Vol 1 No 1 April 2023
e-ISSN 2988-7283

Editorial

Analysis Using R Software: A Big Opportunity for Epidemiology and Public Health Data Analysis

Rinaldi Daswito[1, 2*], Besral[2], Radian Ilmaskal[2, 4]

[1]Health Polytechnic MoH Tanjungpinang, Indonesia

[2]RRZ Scientific Publishing

[3]Department of Biostatistics, School of Public Health University of Indonesia

[4]Alifah Padang Health Science College

*Email the corresponding author: rinaldi@poltekkes-tanjungpinang.ac.id

Abstract

R is a programming language, open-source, developed by various of the world's most active statisticians with powerful function and visualization for data analysis from simple to complex data such as machine learning and artificial intelligence. Data visualization technologies have the ability to assist public health professionals with decision-making. Visualization appears to help decision making by increasing the quantity of information communicated and reducing the cognitive and intellectual strain of processing information. There are numerous commercially available statistical software packages that are widely utilized by epidemiologists worldwide. For industrialized nations, the price of software is not a significant issue. However, for underdeveloped nations, the true expenses are frequently excessive. Some academics in developing nations rely on software that has been illegally copied a copy of the software program. There are several benefits to using R, including the possibility of using software packages for free (open source) and the volume and availability of software packages. It is simple to retain and repeat commands on the same data analysis with multiple data frames, facilitating the work of health monitoring officers who frequently analyze data with similar variables but at different times.

Keywords: R Software, Data Analysis, Epidemiology, Public Health

INTRODUCTION

R is a programming language with numerous statistical functions built in. This programming language is easily extensible using user-written functions. R is an open-source program that permits several parties to contribute to its development. R is the ideal platform for beginning your data science and data engineering journey since R's environment was built from the ground up to enable data science. In addition to being a programming language, R provides an interactive environment for conducting data research; its language is significantly

more versatile than other programming languages. (Greenacre, 2022) (Harrison & Riinu, 2020)(Wickham & Grolemund, 2017).

R is a GNU General Public License project established by John Chamber and friends at Bell Laboratories (previously AT&T, now Lucent Technologies). The app's name is partially derived from the (first) names of the original two R authors, Robert Gentleman and Ross Ihaka, and the concept is partially derived from the name of the language developed by Bell Labs, "S." (Khan, 2013). By the beginning of the 1990s, “S” had gained widespread popularity among statisticians, and various implementations were in place. One of these is the commercially licensed S-PLUS, which includes a graphical user interface (GUI). Ross Ihaka was the recently appointed Professor of Statistics at the University of Auckland (New Zealand) during this time, where he met Robert Gentleman, a Professor from the University of Waterloo (Canada). Their significant meeting initiated the creation of programming languages for statistical analysis and instruction (Giorgi et al., 2022).

In 1993-1994, when R was at its height of popularity, many academic users began contributing voluntarily to the project. In June 1995, Ihaka and Gentleman released the R source code under the GNU general public license, after being persuaded by the nascent community (particularly Martin Maechler of ETH Zurich) to release the full language as free software. This assures that R software remains freely accessible and cannot be exploited for profit. This choice to make R freely available as a tool for all mankind contributed to its enormous popularity in the years that followed, and the R programming language eventually surpassed S and other S derivatives (Giorgi et al., 2022).

After its debut as open-source software, R quickly established itself as an accessible and potent data management, analysis, and visualization tool. Internet development enables the construction of an international community of R users, including both statisticians and non-statisticians. In 1997, ETH Zurich hosted the first online discussion room for R in the form of three mailing lists: r-announce, r-devel, and r-help. Official mailing lists, particularly r-devel and r-help, are still in use today, despite being supplanted by newer web aggregators capable of ranking questions and answers as well as popularity, such as the programming-oriented Stack Overflow and the statistics-oriented Cross Validated (Giorgi et al., 2022).

The expanding R community began creating libraries, which are specialized additions to R's core code, necessitating an official packaging method and software repository. In 1997, Kurt Hornik announced the establishment of CRAN, the Comprehensive R Archive Network, on the r-announce mailing list. On February 29, 2000, R reached its first stable release version (1.0.0). As the breadth and scope of the R community increased, so did the need to maintain CRAN server mirrors and support core developers; consequently, the R Foundation was founded in 2003 to support it. Today, the Statute R Foundation continues to promote and administer R projects and serve as a resource for the R community.

In April 2020, R 4.0.0 was released, seven years after R 3.0.0 and more than twenty years after R 1.0.0. This version includes further memory allocation optimizations (such as when an object is copied to another object) and enhancements geared mostly at novice users. A great addition to R version 4 is the ability to import string data as character objects. Until recently, the default memory-saving parameter imported string data as a factor, which had advantages but was not clear for novices.

Bioconductor was founded in 2001 under the direction of Robert Gentleman with the general objective of building and developing R tools for biological data analysis that statisticians and computational biologists may use. Bioconductor got financial and institutional backing from the start and expanded rapidly, releasing its first stable version

(1.0.0) in 2002. The Bioconductor project is now the largest collector of computing tools for evaluating biological data, particularly quantitative data.

On June 10, 2007, programmer and statistician Hadley Wickham, then a PhD student at Iowa State University, developed a graphical R package that would not only become more popular than basic R plotting functions but also spark a grammatical revolution in the R community. ggplot2, in its most basic form, is a R package (downloadable from CRAN) that adds convenient functions for graphically exploring data, and the elegant "look and feel" of its graphics has become very popular. Based on the grid and grid core packages, Ggplot2 can currently be regarded as a potent substitute for R's fundamental charting capabilities. Despite variations in data input formats and methods, ggplot2 and basic R are capable of producing similar graphical output, particularly since the R 4.0.0 upgrade, which introduced a more aesthetically acceptable standard color palette.

Currently, RStudio manages and promotes the tidyverse package, which includes popular programs such as ggplot2 and dplyr. RStudio was not only conceptualized and envisioned as an editor for writing and running R code but also as an expanding universe for the growth of R and for broadening the scope of programming languages beyond statistical analysis. The RStudio team produced two milestone packages in particular to fulfill this goal. Shiny, which was released in November 2012, extends R's functionality to the web and enables the language to build reliable and current online resources. The second is markdown, which was officially released in 2014 and aims to provide a reproducible pipeline and convey results by embedding R code and results in dynamic documents.

R was historically developed by some of the world's most active statisticians, including in the most current machine learning and artificial intelligence breakthroughs (AI). In reality, some of the founding fathers of machine learning, including Trevor Hastie and Robert Tibshirani, were R programmers and continue to be active producers of R libraries. CRAN contains several hundred R libraries that execute a variety of machine learning tasks and approaches (Giorgi et al., 2022). In other words, R is sophisticated software that aids humanity, particularly academics in data analysis, data science, and artificial intelligence, including in the public health field.

Using of R

R offers a variety of statistical methods, including linear and nonlinear modeling, classical statistical testing, time-series analysis, classification, and clustering, among others. R additionally includes graphic engineering tools for the graphical representation of processed data (GNU, 2023).

As a health data scientist, it is crucial to have a strong grasp of statistical programming languages and the ability to work in a transparent and reproducible manner. To be able to work with health data, it is essential to be familiar with and capable of utilizing statistical software to manipulate, analyze, and visualize data. R is an open-source and free piece of statistical software. It is appealing not only because of its monetary benefits but also because of its vast community of users and creators. There are also other well-documented advantages to using R, such as the fact that the majority of academic statisticians use R, that R is platform-independent (it can be used on Windows and Linux/Mac), allowing researchers on different platforms to collaborate, as well as the abundance of help and resources available, such as books and online forums. R also supports and interacts with a number of tools to document your code, making it efficient and reproducible (Harrison & Riinu, 2020)(Greenacre, 2022).

Health Data and Decisions

Data visualization technologies have the ability to assist public health professionals with decision-making. This article highlights the science and evidence around data visualization and its effect on decision-making behavior influenced by cognitive processes such as comprehension, attitudes, and perceptions. Visualization appears to help decision making by increasing the quantity of information communicated and reducing the cognitive and intellectual strain of processing information. However, awareness of specific data visualization interventions for public health leaders' decision-making is inadequate, and there is little guidance for identifying the qualities and responsibilities of participants. Depending on the control of confounding factors, this review reveals that favorable benefits of data visualization can be identified in attitudes, perceptions, and decision-making (Park et al., 2022).

Resource Availability for Application Development and License Purchase

The importance of data analysis in epidemiological research cannot be overstated. The capacity of computing facilities is continuously expanding, and cutting-edge epidemiological research is also progressing in the same manner. Today, there are numerous commercially available statistical software packages that are widely utilized by epidemiologists worldwide. For industrialized nations, the price of software is not a significant issue. However, for underdeveloped nations, the true expenses are frequently excessive. Some academics in developing nations rely on software that has been illegally copied a copy of the software program.

Advantages and Disadvantages of using R

There are various advantages to using R, including the possibility of using software packages for free (open source) and the quantity and availability of software packages. The EpiInfo package, for example, is free and useful for data entry and basic data analysis. R can be used by advanced data analysts. Data processing in longitudinal investigations, for example, can accommodate repeated measurements and multilayer modeling. In addition, a variety of graphing options are available (Epical Book). R, a free and relatively new software, is very impressive. This program, which is supported by leading statisticians from around the world, has almost everything an epidemiological data analyst needs. However, R is more difficult to learn and use than comparable statistical tools for epidemiologic data analysis, such as STATA.

Opportunities for the Use of R in the Fields of Epidemiology and Public Health

Analyzing epidemiological data has always been difficult, particularly for scientists with a background in the biological sciences rather than mathematics. Due to the often high size of epidemiological data sets, manually calculating simple statistics such as the mean or standard deviation is laborious. Even locating a statistician becomes challenging for many individuals in their environment. Even with exploratory data analysis and simple descriptive data analysis, there are so many data sets that have yet to be studied that they may wait forever.

The command-line interface (CLI) is the ideal user interface for novices since it permits direct control over calculations and is flexible. However, proficiency in this language is necessary. Consequently, CLI can be challenging for beginners. It is simple to retain and repeat commands on the same data analysis with multiple data frames, facilitating the work of health monitoring officers who frequently analyze data with similar variables but at different times. Therefore, a user of R must comprehend what he is doing in order to evaluate the data; without this knowledge, it is nearly impossible to examine the data. As open-source software,

hackers can uncover software vulnerabilities or loopholes more quickly than with closed-source software, making it more susceptible to bug assaults.

DECLARATION OF CONFLICTING INTEREST

None

FUNDING

None

ACKNOWLEDGMENT

None

REFERENCES

Giorgi, F. M., Ceraolo, C., & Mercatelli, D. (2022). The R Language: An Engine for Bioinformatics and Data Science. Life, 12(5), 648. https://doi.org/10.3390/life12050648

GNU. (2023). What is R? https://www.r-project.org/about.html

Greenacre, M. (2022). R for Health Data Science. Journal of the Royal Statistical Society Series A: Statistics in Society. https://doi.org/10.1111/rssa.12851

Harrison, E., & Riinu, P. (2020). R for Health Data Science. In R for Health Data Science. CRC press. https://doi.org/10.1201/9780367855420

Khan, A. (2013). R-software: A newer tool in epidemiological data analysis. Indian Journal of Community Medicine, 38(1), 56. https://doi.org/10.4103/0970-0218.106630

Park, S., Bekemeier, B., Flaxman, A., & Schultz, M. (2022). Impact of data visualization on decision-making and its implications for public health practice: a systematic literature review. Informatics for Health and Social Care, 47(2), 175–193. https://doi.org/10.1080/17538157.2021.1982949

Wickham, H., & Grolemund, G. (2017). R for Data Science: Visualize, Model, Transform, Tidy, and Import Data. In O'Reilly Media.