an input for valid values instead of an arbitrary value range.

The fundamental mechanism of scrambling is to make any response within a valid range equally likely to occur, regardless of treatment arm and subset and without having to coerce the sequence of nominal events. Each parameter is processed individually by its generic definition enabling us to create a data structure where the subject profile is inconsistent with the subject's true clinical events but valid within the context of the generic parameter definitions. Any important inter-relationships are constructed during post-processing.

We assume that the study database contain variations on three generic variable types; continuous, categorical or discrete, and text.

CONTINUOUS DATA

The value range $[x_{\min} + \epsilon_1, x_{\max} + \epsilon_2]$ for generating scrambled data for a continuous variable x is based on the input data range $[x_{\min}, x_{\max}]$. The noise factors ϵ_1 and ϵ_2 are user-defined proportions of the input data range.

The lower randomization bound $r_{\min}$ and upper randomization bound $r_{\max}$ are programmatically constructed as

$$r_{\min} = x_{\min} + \varepsilon_1 = x_{\min} + \text{noise} \cdot (x_{\max} - x_{\min}) \cdot (U_{[0,1]} - 0.5)$$

and

$$r_{\max} = x_{\max} + \varepsilon_2 = x_{\max} + \text{noise} \cdot (x_{\max} - x_{\min}) \cdot (U_{[0,1]} - 0.5)$$

where $0 \leq \text{noise} \leq 1$ is a user-defined parameter to adjust the randomization bounds, blur the edges if you will. Let $\text{noise} = 0$ to resolve the lower and upper randomization bounds equal to the true data minimum and maximum values, respectively.

The scrambled value $x_{\text{scrambled}}$ is derived as

$$x_{\text{scrambled}} = r_{\min} + (r_{\max} - r_{\min}) \cdot U_{[0,1]}$$

where $x_{\text{scrambled}}$ is in the interval from $r_{\min}$ to $r_{\max}$.

If the valid value range is restricted to positive numbers, the lower randomization bound $r_{\min}$ is forced to be greater or equal to zero.

CATEGORICAL DATA

The method to scramble categorical data values is based on a similar approach to that of continuous data. Each of the m number of valid discrete values for a variable x is assigned an ordinal number $n = 1, 2, \dots, m$. The scrambled value $x_{\text{scrambled}}$ is constructed such that

$$x_{\text{scrambled}} = x_q$$

where $q = f((m - 1) \cdot U_{[0,1]} + 1)$ and $f(z)$ returns the greatest integer of z .

The source for valid discrete values of x vary but will most likely be consistent throughout the study. To capture any non-standard values or in a simple way highlight any potential data issues, you can combine structures within the Case Report Form and current non-missing values rather than just solely relying on the expectations set out in the Case Report Forms.

TEXT, DESCRIPTIONS AND COMMENTS

Similar to categorical data, constructing a scrambled data subset of text is straightforward and relies on the same principle. Two important areas for consideration is the variable content as well as the contents association to other parameters within the same data set.

As an example, consider the MedDRA coded adverse event terms System Organ Class, High Level Group Term, ..., Preferred Term and Lowest Level Term. As the values are highly associated, the expectation is for a Preferred Term to most likely be associated with the same System Organ Class. If each variable is processed individually, irrespective of associations, the expected association will not be retained. Therefore, we employ two variations on the scrambling mechanisms for processing text.

The most simple method has each false word, phrase, or sentence assigned an ordinal value n such that $n = 1, 2, \dots, m$. The scrambled value $x_{\text{scrambled}}$ is then

$$x_{\text{scrambled}} = t_q$$

where $q = f((m - 1) \cdot U_{[0,1]} + 1)$, $f(z)$ returns the greatest integer of z and t is any entry in the list of false words, phrases or sentences. The associations mentioned above are not retained, as no formal relation between the true data value and the scrambled value is composed.

In order to retain the above-mentioned associations, scrambling is performed over the unique list of text strings x' for a variable x . Each of the m number false words, phrases, or sentences is assigned an ordinal value n such that $n = 1, 2, \dots, m$. The scrambled value $x'_{\text{scrambled}}$ is then

$$x'_{\text{scrambled}} = t_q$$

where $q = f((m - 1) \cdot U_{[0,1]} + 1)$, $f(z)$ returns the greatest integer of z and t is any entry in the list of false words, phrases or sentences. Using the relationship between x and x' , we can construct the scrambled text and retain any formal relations.

The false text mass used in the above two methods can be simple to complex and of any size. As the vocabulary origins can be the current texts, a recycled data library or a list of fictitious words, the choice should consider the sensitivity of specific phrases and specific needs for program development. The size and reliance on real or fictitious vocabulary are very important considerations.

The larger the text mass, the less likely any subset will have more than one occurrence. The list of false words, phrases or sentences should be constructed such that the resulting scrambled text is applicable. A limited vocabulary may be the most appropriate if summaries are to be performed, hence the adverse event example, as certain output may only report words, phrases, or sentences with ten or more occurrences within a group.

"Splendida fugilente oculi" can provide us with the fictitious vocabulary mentioned previously if the possibility of key words or phrases occurring may be significant. The vocabulary of words, phrases or sentences can be constructed to be dissimilar to English or any other language. As the phrase does not provide any indication of the underlying text, it can efficiently represent any response and observation for sensitive data. Any clear-text key words required for programming purposes can be randomly added in post-processing.

CONCLUSION

There are several approaches to swap, hide, mimic, replace, create, and scramble study data. Each has its advantages as well as several disadvantages. On most occasions, a simple dummy randomization list may suffice, but specific study requirements may require a more involved approach.

As efforts increase, other aspects of study management, coordination, communication, and activities will change in scope and scale. Common to all methods discussed is the increased focus and importance of planning, follow-up and quality assurance tasks

There are many factors that influence selection of an approach. Business requirements and ethical considerations are key drivers in protecting studies while providing the possibility to employ common study designs and capture instruments. Designed profiles, recycled data and the more comprehensive scrambling are techniques that can fulfill requirements not met by the ever-popular dummy randomization lists.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. The author can be contacted through www.limelogic.com/papers.