

# Fundamentals of Data Collection in Clinical Studies: Simple Steps to Avoid “Garbage In, Garbage Out”

The International Journal of Lower  
Extremity Wounds

1–5

© The Author(s) 2020



Article reuse guidelines:

sagepub.com/journals-permissions

DOI: 10.1177/1534734620938234

journals.sagepub.com/home/ijl



José G. B. Derraik, PhD<sup>1,2,3,4</sup>, Wason Parklak, PhD<sup>1</sup>,  
Benjamin B. Albert, MBChB, PhD<sup>2</sup>, Kongsak Boonyapranai, PhD<sup>1</sup>,  
and Kittipan Rerkasem, MD, PhD<sup>1,5</sup>

## Abstract

Many fundamental steps underpin the delivery of high-quality clinical research. In this article, we provide a brief commentary on some important aspects associated with the collection and management of data during clinical studies, which, if overlooked, will lead to poor-quality research. In particular, we discuss the key aspects that should help early career researchers maximize the relevance and impact of their clinical research.

## Keywords

accuracy, cleaning, collection, data, entry, error

Many steps are critical for the successful delivery of any clinical research project. These begin at conception of the hypothesis and research question(s), and also include study design, data collection, analysis, and interpretation. Above all, robust study design is paramount; if the study is not properly planned, it may be impossible to correct data collection issues once the study is already underway. Inevitably, a poorly designed study will not be sufficient to test the proposed hypothesis, and no statistician will be able to salvage the work afterwards. As the old saying goes, when it comes to data, garbage in equals garbage out.

Good clinical research starts from a plausible hypothesis supported by contemporary scientific knowledge that makes a testable prediction. The study must then be designed to test that prediction. To ensure that the study is robust and capable of testing the proposed hypothesis, clinical researchers should consult or involve a biostatistician during the early planning stages, so that input can be given into study design. Sample size is a key consideration, but a statistician can advise on other important aspects, such as the type of data to collect (eg, continuous vs discrete), proper randomization procedures, and how the data should be analyzed. Statisticians are not magicians, and none will be able to save a study whose data collection was not properly planned at the outset.

Optimizing the quality of the data collected is fundamental for high-quality research. In this short report, we provide a brief commentary on important aspects associated with the collection and management of data during clinical studies, which, if overlooked, will inevitably lead

to poor-quality research. In particular, we discuss some key aspects that should help early career researchers maximize the relevance and impact of their clinical research.

## Data Collection

Data form the foundations of any clinical research, and high-quality data can only be collected if the study is well designed. Thus, it is important to pay particular attention to certain aspects about data, such as opportunity for collection, quantity, quality, and power.

<sup>1</sup>NCD Center of Excellence, Research Institute for Health Sciences, Chiang Mai University, Chiang Mai, Thailand

<sup>2</sup>Liggins Institute, University of Auckland, Auckland, New Zealand

<sup>3</sup>Department of Women's and Children's Health, Uppsala University, Uppsala, Sweden

<sup>4</sup>Children's Hospital, Zhejiang University, School of Medicine, Hangzhou, China

<sup>5</sup>Department of Surgery, Faculty of Medicine, Chiang Mai University, Chiang Mai, Thailand

## Corresponding Authors:

José Derraik, Liggins Institute, University of Auckland, Private Bag, 92019 Auckland, New Zealand.

Email: jderraik@gmail.com

Kittipan Rerkasem, Division of Vascular and Endovascular Surgery, Department of Surgery, Faculty of Medicine, Chiang Mai University, Chiang Mai, 50200, Thailand.

Email: rerkase@gmail.com

### Opportunity, Quantity, and Quality

Many types of data can only be gathered once in a lifetime; thus, all opportunities to collect data must be considered, and the most appropriate chosen. For example, cord blood samples or clinical parameters at birth can only be collected in the delivery room. If the sampling opportunity is missed or mismanaged, the opportunity to collect such data is lost for the lifetime of the individual.

Similarly, where anonymous surveys are used, there is only one opportunity to collect data. It is impossible to reapply the survey or to clarify responses, as the participants cannot be contacted again. Thus, questionnaires must be thoroughly tried and tested before they are applied. Failure to include key questions, or including a question that is poorly written or difficult to answer cannot be corrected, and could therefore jeopardize the entire study.

Where biological samples are collected, especially in situations where the opportunity occurs only once, it is essential to ensure they are of sufficient quantity, and that the appropriate procedures are followed to ensure the best quality. For example, one must collect a sufficient volume of blood to enable measurement of all key biochemical parameters. In addition, appropriate procedures must be in place to ensure that samples are properly handled (eg, freezing at the appropriate temperature and within the required time frame).

However, while it is tempting to maximize the amount of data collected from participants, it is important to achieve the right balance. Collecting excessive data or samples can be burdensome for participants, and potentially lead to poorer quality responses, failure to complete the whole assessment, greater attrition rate, or a bad experience that reduces the likelihood of future participation in research. Therefore, proper consideration at the planning stages is important to achieve a balance. This would allow researchers to optimize the quantity and quality of the data to be collected, while making involvement in the study more acceptable for participants.

### Power

Early in the design of clinical studies, investigators must perform appropriate sample size and power calculations based on the study's hypothesis. Even if a study is generally well designed, it is of limited value if it is insufficiently powered to test its hypothesis. The purpose of the sample size calculation is to identify the minimum number of participants required to test the study's prediction about the primary outcome, that is, the key endpoint of a clinical study. This calculation can even help determine what the

primary outcome should be, as if a proposed endpoint would need an unattainable sample size, then a different endpoint may be more appropriate to select as the primary outcome.

An important issue to consider is that in some situations, the available sampling pool (ie, those potential participants who could be recruited) may be too small. This could apply to rare conditions or very specific target groups. In such cases, in order to increase the sample size (ie, the study's power) it may be necessary to sample over a longer period of time, or carry out multicenter studies within a region, a country, or even in different countries.

### Data Accuracy

#### Preventing Errors

Beyond study design, ensuring the accuracy of the data collected is paramount. First, detailed standard operating procedures must be prepared to safeguard not only the quality of individual sample collection, but also consistency in data collection. For instance, if different investigators are measuring anthropometric parameters (eg, height or waist circumference), it is fundamental that all measurements are made in a consistent manner. This is particularly important for longitudinal studies where the same participant is measured on more than one occasion.<sup>1</sup>

For each piece of data recorded, it is important to have considered the best way to obtain it. For example, where measurements are recorded (eg, height or blood pressure), errors can be reduced by repeating the measurement multiple times. A detailed discussion of individual techniques is outside the scope of this commentary, but many techniques have been covered by previous authors.<sup>1-3</sup> Nonetheless, beyond potential measurement errors, the greatest threat to data accuracy is human errors during data handling, in particular data entry.

In our experience, the most common causes of data entry errors are the following:

- *Misplacement of values*—eg, recording height where the weight should have been recorded, or vice-versa
- *Typographical errors*—eg, misplacing a decimal point or missing an integer
- *Wrong unit*—eg, entering height in meter instead of centimeter, or vice-versa, so that the unit of measurement is inconsistently recorded in the database

There is no doubt that the best strategy is to prevent these from occurring in the first place. Most software used for data entry will allow the creation of data

validation rules, which can prevent the vast majority of data entry errors. For example, if height is to be recorded in centimeter in a study involving adolescents, it is possible to create a validation rule that would prevent anyone from entering a value outside a predefined range, such as 100 to 220 cm. It would no longer be possible to enter a height of 1.65 m, as the system would prevent this action, greeting the user with an error message describing the required range (and unit). Data validation rules are extremely useful, but may not be applicable for parameters where it is not clear what the biologically plausible range is for a given population.

Beyond the adoption of data validation rules, we strongly encourage a 2-tiered approach for data entry. This means that once an investigator finishes the data entry for a given participant, a second investigator double-checks all values recorded in the system, before the data entry step is considered complete. Furthermore, where clinical data are being recorded by an investigator on hard copies, it is important that someone double-checks the information collected while the participant is still available, which would make it feasible to clarify any potential recording errors. The same applies to forms completed by the participants themselves.

Following data entry, human errors can also occur during data handling, which could have disastrous consequences for a clinical study. In particular

- *Translocation error*—eg, misplacing data when copying values from one spreadsheet into another, or even from one column/row into an adjacent column/row. This could, for example, lead to a mismatch between the data and the participant to which the data should refer, possibly having a flow-on effect on the data from multiple participants.
- *Data editing error*—eg, accidentally deleting a row, a column, or values during manipulation of electronic data.

Unfortunately, there is no simple solution to prevent these from happening. However, we encourage the creation of back-ups, whenever any major manipulation of data is deemed necessary.

### Identifying Errors

The first author of this commentary has been involved in the analyses of nearly 200 studies. To this date, he has not seen one single database that did not contain errors. Humans are not machines, and mistakes occur almost invariably, despite the adoption of safety measures such as data validation rules. Importantly, even in the absence of human error, biological samples can be compromised. For example, hemolysis of blood samples is a relatively

common occurrence in clinical practice, which can lead to changes in many biochemical parameters.<sup>4</sup> Therefore, even if the laboratory data are entered correctly, the data themselves may contain biologically implausible values where the samples or the biochemical analysis might have been compromised.

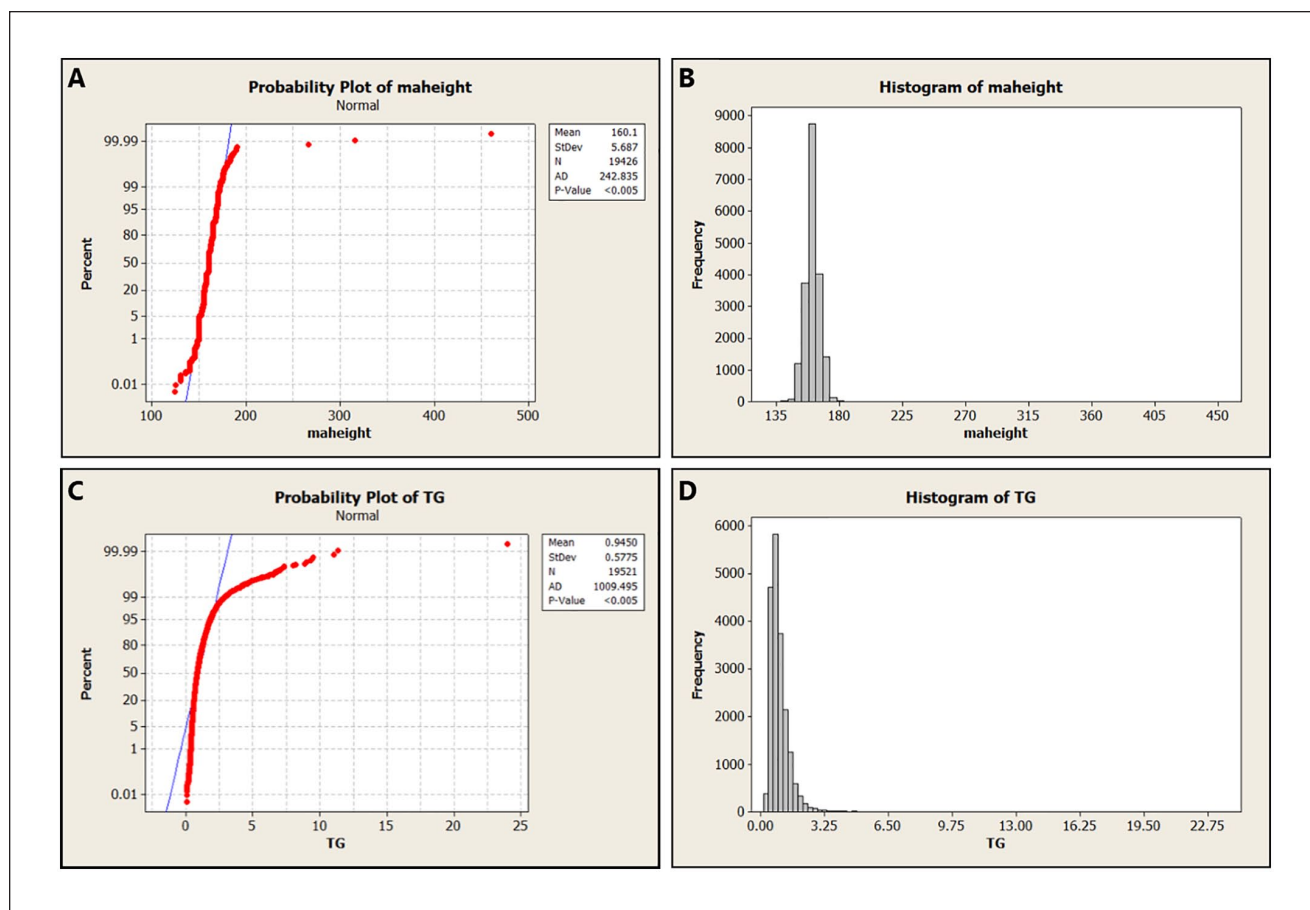
Irrespective as to the cause of data errors, it is fundamental that thorough data checking and ‘cleaning’ are carried out before the data are relayed to a biostatistician for analysis. This is particularly important where the biostatistician is someone outside the research team, or someone without a clinical background in your area of study. More often than not, the biostatistician will not be able to distinguish a real outlier from a biologically implausible value. Such distinction can only be made with good clinical understanding of the variable in question. As a result, it is necessary that members of the research team collaborate in data ‘cleaning’. This is important, as a value should not be deleted by the statistician or researcher just to make a scatter plot or a regression line look prettier; it is necessary that all stand-out values are individually examined for plausibility.

Visual examination of the data set is likely to be the easiest and most efficient way to identify potential errors. Simple histograms can be quickly created in most statistical software or other packages used for data management (such as Microsoft Excel). Normal probability plots are another alternative, where it is easy to identify extreme values.

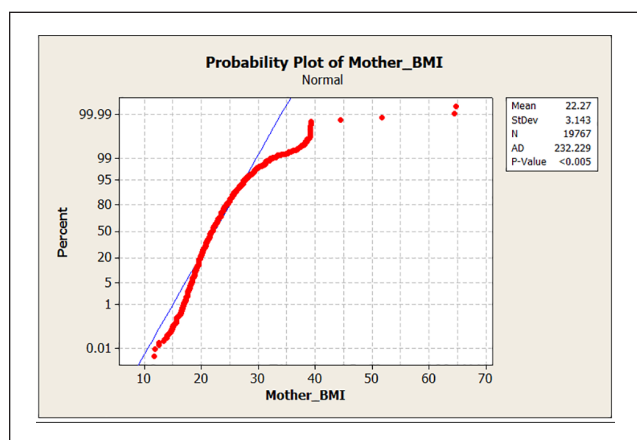
As an example, Figures 1A and 1B show the normal probability plot and the histogram, respectively, created from a data set on maternal height. Looking at both figures, it is possible to immediately identify that there are values for maternal height that are biologically implausible (ie, 260 to 460 cm).

However, simply looking at the shape and layout of a normal probability plot or histogram is not enough to identify or rule out data errors. For instance, the values for triglycerides shown in Figures 1C and D are all real, even though these graphs look very similar to Figures 1A and B. Thus, data ‘cleaning’ needs to be performed by (or with input from) someone with clinical understanding of each given parameter.

Importantly, for parameters that are calculated from other pieces of data such as body mass index (BMI) or Z-scores, it may not be possible to quickly identify data errors based just on visual inspection of a graph. It is necessary to examine the individual data from which the parameter was calculated, in order to ascertain whether these are biologically implausible. To illustrate this point, Figure 2 shows BMI data from a population of mothers. The 2 data points in the upper end of the spectrum represent nearly identical BMI values (64.39 and 64.74 kg/m<sup>2</sup>); however, while the latter was a real value (height = 174 cm; weight = 196 kg), the former was derived based on a



**Figure 1.** Examples of normal probability plots and histograms created during a data ‘cleaning’ process. (A, B) maternal height in centimeters; (C, D) triglycerides (mmol/l) from a paediatric population.



**Figure 2.** Normal probability plot of body mass index (BMI) data from a population of otherwise healthy mothers. The nearly overlapping BMI values above 60 kg/m<sup>2</sup> consist of a real BMI value and one derived from a data entry error.

height of just 89 cm, which was clearly a data entry error for this population (Figure 2).

## Conclusions

There is no doubt that proper planning is fundamental for the successful delivery of any clinical study. Failure to plan effectively risks the study being unable to answer the question it was set out to test, wasting the considerable investment of time, money, and participant goodwill. Rather than learning by making such costly mistakes, we strongly encourage early career researchers to carefully plan all stages of their study, from experimental design to data analysis, identifying any potential pitfalls, as well as the measures necessary to prevent problems that would compromise data collection and/or their accuracy. Here, we have briefly discussed a number of steps that can be taken (particularly during the early stages of a research project) that would minimize the

likelihood of missing crucial samples while also safeguarding the accuracy of any data collected.

### Acknowledgments

We thank Éadaoin Butler (Liggins Institute, University of Auckland) for editorial review of this manuscript. The figures were prepared using Minitab v16 (Pennsylvania State University, State College, PA).



### Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

### Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was supported in part by Chiang Mai University.

### ORCID iDs

Benjamin B. Albert  <https://orcid.org/0000-0003-0498-3473>  
Kittipan Rerkasem  <https://orcid.org/0000-0003-0784-2449>

### References

1. Whitney CW, Lind BK, Wahl PW. Quality assurance and quality control in longitudinal studies. *Epidemiol Rev.* 1998;20:71-80.
2. Lane SJ, Heddle NM, Arnold E, Walker I. A review of randomized controlled trials comparing the effectiveness of hand held computers with paper methods for data collection. *BMC Med Inform Decis Mak.* 2006;6:23.
3. Nahm M. Data quality in clinical research. In: Richesson R, Andrews J, eds. *Clinical Research Informatics Health Informatics*. Springer; 2012:175-201.
4. Heyer NJ, Derzon JH, Wings L, et al. Effectiveness of practices to reduce blood sample hemolysis in EDs: a laboratory medicine best practices systematic review and meta-analysis. *Clin Biochem.* 2012;45:1012-1032.