Statistical tools and packages for data collection, management, and analysis - A brief guide for health and biomedical researchers

Vishal Deo, Priya Ranganathan[1]

National Institute for Research in Digital Health and Data Science, Indian Council of Medical Research, New Delhi, [1]Department of Anaesthesiology, Tata Memorial Centre, Homi Bhabha National Institute, Mumbai, Maharashtra, India

Abstract Previous articles in this series have looked at various aspects of planning, designing, conducting and interpreting biomedical research. In this article, we offer an overview of some tools and resources available to health and biomedical researchers, to help them with their research.

Keywords: Analysis, data collection, software tools, statistical data

Address for correspondence: Dr. Priya Ranganathan, Department of Anaesthesiology, Tata Memorial Centre, Homi Bhabha National Institute, Mumbai - 400 012, Maharashtra, India.
E-mail: drpriyaranganathan@gmail.com
Received: 31-08-24, **Accepted:** 05-09-24, **Published:** 09-10-24.

INTRODUCTION

The quality of a research study depends largely on the quality of its data. Moreover, the quality of a dataset is determined by its characteristics such as validity, accuracy, completeness, and consistency.[1,2] Efforts toward collecting good-quality data start right at the stage of developing a study protocol. To achieve high quality of data, a study must have an appropriate design with adequate sample size, and an efficient data collection mechanism with inherent data quality checks. Research study designs and principles of sample size calculation for different types of research studies have been discussed in previous articles.[3-8] In this article, we will provide an overview of some free or low-cost resources available for various research activities such as power and sample size calculation, randomization, data capture, data analysis, and visual representation of data. This is not meant to be a comprehensive overview since there are a very large number of resources, all of which cannot be covered.

POWER AND SAMPLE SIZE CALCULATIONS

Methods for sample size estimation are based on complex statistical theories spanning the concepts of probability, hypothesis testing, and confidence intervals. As a result, more often than not, sample size calculation may act as a deterrent for medical and public health researchers while developing a research protocol. However, for most of the standard study designs and research outcome measures, the resultant formulae for obtaining the estimate of adequate sample size are well established and popularly known in the literature. In addition, several web-based and software-based tools are available which provide platforms for the easy implementation of these formulae. Power of the test (or the concerned study) and sample size are

Access this article online	
Quick Response Code:	Website: www.picronline.org
	DOI: 10.4103/picr.picr_160_24

How to cite this article: Deo V, Ranganathan P. Statistical tools and packages for data collection, management, and analysis - A brief guide for health and biomedical researchers. Perspect Clin Res 2024;15:209-12.

functions of each other and knowledge of one is required to calculate the other. Accordingly, these tools provide options for calculating both sample size and power. A simple exploration on a search engine provides an overwhelming list of open access sample size and power calculators. However, one must be cautious in choosing a calculator. It is advisable that researchers use tools which are developed or hosted by a recognized institution/organization, and which preferably provide well-documented references for the methods. In addition, while these tools may act as a guide for researchers, it is always best to get the calculations verified by a qualified statistician.

RANDOM ALLOCATION

An important part of any clinical trial is the method used to allocate participants to the various treatment groups, which includes the process of generating a random number sequence (randomization) and the process used to conceal the sequence (allocation concealment). The ideal software used for the random allocation process should allow generation of study identity (to conceal participant identity), methods for variations of randomization (permuted block, stratified randomization, unequal allocation, etc.), and a mechanism to break the code if the protocol mandates it. Martin Bland has a web page (https://www-users.york.ac.uk/~mb55/guide/randsery.htm) which lists various resources (randomization software and randomization programs) for random allocation.

ELECTRONIC DATA CAPTURE AND MANAGEMENT

Electronic tools may be used either as an alternative to, or synchronously with paper data collection forms. A good electronic data capture system should be user-friendly, allow data validation, have an audit trail, export easily to a variety of statistical software, and have security features to protect data confidentiality. Additional provisions such as automatic creation of data dictionary or metadata and assignment of role-based access rights to research personnel improve the quality and ease of data collection and management.

DATA ANALYSIS

Data analysis may involve a wide range of statistical techniques depending on the study objectives, outcome measures, and the nature of the data. By and large, statistical data analysis can be broadly classified into six processes: data cleaning, descriptive analysis, estimation and hypothesis testing (statistical inference), correlation and regression analysis, nonlinear modeling, and multivariate analysis. Descriptive analysis helps to characterize the data through summary statistics and data visualization. It is fundamental toward developing a preliminary understanding of the data and identifying possible patterns and associations. It also helps in detecting data inconsistencies such as outliers, missing values, and violation of distributional assumptions, among others. While data cleaning techniques may be considered as precursors to data analysis, they are also employed for removing data inconsistencies identified through descriptive analysis. Some data collection and management tools, such as Research Electronic Data Capture, Epi Info, Census and Survey Processing System, etc., have built-in options for basic data analysis such as summary statistics, graphical visualization, and hypothesis tests.[9-11] Relatively more analysis options are available in spreadsheets such as Microsoft Excel (MS Excel) and Google sheets. MS Excel, which is a part of the Microsoft Office suite, is a popular spreadsheet application for collecting and storing data, and has an elaborate list of functions, calculations, pivots, and charts for analyzing the data. Availability of various add-ins in MS Excel, like the Analysis ToolPak, Solver, etc., make it possible to perform additional tasks such as hypothesis tests, correlation and linear regression analysis, basic time-series analysis, and optimization.

To carry out a more comprehensive data analysis, statistical software packages such as IBM-SPSS (IBM Corp., Armonk, New York, USA), Stata (StataCorp LLC, Texas, USA), R (R Foundation for Statistical Computing, Vienna, Austria), Python (Python Software Foundation, Lafayette Boulevard, Virginia, USA), SAS (SAS Institute Inc., North Carolina, USA), etc., can be used. Some of these packages, such as IBM-SPSS, Stata, and SAS have a menu-driven and user-friendly interface for nontechnical users which enables them to run even highly complex statistical analysis with just a few clicks.[12-14] Users who are comfortable in programming may prefer using commands and codes in Stata and SAS to widen their scope of analysis. Statistical packages such as SAS, IBM-SPSS, and Stata offer more advanced analysis options but are expensive and require some expertise.

For technical users, R and Python are among the most popularly used programming languages for statistical computing and graphics. Both are freely available software, with extensive coverage of statistical and computational methods through numerous packages.[15,16] R is specifically designed as an environment for statistical analysis and is an integrated suite of software facilities for data manipulation, calculation, and graphical display.[15]

Many other statistical software and tools are available as well. The choice of software for analysis should

Table 1: Some easily accessible web-based resources and software tools with their features

Resource	Features
PS: Power and sample size calculation https://biostat.app.vumc.org/wiki/Main/PowerSampleSize https://cqsclinical.app.vumc.org/ps	Free software Allows calculations for studies with continuous, dichotomous or time-to-event outcomes Has downloadable versions for iOS, Windows and Linux operating systems In addition, there is a web-based program Allows sample size calculation, power calculation, and detectable alternative hypothesis for a given sample size and power
REDCap https://projectredcap.org	Free for consortium members Server-based Create data entry forms and research databases Exports to most statistical software tools Allows multiple simultaneous projects and simultaneous access from multiple sites Mobile-based applications to increase functionality Facilitates electronic consent Audit trail Multiple language options
MS Excel	Part of Microsoft Office suite Spreadsheet for creating database and formatting Data files are compatible with all data analysis software and programming languages Has many built-in formulae for data analysis Provides option for user-defined formulae A wide range of visualization options through graphs Add-in features such as "Analysis ToolPak" Prompt-based AI co-pilot option available in recent versions in Microsoft 365
Google Sheets	Similar to Excel in many ways Differences are Needs internet connectivity to activate all features Allows collaboration - sharing of databases in real-time Has revision history - earlier versions can be accessed
Sealed Envelope https://www.sealedenvelope.com/	Randomization software Has a free option for first 50 randomizations Also has an online database application (Red Pill) for EDC and electronic patient reported outcomes Offers basic power and sample size calculation for trials with binary and continuous outcomes
MedCalc https://www.medcalc.org/	A statistical software package Requires purchase of license Trial version is free Data management options with integrated spreadsheet Includes more than 220 statistical tests, procedures and graphs, with ROC curve analysis, method comparison and quality control tools
Epi Info https://www.cdc.gov/epiinfo/index.html	Developed by Centers for Disease Control and Prevention Free resource: Can be used to create forms, collect data, and perform epidemiologic data analysis and visualization (graphs and maps) Used in epidemiology/public health Mobile version for use with tablets or smartphones to conduct epidemiologic studies in the field
NVivo https://lumivero.com/products/nvivo/	Used for the analysis of unstructured text, for example, interviews or focus groups Helps in transcribing and coding data and sorting into thematic areas Used in qualitative and mixed-methods research Trial version is free
CSPro https://www.census.gov/data/software/cspro.html	Free public domain software package for entering, editing, tabulating, and disseminating census and survey data Developed and supported by the U.S. Census Bureau and ICF Macro Supports data collection on android devices (phones and tablets) CSEntry Android App works in collaboration with the desktop version of of CSPro Supports smart data transfer from Android or Windows devices to a server running CSWeb Also contains a sophisticated programming language to create highly customized applications like quality checks Has freely available extensive learning resources (https://www.census.gov/programs-surveys/international-programs/events/training.html)

Contd...

Table 1: Contd...

Resource	Features
ODK https://getodk.org/	An open-source software for creating forms and collecting data Users can self-host and self-support it for free but requires technical expertise ODK Cloud is a paid version. It is the same ODK software, but fully hosted, managed Powerful data collection forms can be built with options for photos, GPS locations, skip logic, calculations, external datasets, multiple languages, and more Works both online and offline through mobile app and web app Offline data are automatically synced once internet is connected Option to connect with applications such as MS Excel, R, Python, or Power-BI to create real time dashboards

REDCap=Research Electronic Data Capture, EDC=Electronic data capture, ROC=Receiver operating characteristic, CSPro=Census and Survey Processing System, ODK=Open Data Kit, MS Excel=Microsoft Excel, AI=Artificial intelligence, PS=Power and Sample Size

be based upon the level of analysis to be performed and the technical capacity of the user. In addition, researchers must ensure credibility of the software or tool, especially if they are not well known in academic circles. Authenticity of developers, validation of software and tools, and information on scientific peer-review or associated publications are some of the aspects to focus on before using any such software. Ideally, statistical analysis software should offer a wide range of statistical techniques, handle multiple types of data file formats, allow hassle free import and export of data files, have adequate options for graphs and visualizations, provide output in tangible formats, and preferably allow tracing and saving of command flow.

For a quick reference, a list of some easily accessible web-based resources and software tools with their features are provided in Table 1.

CONCLUSION

There are several free and paid statistical software tools available to perform the tasks of sample size and power calculation, randomization, data collection, data management, and data analysis. In general, criteria for choosing a tool should include cost considerations, credibility of the tool, ease of use as per the technical capacity of the user, range of functions, and features such as audit trail and validation. Availability of such a wide range of options enables the researchers in the field of health and clinical research to effectively plan and execute their studies. However, researchers must ensure appropriateness of methods before implementing them through these tools and packages.

Financial support and sponsorship
Nil.

Conflicts of interest
There are no conflicts of interest.

REFERENCES

1. Schmidt CO, Struckmann S, Enzenbach C, Reineke A, Stausberg J, Damerow S, *et al*. Facilitating harmonized data quality assessments. A data quality framework for observational health research data collections with software implementations in R. BMC Med Res Methodol 2021;21:63.
2. Davis JR, Nolan VP, Woodcock J, Estabrook RW, editors Institute of Medicine (US) Roundtable on Research and Development of Drugs, Biologics, and Medical Devices. Assuring Data Quality and Validity in Clinical Trials for Regulatory Decision Making: Workshop Report. Washington (DC): National Academies Press (US); 1999.
3. Aggarwal R, Ranganathan P. Study designs: Part 2 – Descriptive studies. Perspect Clin Res 2019;10:34-6.
4. Ranganathan P, Aggarwal R. Study designs: Part 3 – Analytical observational studies. Perspect Clin Res 2019;10:91-4.
5. Aggarwal R, Ranganathan P. Study designs: Part 4 – Interventional studies. Perspect Clin Res 2019;10:137-9.
6. Ranganathan P, Pramesh CS, Aggarwal R. Equivalence trials. Perspect Clin Res 2022;13:114-7.
7. Ranganathan P, Pramesh CS, Aggarwal R. Non-inferiority trials. Perspect Clin Res 2022;13:54-7.
8. Ranganathan P, Deo V, Pramesh CS. Sample size calculation in clinical research. Perspect Clin Res 2024;15:155-9.
9. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap) – A metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009;42:377-81.
10. Centers for Disease Control and Prevention (CDC). Epi Info™ User Guide. USA. Available from: https://www.cdc.gov/epiinfo/support/userguide.html. [Last accessed on 2024 Aug 28, Last updated on 2022 Sep 16].
11. United States Census Bureau. Census and Survey Processing System (CSPro). USA Available from: https://www.census.gov/data/software/cspro.html. [Last accessed on 2024 Aug 28, Last updated on 2024 Jan 24].
12. IBM SPSS Statistics. Simplify Data Analysis with an Intuitive, Easy-to-Use Statistical Solution for Data-Driven Decisions. Available from: https://www.ibm.com/products/spss-statistics. [Last accessed on 2024 Aug 28].
13. StataCorp LLC. Stata's Interface. Available from: https://www.stata.com/features/overview/graphical-user-interface/. [Last accessed on 2024 Aug 28].
14. SAS Institute Inc. Overview of the SAS Interface. Available from: https://documentation.sas.com/doc/en/pgmsascdc/9.4_3.5/hostwin/p0j67928u0uygqn17gi0dasyot6p.htm. [Last accessed on 2024 Aug 28].
15. The R foundation. What is R? Available from: https://www.r-project.org/about.html. [Last accessed on 2024 Aug 28].
16. The Python Software Foundation. What is Python? Executive Summary. Available from: https://www.python.org/doc/essays/blurb/?external_link=true. [Last accessed on 2024 Aug 28].