Research Fundamentals: Data Collection, Data Analysis, and Ethics

Bismah Jameel, BSc Student [1], Umair Majid, MSc [2,3]*

[1] School of Interdisciplinary Science, McMaster University, Hamilton, Ontario, Canada
[2] Editorial and Advisory Board Member, URNCST Journal, Toronto, Ontario, Canada
[3] Instructor, McMaster University, Hamilton, ON

*Corresponding Author: majidua@mcmaster.ca

Abstract
This is the third and final article of a three-part series that follows up the discussion on the fundamentals of writing research protocols for quantitative clinical research studies. In this editorial, the authors discuss key elements of data collection, data analysis, and the ethical considerations and implications that come with clinical research. This editorial is the concluding segment on providing guidelines for undergraduate researchers interested in publishing their protocol in the Undergraduate Research in Natural and Clinical Sciences and Technology (URNCST) Journal.

Keywords: data collection; questionnaires; analysis; descriptive statistics; clinical ethics; informed consent; research ethics board; undergraduate research; research protocol; clinical research

Introduction
The final segment of the three-part series focuses on data collection, data analysis, and the ethics of clinical research studies. The authors first describe data collection in clinical research and then explore how undergraduate investigators may collect valid, useful data through questionnaires and how to analyze quantitative data appropriately. The authors conclude this paper by identifying the ethical considerations in conducting clinical research.

Data Collection
Data collection is the process of gathering information on variables of interest from a sample of research participants. There are two types of data collection:

1. Primary data collection refers to data that is collected from research participants directly by the investigators of a study and the data is used for that study.
2. Secondary data collection refers to data that is collected by investigators from research papers that are already published online. Secondary data is used by these investigators in a secondary research study (e.g., review of primary research).

Questionnaires
Questionnaires (also known as surveys) are a systematic method of data collection using structured or open-ended questions [1]. Structured questionnaires elicit close-ended responses (i.e., yes or no) whereas open-ended questionnaires prompt the research participant to respond freely. Questionnaires are most commonly used in observational studies to collect data and insight about the relationship between exposures and outcomes [2]. For example, the Pain Disability Index (PDI) is a commonly utilized questionnaire that measures patients’ self-perception of how their pain interferes with their functioning in seven domains of life (family/home, recreation, social activity, occupation, sexual behaviour, self-care, and life support) [1].

Ideally, clinical research studies should employ objective measures such as body mass index, results of magnetic resonance imaging scans, and serum cholesterol levels. However, questionnaires could be employed as substitutes to objective measures when 1) objective measures that measure the variable of interest are unavailable or 2) objective measures are not feasible in the research study or dangerous to the research participants. For example, measuring how pain interferes in the domains of life is an individual experience that may be measured accurately through a questionnaire such as the PDI [3]. An objective measure for pain may not be available because it is a subjective experience; only patients can determine how much pain they feel. In addition to these characteristics, questionnaires have several advantages and disadvantages as shown in Table 1.

Since questionnaires are used as methods for measuring more subjective variables [2], there is a considerable degree of uncertainty in their reliability and validity. Table 2 describes four components of effective questionnaires [1].

There are many challenges associated with the use and administration of questionnaires in clinical research studies...
such as biases in their design and administration [4]. Table 3 provides a description of some of the many biases commonly encountered in research involving questionnaires.

### Table 1: Advantages and Disadvantages to Employing Questionnaires in Clinical Research

<table>
<thead>
<tr>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Reach a large number of participants</td>
<td>• Provide limited insight into the variables because questionnaires are structured</td>
</tr>
<tr>
<td>• Reach a wider geographical area</td>
<td>• Participants may not provide the correct response to questionnaires due to their lack of understanding or due to response bias</td>
</tr>
<tr>
<td>• Administered in a shorter amount of time</td>
<td>• Questionnaires may not be a valid or reliable measure of the variables</td>
</tr>
<tr>
<td>• Can provide quantifiable answers</td>
<td>• Loss-to-follow-up or low response rate</td>
</tr>
<tr>
<td>• Relatively easy to analyze and interpret</td>
<td></td>
</tr>
<tr>
<td>• Less expensive compared to other types of data collection (e.g., interviews)</td>
<td></td>
</tr>
<tr>
<td>• Reduces interview bias</td>
<td></td>
</tr>
</tbody>
</table>

### Table 2: Four Components of Effective Questionnaires

<table>
<thead>
<tr>
<th>Component</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Validity</td>
<td>The questionnaire measures what it intends to measure</td>
</tr>
<tr>
<td>Reliability</td>
<td>Findings from the questionnaire are consistent</td>
</tr>
<tr>
<td>Feasibility</td>
<td>The questionnaire is appropriate to the context of the study and in reaching the intended population</td>
</tr>
<tr>
<td>Acceptability</td>
<td>The population of interest would be willing to complete the questionnaire</td>
</tr>
</tbody>
</table>

### Table 3: Some Biases Commonly Encountered in Research that Uses Questionnaires

<table>
<thead>
<tr>
<th>Type of Bias</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Social Desirability</td>
<td>An attempt (consciously or unconsciously) to appear healthier, more adherent to treatment, and more “normal” (relative to social norms) and knowledgeable</td>
</tr>
<tr>
<td>Fear of Disclosure</td>
<td>Avoid disclosing pertinent information in fear of stigma or negative consequences to personal or professional life domains</td>
</tr>
<tr>
<td>Attrition</td>
<td>Loss-of-follow-up - as time passes, fewer participants will complete and return the questionnaire, affecting sample size</td>
</tr>
<tr>
<td>End Aversion</td>
<td>Participants tend to avoid the extreme answers of questions with Likert scales</td>
</tr>
<tr>
<td>Recall Bias</td>
<td>Participants tend to remember events that are more recent or more emotional than events that are not</td>
</tr>
</tbody>
</table>

Designing a data collection plan in the beginning stages of a research study is an important step for investigators. A data collection plan provides a description of the questionnaires, how they will be administered in the population of interest, how the data will be managed, and any other resources that will be utilized for the data collection process. When designing a data collection plan, investigators should consider their answers to the following questions:

1. What are my research objectives and question?
2. Have I obtained ethics approval for data collection from my research ethics board? What are the variables of interest?
3. How are the variables of interest conceptualized and operationalized [5]?
4. Is there an existing questionnaire in the literature that measures these variables?
5. Is this questionnaire valid, reliable, feasible, and acceptable?
6. How should the questions be presented to the sample population (appropriateness)?
7. Will the questionnaire answer my research question (specificity)?
8. Will the questionnaire detect differences in the variables of interest (sensitivity)?
9. How frequently do I need to administer my questionnaire?
10. Do I need any internal or external support to administer my questionnaire?
11. Is there anything else I am missing in my data collection plan?

### Data Analysis Definitions and Concepts

Research protocols should contain a section that describes the statistical procedures the investigators intend to conduct to answer their research question. Although the type and method of statistical analysis may differ between
research protocols, the section on data analysis generally includes a discussion on the descriptive features of the data and the primary statistical analysis that will clarify associations and relationships between variables [6]. Other information to include are assumptions about the shape and distribution of the data; the process of verifying the assumptions underlying statistical testing; the chosen statistical software packages; the purposes and approaches of primary, secondary, and ad hoc analyses; and the parameters of outcome measures [6].

**Definitions and Concepts**

Generally, there are two categories of statistics. **Descriptive statistics** organizes and summarizes raw data to provide an overview of the general features of a data set. This category of statistics provides a visual presentation of data through histograms, boxplots, and scatterplots. On the other hand, **analytic statistics** performs computations on a data set to discern the statistical significance between variables. This category of statistics seeks to determine the statistical relationships and associations between one or more variables of interest [7].

Measures of central tendency, standard deviation, and variance are some of the variables important in descriptive statistics. **Table 4** describes the measures of central tendency. The **variance** and **standard deviation** measure the spread and dispersion of a data set relative to its mean. These two statistical values are related through the square root calculation. The standard deviation is a more commonly used because it is in the same units as the mean, which eases comparison between the two values. Variance, on the other hand, is advantageous for developing math theory and manipulating statistical formulae, which is beyond the scope of this article [7].

**Hypothesis Testing**

There are many reasons to conduct research inquiry. These reasons include understanding and clarifying natural or theoretical phenomena, building the scientific foundation of understanding, engaging in knowledge translation, satisfying personal or collective curiosity, answering pressing questions, and enhancing patient care or how to perform research.

In statistical hypothesis testing, the objective is to produce findings that can be generalized to other populations and contexts. Medical health professionals, for example, continuously survey research studies conducted in other jurisdictions, health care systems, and communities so that they may apply the findings in their own context. In clinical research, findings that are not transferable beyond the circumstances of the study may only offer minimal value to evidence-based practice.

**Statistical inferencing** is the process of drawing conclusions about an entire population based on a sample from it [8]. For example, to investigate whether or not there is a statistical relationship between postpartum depression and a second pregnancy, researchers may recruit a sample of pregnant women that supposedly represent the entire population. In this case, researchers do not know the mean of the prevalence of postpartum depression of the entire population of pregnant women but they can calculate the mean of the sample in their research study. Statistical inferencing uses the mean of the prevalence of postpartum depression in the sample of pregnant women to make conclusions about the mean of the entire population and its relationship to a second pregnancy.

**Statistical hypothesis testing** determines whether or not the difference between the sample and population mean is too large to be attributed to random variation [9]. Hypothesis testing relies on the null and alternate hypotheses, type I and II errors, significance and power level, and the p-value, which were discussed in the second editorial of this three-part series on writing research protocols [8]. Hypothesis testing determines whether or not the data is in favour of the null hypothesis (i.e., no significant difference between population and sample means) or alternate hypothesis (i.e., there is a significant difference between population and sample means). If there is no significant difference, then the sample is characteristically identical to the population on the variables of interest.

On the other hand, if there is a significant difference between the sample and population means, then the two populations are distinct due to one or more variables of interest. For example, in an investigation of a therapeutic intervention that increases the levels of belongingness in victims of intimate partner violence (IPV), if a sample of IPV victims experience a statistically significant increase in belongingness by the end of the intervention, then the sample mean of belongingness is different from the population mean, and therefore, the sample and population are significantly different due to the therapeutic intervention [10].

<table>
<thead>
<tr>
<th>Measure</th>
<th>Definition</th>
<th>Example</th>
</tr>
</thead>
</table>
| **Mean**         | The average of a dataset                        | Dataset: 1, 1, 5, 6, 8, 9  
Mean: 5           |
| **Mode**         | The observation that occurs most frequently in a dataset | Dataset: 1, 1, 5, 6, 8, 9  
Mode: 1            |
| **Median**       | The 50th percentile of a dataset               | Dataset: 1, 1, 5, 6, 8, 9  
Median: 5.5    |
This is the reasoning used in clinical research to make conclusions and interpretations based on statistical computations of significance.

Comparison of Two Means

There are two basic methods to perform a statistical comparison of two means. The method used by investigators depend on the type of sample in the research study. A paired significance test (comparison of two means of one population) computes the statistical difference between two or more observations of the same population but at different times [7]. For example, a study investigating an intervention that aims to reduce systolic blood pressure in older individuals would compute the statistical difference between the sample means of systolic blood pressure before and after the intervention.

The second type of statistical comparison of two means is the independent significance test (comparison of two means of different populations). In this test, the mean is computed for one or more samples in the research study and compared to obtain a statistically significant difference [7]. These samples have different characteristics on variables of interest in the research study. For example, in a study comparing the effect of the blood pressure-reducing intervention on a group of older individuals, the statistical comparison for an independent significant test would be between two different samples of older individuals where one would receive the intervention (experimental group) and another group would not receive the intervention or receive a placebo (control group). This significance test has the advantage of employing a control group, which may reduce the risk of samples being significantly distinct due to random variations in the sampling or recruitment process. Moreover, using a control group that is similar to the experimental group in all characteristics except for the variables of interest is generally perceived to have superior methodological properties because this process supports investigators to make confident conclusions on which variables of interest are statistically meaningful [9].

Linear Regression

Linear regression is the simplest form of statistical modeling used to provide a meaningful summary description about the relationship between a dependent (outcome) and an independent (predictor) variable [11]. Complex statistical modeling such as logistic, multivariable, stepwise (backward and forward) regression compare dependent variables to multiple independent variables. However, these forms of modeling are beyond the scope of this article. As the name implies, linear regression is often used to establish a dose-dependent response or evaluate whether the strength of an intervention can impact the outcome on the dependent variable.

Association, relationship, link, and other terms are often used interchangeably in research studies. Statistically, an association between two variables means that if there is a change in one variable, then there is a predictable change in the other variable. Correlation, on the other hand, is a statistical term that is familiar to many undergraduate investigators that describes a linear association between two variables [11]. If two variables do not have a linear relationship (e.g., inverted U relationship), however, then their association is not considered a correlation. Correlation is represented through the Pearson’s R or Spearman Rank Correlation Coefficients. The value of the correlation coefficient ranges from -1 to 1, which represent a negative to positive association between two variables, respectively [9].

In linear regression, the basic assumption is that each unit increase in an independent variable corresponds to a fixed and predictable increase (or decrease) in a dependent variable. The beta-coefficient is the value obtained from regression modeling that represents the magnitude and direction of the increase or decrease in a dependent variable. In statistical tests of significance using regression, the null hypothesis is that the beta-coefficient will be zero, meaning that there is no change in the dependent variable when there is a change in the independent variable, and therefore, there is no association between these variables. On the other hand, the alternate hypothesis in regression statistics is that the beta-coefficient will not be zero.

Ethics in Clinical Research

Introduction to Ethics

“Ethics is knowing the difference between what you have a right to do and what is right to do” – Potter Stewart

One of the most important aspects of conducting a research study involves a consideration of the moral and ethical implications of the research protocol, methodology, and the expected study findings. Throughout the life cycle of a research study, investigators should respect research participants’ freedom of autonomy and reflect upon the potential harms and risks posed to them [12]. Moreover, investigators need to ensure that the benefits to participation in the research study outweigh the potential risks to physical, psychological, and social health or status [12].

Informed Consent

Informed consent is the process of communicating essential information about a research study and medical or therapeutic intervention that may be important to a potential research participant to know about the risks and benefits to participation and make an informed, rational decision of whether or not participation in a research study is in their best interests [13]. In a clinical research study, an adequate informed consent procedure includes information about the nature of the research project, specific procedures of the research study, and the potential risks and benefits to participants [14]. Table 5 describes the components of informed consent in more detail.
The design and administration of informed consent forms (ICFs) is required to receive ethics approval to conduct a research study [14]. ICFs describe the nature, procedures, and risks and benefits to research participation. They are written in plain language that carefully consider the demographic characteristics of the potential research participants. As a rule of thumb, ICFs should be written in grade 6 or 8 level English. Moreover, when writing an ICF, investigators should make the following considerations:

- Avoid technical jargon and complicated sentences
- Increase comprehension of the ICF by spending more time speaking with the potential research participants, using question and answer format to frame the ICF and informed consent discussion, and revisit the ICF details over multiple interactions with participants
- Assess participants’ comprehension by asking them to summarize the key aspects of the research study and the information contained in the ICF

Use ICFs from past or current research studies as a template for designing ICFs for future research studies

Even after receiving informed consent for participation, investigators should revisit and remind participants about the boundaries of informed consent. Participants should always be made aware that they may withdraw consent at any time.

Undue influence refers to an external pressure (e.g., coercion or financial incentive) that results in consent to participate in a research study that is not entirely voluntary. For example, patients who are recruited by their physicians may feel that choosing to not participate in a research study may adversely affect their future health care. In such cases, someone other than the physician or investigator should recruit patients and reassure the patients that there will be no consequences to their health care, status, or treatment if they decide to not participate in the research study.

### Table 5: Specific Components of Informed Consent

<table>
<thead>
<tr>
<th>Information</th>
<th>Components</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nature of Research Project</td>
<td>(1) Explicitly state that a research study is being conducted</td>
</tr>
<tr>
<td></td>
<td>(2) Describe the purpose(s) of the research study</td>
</tr>
<tr>
<td></td>
<td>(3) Explain who is being included as a research participant</td>
</tr>
<tr>
<td>Procedures of Study</td>
<td>(1) Participants need to know what they will be asked to do (duration and frequency) in the research study</td>
</tr>
<tr>
<td></td>
<td>(2) Procedures that are not the standard of care should be identified</td>
</tr>
<tr>
<td></td>
<td>(3) Blinding and randomization should be explained in lay terms</td>
</tr>
<tr>
<td></td>
<td>(4) Topics in the interviews/questionnaires should be explained in lay terms</td>
</tr>
<tr>
<td>Risks and Potential Benefits</td>
<td>(1) Medical, psychosocial, and economic risks and benefits should be described in simple terms</td>
</tr>
<tr>
<td></td>
<td>(2) Alternatives to participation should be identified</td>
</tr>
<tr>
<td></td>
<td>(3) Explicitly state that it is not known whether the intervention under investigation is more effective than the standard of care (i.e. equipoise)</td>
</tr>
</tbody>
</table>

### Research Ethics Approval by a REB/IRB

Generally, a research protocol must be approved by a Research Ethics Board (REB) (also known as an Institutional Review Board or Research Ethics Committee) before the research study is conducted [15]. REBs are governing bodies with the authority from the federal government to ensure that research investigations fulfill the guidelines of ethical and legal practice [16]. An REB comprises of a team of researchers, professionals, and community members with knowledge and experience in research, ethics, and law.

When assessing the ethics of a clinical research study, an REB looks for certain things such as the balance between risks and benefits, the fairness of recruitment participation, the informed consent process, and measures to respect participant privacy and confidentiality [14]. The aim of REBs is to provide investigators with encouragement and suggestions on how to address ethical concerns related to participation in their research study. The decisions made by an REB are not final and the research team may discuss with the REB if they disagree with any suggestions or recommendations.

The Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans (TCSP-2) is a guideline that informs ethical research practice in Canada. This document contains useful definitions and concepts that may support investigators to develop research protocols in a manner that is ethically appropriate and cognizant of the population of interest.

Some research may include minimal risk to participants, which TCPS-2 defines as “if potential subjects can reasonably be expected to regard the probability and magnitude of possible harms implied by participation in the re-

DOI Link: https://doi.org/10.26685/urncst.39
search to be no greater than those encountered by the subject in those aspects of his or her everyday life that relate to the research” [17]. Investigators conducting research studies with minimal risk must make it explicit in the ethics application. Their research study may undergo an expedited review where one or two members of the REB reviews the research study as opposed to the full team [18]. Some examples of clinical investigations that may undergo an expedited review are collecting saliva, minor changes in previously approved protocols, and renewing research studies for a follow-up questionnaire. In some instances, a research study may be exempted from review if the risk is very minimal and involves questionnaires that will have no impact on the research participant’s mental health, social or legal status.

Investigators may use the following as a guide for writing an ethics application to a REB:

1. **Research the REB**: Most institutions have their own REB and unique ethics application form. Investigators should identify the requirements of their REB, which usually includes certain sections or a complete research protocol, a letter explaining the purpose of the research study, and supplementary documentation such as ICFs, interview guides, data collection sheets, and templates of questionnaires.

2. **Draft the Research Protocol**: Sometimes, REBs may not require a full research protocol. However, they will require specific information about the methodology and process of the research study such as information about potential research participants, sampling and recruitment methods, interventions and data collection procedures, safety measures to prevent undue harm, potential risks and benefits to research participants, informed consent process, and steps to maintain privacy and confidentiality of participants. The measures adopted to minimize potential risks to research participants should be clearly outlined. Moreover, the ethics application should also describe who will have access to confidential data, and where and for how long the information will be stored. For a more complete description on writing these sections of a research protocol, refer to the first two editorials in the series on Research Fundamentals [5] [8].

3. **Complete a REB Application**: Each REB will have its own process for obtaining ethics approval. Some will require the completion of an online application, which may vary in length and require additional documentation. The online application prompts investigators to describe the purpose, procedures, methodology, participants, and ethical considerations.

4. **Review the Protocol and Ethics Application**: Typically, undergraduate investigators working on a clinical research study should collaborate with a principal investigator (PI) affiliated with a health care or education institution. The undergraduate investigator should ensure that the PI has reviewed the protocol and ethics application thoroughly and that it contains all the necessary information for obtaining ethics approval.

**Conclusion**

This guest editorial series concludes the discussion on writing research protocols. The first article in this three-part series discussed framing the research question, outcomes, and background. The second editorial elaborates on the study design, population, and sample size. This article is the third and final editorial of the series and discussed the foundations of data collection, data analysis using statistics, and the ethics of clinical research.

**List of Abbreviations Used**

1. ICFs: Informed Consent Forms
2. IPV: Intimate Partner Violence
3. PDI: Pain Disability Index
4. REB/IRB: Research Ethics Board/Institutional Review Board

**Conflicts of Interest**

The authors declare that they have no conflicts of interest.

**Authors’ Contributions**

BJ and UM wrote the entire publication based on their experiences and knowledge in addition to the methodological literature on writing effective manuscripts, proposals and protocols.

**Acknowledgements**

None.

**Funding**

This study was not funded.

**References**


Do you research in earnest? Submit your next undergraduate research article to the URNCST Journal!
| Open Access | Peer-Reviewed | Rapid Turnaround Time | International |
| Broad and Multidisciplinary | Indexed | Innovative | Social Media Promoted |
Pre-submission inquiries? Send us an email at info@urncst.com | Facebook, Twitter and LinkedIn: @URNCST
Submit YOUR manuscript today at https://www.urncst.com!