African Journal of Emergency Medicine 10 (2020) S135–S139

Contents lists available at ScienceDirect

African Journal of Emergency Medicine

journal homepage: www.elsevier.com/locate/afjem

Research primer

Acquiring data in medical research: A research primer for low- and middle-income countries

Vicken Totten[a,*], Erin L. Simon[b], Mohammad Jalili[c], Hendry R. Sawe[d]

[a] Kaweah Delta Health Care District (KDHCD), KDHCD Department of Emergency Medicine, Visalia, CA, USA
[b] Cleveland Clinic Akron General, Department of Emergency Medicine, Akron, OH, USA
[c] Department of Emergency Medicine, Tehran University of Medical Sciences, Tehran, Iran
[d] Muhimbili University of Health and Allied Sciences, Dar es Salaam, Tanzania

ARTICLE INFO

Keywords:
Data collection
Sampling
Variables
Data storage

ABSTRACT

Without data, there is no new knowledge generated. There may be interesting speculation, new paradigms or theories, but without data gathered from the universe, as representative of the truth in the universe as possible, there will be no new knowledge. Therefore, it is important to become excellent at collecting, collating and correctly interpreting data. Pre-existing and new data sources are discussed; variables are discussed, and sampling methods are covered. The importance of a detailed protocol and research manual are emphasized. Data collectors and data collection forms, both electronic and paper-based are discussed. Ensuring subject privacy while also ensuring appropriate data retention must be balanced.

African relevance

- To get good quality information you first need good quality data
- Data collection systematically and reproducibly gathers and measures variables to answer research questions.
- Good data is a result of a well thought out study protocol

The International Federation for Emergency Medicine global health research primer

This paper forms part 9 of a series of 'how to' papers, commissioned by the International Federation for Emergency Medicine. It describes data sources, variables, sampling methods, data collection and the value of a clear data protocol. We have also included additional tips and pitfalls that are relevant to emergency medicine researchers.

Background

Data collection is the process of systematically and reproducibly gathering and measuring variables in order to answer research questions, test hypotheses, or evaluate outcomes.

Data is not information. To get good quality information you first need good quality data, then you must curate, analyse and interpret it. Data is comprised of variables. Data collection begins with determining which variables are required, followed by the selection of a sample from a certain population. After that, a *data collection tool* is used to collect the variables from the selected sample, which is then converted into a data spreadsheet or database. The analysis is done on the database.

Sometimes you gather data yourself. Sometimes you analyse data others collected for different purposes. Ideally, you collect a universal sample, that is, 100%. In real life, you get a limited sample. Preferably, it will be a truly random sample with enough power to answer your question. Unfortunately, you may have to settle for consecutive or convenience sampling. Ideally, your data collectors would be blinded to the outcome of interest, to prevent bias. However, real life is full of biases. Imperfect data may be better than no data; you can often get useful information from imperfect data. Remember the enemy of good is perfect.

Why is good data important?

Acquiring data is the most important step in a research study. The best design with bad data is useless. Bad design produces bad data. The most sophisticated analysis cannot be performed without data; analysing bad data produces erroneous results. Analysis can never be better than the quality of the data on which it was run. Good data has integrity. Data integrity is paramount to learning "Truth in the Universe". Good data is as complete and as clean, as you can reasonably make it. Clean data 'has integrity' when the variables access as much relevant information as possible, and in the same way for each subject.

* Corresponding author.
E-mail addresses: Vicken.totten@gmail.com, Vicken.Totten@case.edu (V. Totten).

https://doi.org/10.1016/j.afjem.2020.09.009
Received 21 February 2020; Received in revised form 8 August 2020; Accepted 5 September 2020
Available online 01 October 2020

Some information is very hard to get. You may have to use proxy variables for what you really want to know. A proxy variable is a variable that is not in itself directly relevant, but that serves in place of an unobservable or immeasurable variable. In order for a variable to be a good proxy, it must have a close correlation, not necessarily linear, with the variable of interest. One example for the variable of a specific illness might be a medication list.

Consequences of bad data include an inability to answer the research question; inability to replicate or validate the study; distorted findings and wasted resources; compromised knowledge and even harm to subjects.

Ensure data quality

Good data is a result of a well-thought-out study protocol, which is the written plan for the study. Good planning is the most cost-effective way to ensure data integrity. Good planning is documented by a thorough and detailed protocol, with a comprehensive procedures manual. Poorly written manuals risk incomplete or inconsistent collection of data, in other words, 'bad data'. The manual should include rigorous, step-by-step instructions on how to administer tests or collect the data. It should cover the 'who' (the subject and the researcher); the 'when' (the timing), the 'how' (methods), and the 'what' (a complete listing of variables to be collected). There should also be an identified mechanism to document any changes in procedures that may evolve over the course of the investigation. The study design should be reproducible: so that the protocol can be followed by any other researcher. All data needs to be gathered in the same way. Test (trial-run) your manual before you start your study. If data is collected by several people, make sure there is a sufficient degree of inter-rater reliability.

To get good data, your sample needs to be representative of the population. For others to apply your results, you need to characterize your population, so others can decide if your conclusions are relevant to their population (see Sampling section, below).

Data integrity demands you supervise your study, making sure it is complete and accurate. You may wish to do interim analyses. Keep copies! Keep both the raw data and the data sheets, for the length of time required by law or by Good Research Practice in your country. This will protect you from accusations of falsification of data.

In real life, you may have to deal with any number of sampling and data collection biases. Some of these biases can be measured statistically. Regardless, all the limitations you can think of should be written in your *limitations* section. The best design you can practically use gives you the best data you can reasonably get. Remember, "you cannot fix with statistics what you fouled up by design."

Before you acquire your first datum, consider: Do you have a developed protocol and a research manual? Have you sought Ethics Board approval? Do you have an informed consent? Do you have a plan to protect the subject's confidentiality? Do you have a plan for data analysis? Where will you safely store and protect the data? If you have collaborators, have you established, in writing, who owns the data, and who has the right to analyse and publish it?

Types of data: qualitative vs. quantitative data

Numerical data is generally called quantitative; if in words or sentences, it is qualitative. Medical research historically has focused on quantitative methods. Generally, quantitative research is cheaper, easier to gather and easier to analyse. For purposes of this chapter, we will focus on quantitative research.

Qualitative research is about words, sentences, sounds, feeling, emotions, colours and other elements that are non-quantifiable. It requires human intellect to extract themes from the sentences, evaluate the fit of the data to the themes, and to draw the implications of the themes. Primary sources for qualitative data include open ended surveys, interviews, and public meetings. Qualitative research is more common in politics and the social sciences, and will not be further discussed here, except to refer you to other sources.

Quantitative research can include questionnaires with closed-ended questions (open ended questions belong in qualitative research). The data is transformed into numbers and will be analysed with parametric and non-parametric statistical tests. In general, you will derive a mean, mode and median; you will calculate probabilities, make correlation and regressions in order to draw conclusions.

Sources of data: primary vs secondary data

To answer a research question, there are many potential sources of data. Two main categories are primary data and secondary data. Primary data is newly collected data; it can be gathered directly from people's responses (surveys), or from their biometrics (blood pressure, weight, blood tests, etc.). It is still considered primary data if you gather data that was collected for other (medical) purposes by extracting the data from medical records. Medical records can be a rich source of data, but data extraction by hand takes a lot of time.

Secondary data already exists; it has already been published or complied. There are extant local, regional, national and international databases such as Trauma Registries, Disease-specific Registries, Public Health Data, government statistics, and World Health Organization data. Locally, your hospital or clinic may already keep statistics on any number of topics. Combining information from disparate databases may sometimes yield interesting results. For example, in the US, the Centers for Disease Control and Prevention keeps databases of reportable diseases, accidents, causes of death and much more. The US Geographic Survey reports the average elevation of American cities. Combining the two databases revealed that, even when gun ownership, drug and alcohol use were statistically controlled for, there was a linear correlation between altitude and suicide rates [2]. Reno et al., reviewed the existing medical literature (also secondary data), and confirmed the correlation and concluded that the mechanisms have yet to be elucidated [3].

Sampling

Collecting good data is often the hardest part of research. Ideally, you would want to collect 100% of the data (universal sampling to reflect target population). One example would be 'all elderly persons with gout'. In real life, you have access to only a subset of the target population (the accessible population). Further, in your study you will be limited to a subset of the accessible population (the study population). Again, in the ideal world, that limited sample would be truly random, and have enough power to answer your question. You can find free random number generators online. In real life, you may have to settle for consecutive or convenience sampling. Of the two, consecutive sampling has less bias. Sometimes it is important to balance your groups. You may have 2 or 3 *treatments* (or interventions) and want to have an equal number of each kind. So, you create *blocks* — of a few times the number of treatments. You randomized within the block. Each time a block is filled, you are assured that you have the right balance of subjects. Blocks are often in groups of six, eight or 12. This is called *balanced allocation*.

If you must get only a convenience sample – for example because you only have a single data gatherer and can get data only when that person is available – you should, at a minimum, try to get some simple demographics from times when the data gatherer is not available, to see if subjects at that other time are systematically different. For example, if you are looking at injuries, people who are injured when drinking on a Friday night might be systematically different from people who are injured on their way to work on a Monday morning. If you can only collect injury data in the morning, your results will be biased.

Variables

Variables are the bits of data you collect. They change from subject to subject and describe the subject numerically. Age (or year of birth); gender; ethnic group or tribe; and geographic location are commonly called *simple demographic variables* and should be collected and reported for most populations.

Continuous variables are quantified on a continuous scale, such as body weight. Discrete variables use a scale whose units are limited to integers (such as the number of cigarettes smoked per day). Discrete variables have a number of possible values and can resemble continuous variables in statistical analysis and be equivalent for the purpose of designing measurements. A good general rule is to prefer continuous variables because the information they contain provides additional information and improves statistical efficiency (more study power and smaller sample size).

Categorical variables are those not suitable for quantification. They are often measured by classifying them into categories. If there are two possible values (dead or alive), they are dichotomous. If there are more than two categories, they can be classified according to the type of information they provide (polytomous).

Research variables are either *predictor* (independent) or *outcome* (dependent) variables. The predictor variables might include such things as "Diabetes, Yes/No", "Age over 65 — Yes/No", and "diagnosis of hypertension" (again, Yes/No). The respective outcome might be "lower limb amputation" or "death within 10 years". Your question might have been, "How much additional risk of amputation does a diagnosis of hypertension add in a person with diabetes?"

Before analysis, variables are coded into numbers and entered into a database. Your Research Manual should describe how to code all the data. When the variables are binary, (male/female; alive/dead) coding them into "0" and "1" makes analysing the data much easier ("1" versus "2" makes it harder). The easiest variables for computers to analyse are binary. In other words, "0" or "1". Such variables are Yes/No; True/False; Male/Female; 65 or over / under 65, etc. The next easiest are ordinal integers: 1, 2, 3, etc. You might create ordinal numbers from categories (0–9; 10–19; 20–29 years of age, etc.), but in order to be ordinal, they require an obvious sequence. Categorical variables do not have an intrinsic order. "Green" "Brown" and "Orange" are non-ordinal, categorical variables. It is possible to transform categorical variables into binary variables, by making columns where only one of the answers is marked with a "1" (if that variable is present) and all the others are marked "0". The form of the variables and their distribution will determine the type of statistical analysis possible. Data which must be transformed or cleaned is more prone to error in the cleaning or transformation process.

There are alternative ways to get similar information. For example, if you wanted to know the HIV status of each of your subjects, you could either test each one, or you could ask them. The tests cost more, however; they are less likely to give biased results. How you gather each variable will depend on your resources and will inform the limitations of your study.

Precision of a variable is the degree to which it is reproducible with nearly the same value each time it is measured. Precision has a very important influence on the power of a study. The more precise a measurement, the greater the statistical power of a given sample size to estimate mean values and test your hypotheses. In order to minimize random error in your data, and increase the precision of measurements, you should standardize your measurement methods; train your observers; refine any instruments you may use (such as calibrating instruments); automate instruments when possible (automated blood pressure cuff instead of manual); and repeat your measurements.

Accuracy of the variable is the degree to which it actually represents what it is intended to (Truth in the Universe). This influences the validity of the study. Accuracy is impacted by systemic error (bias). The greater the error, the less accurate the variable. Three common biases are: observer bias (how the measurement is reported); instrument bias (faulty function of an instrument); and subject bias (bad reporting or recall of the measurement by the study subject).

Validity is the degree to which a measurement represents the phenomenon of interest. When validating an abstract concept, search the literature or consult with experts so you can find an already validated data collection instrument (such as a questionnaire). This allows your results to be comparable to prior studies in the same area and strengthens your study methods.

Research manual

Simple research with limited resources does not need a research manual, just a protocol. Nor is there much need if the primary investigator is the only data gatherer and analyser. However, if several persons gather data, it is important that the data be gathered the same way each time.

Prevention is the most cost-effective activity that will ensure the integrity of data collection. A detailed and comprehensive research manual will standardize data collection. Poorly written manuals are vague and ambiguous.

The research manual is based off your protocol. The manual should spell out every step of the data collection process. It should include the name of each variable and specific details about how each variable should be collected. Contingents should be written. For example: "If the patient does not have a left arm, the blood pressure may be taken on the right arm. If the patient has no arms, leg blood pressures may be recorded, but put an '*' beside the reading." The manual should also include every step of the coding process. The coding manual should describe the name of each variable, and how it should be coded. Both the coder and the statistician will want to refer to that section. The coding section should describe how each variable will be entered into the database. Test the manual to make sure everyone understands it the same way.

Think about various ways a plan can go wrong. Write them down, with preferred solutions. There will always be unexpected changes. They should be added into the manual on a continuing basis. An ongoing section where questions, problems and their solutions are all recorded will increase the integrity of your research.

Data collection methods

Before you start data collection, you need to ask yourself what data you are going to collect and how you are going to collect them. Which data, and the amount of data to be collected needs to be defined clearly. Different people (including several data collectors) should have a similar understanding of each variable and how it is measured. Otherwise, the data cannot be relied on. Furthermore, the decision to collect a piece of data needs to be justified. The amount of data collected for the study should be sufficient. A common mistake is to collect too much data without actually knowing what will be done with it. Researchers should identify essential data elements and eliminate those that may seem *interesting* but are not central to the study hypothesis. Collection of the latter type of data places an unnecessary burden on both the study participants and data collectors.

Different data collection approaches which are commonly used in the conduct of clinical research include questionnaire surveys, patient self-reported data, proxy/informant information, hospital and ambulatory medical records, as well as the collection and analysis of biologic samples. Each of these methods has its own advantages and disadvantages.

Surveys are conducted through administration of standardized or *home-grown* questionnaires, where participants are asked to respond to a set of questions as yes/no, or perhaps on a Likert type scale. Sometimes open-ended responses are elicited.

Medical records can be important sources of high-quality data and

may be used either as the only source of data, or as a complement to information collected through other instruments. Unfortunately, due to the non-standardized nature of data collection, information contained in the medical records may be conflicting or of questionable accuracy. Moreover, the extent of documentation by different providers can vary significantly. These issues can make the construction or use of key study variables very difficult.

Collection of biological materials, as well as various imaging modalities, from the study participants are increasingly being used in clinical research. They need to be performed under standardized conditions, and ethical implications should be considered.

Data collection tool

You may need to collect information on paper. If you do, it is useful to have the actual code which should be entered into the computerized database written on the forms themselves (as well as in the manual). If you have access to an electronic database such as REDcap [a web-based application developed by Vanderbilt University to capture data for clinical research and create databases and projects [4], you can enter the data directly as you get them (*male*; *female*) and the database will automatically convert the data into code. This reduces transcribing errors. Another common electronic database is Excel, which can also be used to manipulate the data. In spite of the advantages of recording data electronically, such as directly into REDcap or Excel, there are advantages to collecting and keeping the original data on paper. Paper data collection forms can be saved for audit or quality control. Furthermore, paper records cannot be remotely hacked. Moreover, if the anonymous electronic database is compromised or corrupted, you can re-create your database.

Data collectors

Good data collectors are worth gold. If they are thorough and ethical, you will get great data. If not, your data may be unusable. Make sure they understand research ethics, the need for protection of human subjects, and the privacy of data. Ideally, your data collectors would be blinded to the outcome of interest, to prevent bias. It is ok to *blind* data collectors to the research question, but they need to understand that collecting every variable the same way for each subject is essential to data integrity.

Data gatherers should be trained in advance of collecting any data. They need to understand informed consent and have the time to explain a study to the satisfaction of the subjects. The importance of conducting a *dry run* in an attempt to anticipate and address issues that can arise during data collection cannot be over-stated. It would even be worthwhile to pilot the research manual, to learn if everyone understands it the same way.

Data storage

Data collection, done right, protects the confidentiality of the subject as well as the data. Data must also be properly stored safely and securely. It is reasonable to back up your data in a different, secure, location. You do not want to go to all the trouble of creating a protocol, collecting your data, only to lose it, or have no way to analyse it!

There are many reasons to keep your data safe and secure. Obviously, you do not want to lose your data. You may wish to use the data again. For example, you may wish to combine it with other data for a different study. An additional reason is that you do not want your subjects to risk a 'loss of privacy'. Still another reason is that institutions and governments may require you to store data for a specified number of years. Know how long you must keep your data. Keep it in a locked cabinet in a secure room, or behind an institutional firewall.

Furthermore, if you keep a *cipher*, that is, a connector between a subject and their study number, keep that cipher separate from the research data. That way, even if someone learns that subject 302 has an embarrassing condition, they will not know *who* subject 302 really is.

These days, almost everyone has access to computers and programs, locally or 'in the cloud'. For statistical analysis, you will need to have your data in electronic form. If you started with paper, consider double entry (two data extractors for each record, then compare the two) for greater accuracy.

Tips on this topic and pitfalls to avoid

Hazard: no research manual

- No identified mechanism to document changes in procedures that may evolve over the course of the investigation.
- Vague description of data collection instruments to be used in lieu of rigorous step-by-step instructions on administering tests
- Only a partial listing of variables to be collected
- Forgetting to put instructions on the data collection sheet about how to code the data when transferring to an electronic medium.

Hazard: no assistant training

- Failure to adequately train data collectors
- Failure to do a Dry Run/Failure to try enrolling a mock subject
- Uncertainty about when, how and who should review gathered data.

Hazard: failure to understand data management

- Data should be easy to understand, and the protocol good enough that another researcher can repeat the study.
- Data audit: keep raw data and collected data
- Failure to keep backups

Annotated bibliography

1. RCR Data Acquisition and Management. This online book is pretty comprehensive. http://ccnmtl.columbia.edu/projects/rcr/rcr_data/foundation/ (Accessed 2019 June 23)
2. Qualitative research – Wikipedia: en.wikipedia.org/wiki/Qualitative_research (Accessed 2019 June 23) – this is a good overview with references so you can delve deeper if you wish.
3. Qualitative Research: Definition, Types, Methods and Examples: https://www.questionpro.com/blog/qualitative-research-methods/ (Accessed 2019 June 23) – this is a good overview with references so you can delve deeper if you wish.
4. Qualitative Research Methods: A Data Collector's Field Guide: https://course.ccs.neu.edu/is4800sp12/resources/qualmethods.pdf (Accessed 2019 June 23) – another on-line resource about data collection.

Additional reading about statistical variables

1. Types of Variables in Statistics and Research: A List of Common and Uncommon Types of Variables. https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/types-of-variables/
2. Research Variables: Dependent, Independent, Control, Extraneous & Moderator. https://study.com/academy/lesson/research-variables-dependent-independent-control-extraneous-moderator.html
3. Knatterud GL. Rockhold FW. George SL. Barton FB. Davis CE. Fairweather WR. Honohan, T. Mowery R. O'Neill R. (1998). Guidelines for quality assurance in multicenter trials: a position paper. Controlled Clinical Trials, 19:477–493.
4. Whitney CW. Lind BK. Wahl PW. (1998). Quality assurance and quality control in longitudinal studies. Epidemiologic Reviews, 20 [1]: 71–80.

Additional relevant information to consider

Consider *who owns the data* before and after collection (this brings up questions of consent, privacy, sponsorship and data-sharing, most of which are beyond the scope of this paper).

Authors' contribution

Authors contributed as follow to the conception or design of the work; the acquisition, analysis, or interpretation of data for the work; and drafting the work or revising it critically for important intellectual content: ES contributed 70%; VT, MJ and HS contributed 10% each. All authors approved the version to be published and agreed to be accountable for all aspects of the work.

Declaration of competing interest

The authors declared no conflicts of interest.

References

1. Dudovskiy John. The ultimate guide to writing a dissertation in business studies: a step by step assistance E-book https://research-methodology.net/.
[2] Brenner Barry, Cheng David, Clark Sunday, Jr Carlos A Camargo. Positive association between altitude and suicide in 2584 U.S. counties. High Alt Med Biol 2011;12:31–5.
[3] Reno E, Brown TL, Betz ME, Allen MH, Hoffecker L, Reitinger J, et al. Suicide and high altitude: an integrative review. High Alt Med Biol 2018;19(2):99–108. https://doi.org/10.1089/ham.2016.0131. Jun. [Epub 2017 Nov 21].
[4] Patridge EF, Bardyn TP. Research electronic data capture (REDCap). J Med Libr Assoc 2018;106(1):142–4. https://doi.org/10.5195/jmla.2018.319.