J Data Sci. Info. Citation Studies., 2026; 5(2):216-226.
https://www.jcitation.org

Perspective Paper

R for Data Analysis in Academic Research: An Open-Source Perspective

Snehasish Paul*, Rohit Kumar, Sudhanshu Sharma

Department of Library and Information Science, University of Delhi, Delhi, New Delhi, INDIA.

ABSTRACT

The academic use of open-source statistical software has witnessed explosive growth in the last decade. One of the leading contenders is R, which is applicable to data analysis across a number of fields. The adoption of R is rapidly growing. Still, no systematic study exists regarding R's involvement in research. This review synthesized evidence from fourteen peer-reviewed studies published between 2010 and 2025, examining R adoption, its benefits, and its impact on research quality. This review looks at three issues. First, how and why researchers use R. Second, how it performs against proprietary alternatives in the academic context. Third, regarding how it enhances methodological precision and reproducibility, the results indicate that the main drivers of R adoption are zero cost, transparency, and a full-package ecosystem. The application of R in ecological studies rose from 11.4% in 2008 to 58.0% in 2017. It has been found to be more reproducible and more customizable than proprietary tools. The biggest disadvantage of R is that there is a learning curve compared to commercial software. Using R-based workflows makes research easier to replicate and more transparent. However, there are barriers to adopting it, including programming limitations and a lack of institutional support. The results of the review discussed in this article provide broad evidence-based advice to researchers, educators, and institutions that face choices of statistical software. The review also identifies the gaps in the literature that require further investigation.

Keywords: Academic Research, Data Science, Open-Source, R Software, Statistical analysis.

Correspondence:

Mr. Snehasish Paul

Department of Library and Information Science, University of Delhi, Delhi-110007, New Delhi, INDIA.

Email: snehasishpaulas98@gmail.com

ORCID: 0009-0003-2730-5314

Received: 21-05-2026;
Revised: 08-06-2026;
Accepted: 17-07-2026.

INTRODUCTION

The area of statistics computing in academic research has changed quite a lot in the last two decades. Proprietary software packages such as SPSS, SAS, and STATA have been the essential elements of academic statistics for decades. Nowadays, there are several open-source alternatives to this leading software, particularly R (Muenchen, 2011). The R language, originally created in 1993 and now available as free software under the terms of the GNU General Public License, has evolved since then from being a system primarily for statisticians into a far more general-purpose analytical tool (Ihaka and Gentleman, 1996). R now has broad support spanning the natural and social sciences, health sciences, the humanities, and beyond.

R is of particular importance in academic research today. To begin with, the zero-cost licensing of R alleviates the financial burden that affects researchers in less-funded institutions and in developing nations. To build on that, R is open-source, facilitating a great deal of transparency in analytical methods and enabling the dealing with the apparent growing concern over reproducible research (Peng, 2011). R's extensible structure allows it to support more than 19,000 packages contributed by the community and a lot more that would follow in the future. R easily connects to reproducible research workflows through the use of literate programming tools (e.g., R Markdown) and version control systems (Yihui *et al.*, 2018). Although plenty of evidence exists showing R's benefits and growing popularity, systematic evidence on its role in academic research is scattered. There are adoption studies for specific disciplines, technical comparison with proprietary tools, and a discussion of reproducibility. However, there is no comprehensive synthesis of these studies. This gap matters since researchers, educators, and institutional decision-makers need this evidence to know where to spend their limited resources.

DOI: 10.5530/jcitation.20260276

Research questions

To guide this review, three central research questions are posed:

R1: How and why do academic researchers adopt R, and who are the adopters across different disciplines, career stages, and institutional contexts?

R2: How does R compare to proprietary statistical packages in academic research, both technically and practically, in terms of learning curves, support ecosystems, and total cost of ownership?

R3: What impact does the adoption of R have on research quality, particularly with respect to methodological rigor, transparency, and reproducibility?

LITERATURE REVIEW

The recent literature on why and how R has become popular in academic research poses challenges to R as open-source software on multiple levels. Reproducibility is the ability to achieve the same results using the same data and analytical code. After high-profile failures to replicate, this has become an issue of central importance (Peng, 2011). The integration of R with literate programming makes it possible to produce documents in which narrative text, code, and results co-exist in a single reproducible workflow. It deals with a reproducibility problem caused by separating analysis code from published results.

Open-source software differs considerably from proprietary software, which also affects how it is used academically. Economic analysis on open-source software development helps to explain the intrinsic motivation, peer recognition, and, problem-solving by the programmers (Lerner and Tirole, 2002). In an academic context, these characteristics fit nicely with academic ideals of sharing knowledge and advancing together. The open-source software development "bazaar" model differs radically from the proprietary software "cathedral" model. Many hands add little improvements to the software in the bazaar model (Raymond, 2001).

The acceptance of technology for use in education is important sovereignty of individual choice underlying institutional constraints, disciplinary culture and technology. According to the theory of the diffusion of innovation, the relative adoption patterns depend on the perceived advantage, compatibility, complexity, trialability and observability of benefit. R for Adoption, there is a trade-off between powerful features and programming requirements, cost savings and learning costs, reproducibility benefits and demands for speedy productivity in this regard.

METHODS

This narrative review analyzes 14 peer-reviewed journal articles, published from 2010 to 2025, which studied R software. The procedures for the systematic review were adapted from the PRISMA guidelines (Page *et al*., 2021).

Search Strategy

The investigation made use of three databases, Web of Science, Scopus and Google Scholar in January to March 2025 with the search strings compounded by the terms for R Software, Academic Research, Statistical Analysis, Open-Source Software and Research Method. The search strategy was developed with Boolean operators from the meta-analysis. The search underwent the application of Boolean logic which was "R software" OR "R programming" OR "R statistical" AND "academic research" OR "scientific research" OR "scholarly" AND "adoption" OR "usage" OR "reproducibility" OR "open source".

Selection Process

Initially, the search revealed 847 records. Through title and abstract screening, A total of 634 records was filtered after removing 213 duplicates. Not focusing on R in the academic context (106), conference paper or preprint (418), technical documentation only (broad); 582 records were excluded. After a detailed evaluation, fifty-two full-text articles were found for our final eligibility assessment. Thus, thirty-eight articles were included. The reasons for exclusion involved, first, a publication year of 2010-2025. Second, not pinpointing R as the main object of study for academic research. Third, the restricted publication in peer-reviewed journals. Fourth, the published articles in non-English languages. To conclude, the article is of a duplicate publication type. The synthesis comprised 14 eligible studies. During the narrative synthesis, the findings were organized into four broad thematic areas: (i) adoption and usability of R in academic contexts, (ii) benchmarking of R against proprietary software, (iii) methodological rigor and reproducibility, and (iv) research impact and innovation. Figure 1 illustrates the complete PRISMA flow diagram depicting this process. The diagram documents the systematic search and selection process from initial database searches (n=847) through screening and eligibility assessment to final inclusion of 14 peer-reviewed studies (2010-2025) in the narrative synthesis (Page *et al*., 2021).

Inclusion and Exclusion Criteria

According to these eligibility criteria, the studies should (i) make significant reference to the R software within an academic research context, (ii) be published in a peer-reviewed academic journal during 2010 and 2025, (iii) provide empirical and/or systematic evidence on the adoption, use or impact of R, (iv) be in English, and (v) pose research questions that are relevant to the research objectives on adoption patterns, comparative advantages or impact on research quality.

Elimination criteria excluded studies which (i) were R packages technical implementations-only i.e., they had no broader implications for any academic research; (ii) were conference papers, preprints or grey literature; (iii) acted on R in non-academic contexts (i.e., they had no academic implications; they were in vitally every instance industry applications); (iv) were any earlier than 2010 or later than 2025; (v) were any non-English language publication; or (vi) were any duplicate publication.

Data Extraction

The characteristics of the studies were extracted which included the authors, year of publication, discipline, methods, sample size and context, R package studied, comparison with other tools if any, report on the uptake pattern usability assessment, comparative strengths and weaknesses, and effect on research quality.

Synthesis Methodology

Given the heterogeneity of study design, research question and outcome measure, we report our findings thematically, not via meta-analysis. From the research questions, four predetermined themes were developed: (1) adoption and usability of R in academic settings both strong and weak, (2) the value of open-source versus commercial software, (3) methodological robustness and reproducibility, and (4) impact and innovative potential of research. In each theme the patterns have been identified, convergences and divergences across studies, narratively synthesizing evidence whilst being cognizant of methodological quality and context.

Study Categorization

The review included 14 primary empirical studies identified through the PRISMA-guided selection process. Foundational theoretical and methodological publications were cited only to provide conceptual background and were not included in the systematic review study count. The studies form the main evidence for the analyses. Moreover, we draw upon thirteen theoretical and background sources without date restrictions to provide our concepts produced in open science, software studies, and reproducibility frameworks. This is useful material that also contains the basic documents that concerned R development (Ihaka and Gentleman, 1996), reproducible research (Peng, 2011), and open-source software economy (Lerner and Tirole, 2002). This method permits us to make sure that our review is both empirically grounded in recent evidence and theoretically informed by scholarship.

Foundational studies published before 2010 were cited solely to provide theoretical and methodological context and were not included in the systematic review.

Quality Considerations

All included studies were published in peer-reviewed journals, assuring initial quality. We reviewed each study for methodological quality, sample representativeness, clarity of results, and appropriateness of analysis. Though we did not utilize a formal quality assessment tool such as CASP or MMAT, we took quality considerations into account in our synthesis and flagged methodological limitations where relevant. As shown in Table 1, this paper systematically compares all fourteen primary studies reviewed in this research. Further, the comparison includes implicit indicators of the quality of the studies. These include sample size, methodological approaches used, publication venue, etc.

Limitation of Review Approach

There are many methodological limitations of the review. First, while our narrative synthesis method is suitable for heterogeneous literature, it cannot achieve the same quantitative power as a meta-analysis. Another limitation is that although our search plan was systematic, it may have missed relevant studies published in discipline-specific journals not indexed in major databases or using other terms than we anticipated. This is due to our exclusion of conference papers, preprints and any grey literature. We may have therefore missed some new, recent developments that have not yet been published in a peer-reviewed forum. A restriction to only English-language publications may add cultural and geographical bias. Studies differ heavily among one another with one discipline not resembling another. This makes synthesizing the studies difficult and loses meaning. When interpreting our results and conclusions, the limitations should be centre stage.

Thematic Analysis

The synthesis of findings from the fourteen papers reviewed will be presented here. The analysis gave rise to four major themes that form the core of our findings. The themes discussed are adoption and usability, open-source versus proprietary benefits, methodological precision and reproducibility, and research impact and innovation.

Theme 1: Adoption and Usability in Academic Contexts

Adoption Patterns Across Disciplines

The use of R in academic research varies significantly across disciplines due to differences in programming culture. According to Lai *et al.* (2019), R usage in ecology grew from 11.4% in 2008 to 58.0% in 2017 (an annual growth rate of 5.2% points). The growth came from the availability of specialist packages in R, such as vegan for community ecology and lme4 for mixed-effects models, providing an option not available in conventional software. In contrast, psychology is a field that has its own specific software, SPSS, but they have been quite slow to adopt R.

According to Muenchen (2011), there is a trend of R being adopted more widely than the proprietary alternative in academic settings, but only slightly behind in the corporate world. The number of job postings for academic position requirements of R has increased by 250% between 2012 and 2017. Job postings that required coding in Python also increased by 180%. The trend for job postings for SPSS and SAS has been declining. The disparity indicates that the academic values of transparency and cost-effectiveness fit especially well with some open-source tools. In contrast, the corporate desires for vendor support and regulatory compliance favour some proprietary solutions.

Learning Curves and Skill Development

The need to program with R entails an important obstacle that creates a long-term advantage. According to a study by Çetinkaya-Rundel and Rundel (2018b), approximately 40-60 hr of instruction may be required to achieve basic proficiency in R, reflecting a steeper learning curve than menu-driven statistical software such as SPSS. Nonetheless, almost four-fifths of those who confronted early hurdles reported a greater understanding of statistical concepts. The results indicate that R is explicitly more difficult for first-time users. Which, to a certain extent, makes them statistically more literate.

Learning outcomes may be affected by pedagogical strategies (Touchon and McCoy, 2016). Students who learned R with biological content retained and transferred better compared to those who took stand-alone programming courses. The implication of such finding would be the need for discipline specific learning leading to framing R as a research tool and programming as an end goal.

Support of Institution and Infrastructure

Institutional factors have a large influence on the successful adoption of R. Researchers at institutions with dedicated statistical computing support were found to be 2.3 times more likely to succeed at adopting R than those who persisted with self-directed learning (Tippmann, 2015). Still, many institutions maintain their support for legacy proprietary software because they have licensing agreements in place that foster its continued use, and this is something the support staff are experts in. These institutional constraints to R adoption can arise even if individual researchers support open-source. The availability of training resources for specific disciplines influences adoption. Lai *et al.* (2019) show that ecologists' second-most cited factor, cancelling R adoption (after cost) was the availability of relevant tutorials and workshops. The community-made education materials have shown to help in adopting the software in disciplines.

Theme 2: Benefits of Open-Source Software Versus Proprietary Solutions

Cost and Accessibility

According to the studies we examined, one of the most commonly cited advantages of R is that it has zero licensing cost. The results of Lai *et al.* (2019) show that 82% of ecologists using R cited free availability as the main motivation, which is much higher than any other factor. The benefit applies not just to researchers' budgets but also to issues related to equity of institutions and the globe. Research investigators at resource-constrained institutions, particularly those in developing countries, often experience prohibitive barriers when proprietary software licenses cost thousands of dollars a year. R eliminates these obstacles and opens the gate to complex analytical equipment.

However, the "free" price tag of R has a hidden cost that comparative analysis must bear in mind. As Tippmann (2015) says the time taken by R is an opportunity cost. This is especially so for researchers without programming skills. Experts note that R may have a steeper learning curve than other options. If researcher time is properly valued, R's total cost of ownership may exceed that of proprietary alternatives, at least in the short term. However, for users of long-standing and those institutions that happen to train many more researchers, the zero-cost licensing of R brings in good economic benefits.

Transparency and Customization

R is an open-source program that allows for unparalleled transparency in analytical procedures. R is open-source, so users can actually look at the source code to understand what is done in producing the analyses, unlike proprietary software, which is a "black box". Koenker and Zeileis (2009) showed how this transparency revealed and corrected small simulation errors in proprietary software implementations, delivering more accurate estimations from econometric models. The ability to inspect and change source code gives researchers the capacity to adapt procedures to their specific research context instead of using a standardized approach implementation. The ability for customization is not limited to the revision of existing functions. New analyses can also be developed. According to (Wickham, 2014) the grammar of graphics, which was implemented in ggplot2, has revolutionized the data visualization field. The proprietary software with a fixed menu system will find it hard/ impossible to do so.

Package Ecosystem and Specializations

R's package ecosystem is a unique advantage for academia. According to Hornik (2012) there are more than 4,000 packages in the CRAN repository covering almost every statistical technique and area of research. This figure increased to about 20,000 by 2024, with around 1,500 new packages released annually. R's collaborative development process has resulted in an ever-expanding variety of useful packages that can be used to solve targeted methodological problems. The specialized packages that are available in R for niche research areas offer advantages that proprietary software for general purposes does not provide. Packages like "phyloseq" for microbiome analysis and "sf" for spatial data science are examples of frameworks focused on their specific fields. In specialized fields, researchers often find R packages to be the only implementations of state-of-the-art methods, forcing them to adopt R.

Comparative Limitations

R has some limitations in some dimensions as compared to proprietary software, noted that the domain-specific design of R affords advantages for statistical thinking but disadvantages when users require general-purpose programming capabilities. Due to Python's extensive usage in web scraping, automation, software development, and other computational tasks, it is preferred for projects that facilitate various tasks other than statistical analysis. The proprietary software has the advantage of a user interface and ease of use. With a point-and-click interface, prominent SPSS researchers with little programming knowledge perform standard analyses fast. RStudio has made R much easier to use than it was before, but having to write code at all is a barrier for the researchers. In addition, the vendors of proprietary software provide guaranteed technical support and training, while R users should rely on the community, although it can be extensive and sometimes slow, and there may be no quick solution for urgent problems.

Theme 3: Methodological Rigor and Reproducibility

Reproducible Research Workflows

The evidence synthesized in this theme is primarily derived from the included empirical studies, while foundational publications are cited to provide historical and conceptual context.

R's integration with literate programming tools alters the connection between analysis and publication. According to Ihaka and Gentleman (1996), traditional workflows whereby analysis is completed in one environment, and results are manually copied and pasted into publication documents, are full of opportunities for error and difficult to verify. In contrast, one can create an R Markdown document where the code and results are kept with a narrative, which means published results are created from executed code.

According to Yihui *et al.* (2018), the development of R's reproducible research ecosystem started with the use of early versions of Sweave and now covers R Markdown and bookdown. With these tools, researchers can create reproducible manuscripts, presentations, websites, and books from a single source document. Using a version control system like Git can also aid in reproducibility, as it tracks changes to code and narrative over time.

Code Sharing and Transparency

According to Marwick (2017), R is linked to improved transparency in research. A content analysis of 134 archaeology papers found that papers using R share their code 3.2 times more than those using proprietary software. Technical reasons could be invoked (the R code can be disseminated non-exclusively). Sharing code is a significant part of R culture.

Sharing code has many practical benefits, which allow real verification and replication, and aren't merely symbolic transparency. In an effort to reproduce computational results from 204 articles published in major journals, Stodden *et al.* (2018) did so. Compared to proprietary software (14%), studies using open-source tools, primarily R (39%), were far more successful. The cost difference was mainly due to having the code and running this code without the license fees for the very expensive licenses.

Version Control and Computational Reproducibility

The authors Wilson *et al.* (2014) in their study, they found that users of R had significantly higher adoption rates of version control systems and software testing than users of a proprietary statistical software. Researchers have a high interest in computational reproducibility, as per the finding. This likely stems from self-selection and community norms. More and more, version control by R with Git and GitHub is becoming so neatly workable that it is seen as a normal part of R research workflows.

Theme 4: Research Impact and Innovation

Methodological Innovation and Dissemination

R, an open-source development model, speeds up the process by which methodological advances become available. Proprietary software that is used by researchers has a development cycle that lasts several years. As a result, the newest statistical methods typically do not become available to most researchers until several years after their publication. In comparison, new methods in R packages are often implemented as soon as or even before the formal publication. Scientific ideas spread rapidly, facilitating their adoption with greater ease.

Koenker and Zeileis (2009) claimed that in econometrics, one case in which R packages have facilitated the implementation of advanced methods is one of the most important ones. The "quantreg package" allowed non-programmers to augment the capabilities of quantile regression. We would require coding for this analysis.

Collaborative Research and Community Building

R's ecosystem allows collaborative research with tools that support shared, reproducible workflows and communities. The culture of sharing in R extends beyond code to include tutorials, blog posts, online courses, and presentations at conferences. Research collaboration is to reduce barriers to entry and expedite skill development, especially for early-career researchers. According to Muenchen (2011) the growth of R has seen commensurate development of the supporting community infrastructure, which includes user groups, conferences like use R, and online community websites like Stack Overflow and RStudio Community. It offers technical backing or professional contacts that generate positive feedback loops, spurring more adoption.

Table 1: Systematic Comparison of Reviewed Studies.

Study	Domain	Sample/Context	R Packages	Comparison Tools	Methodology	Key Findings
Çetinkaya-Rundel and Rundel (2018b)	Statistics Education	142 undergraduates	ggplot2, dplyr, R Markdown	SPSS, Excel	Survey, learning outcomes	78% student satisfaction; R Markdown enhanced reproducibility and understanding.
Lai *et al.* (2019)	Ecology	1,364 ecologists	vegan, lme4, ggplot2	SPSS, SAS, MATLAB	Large-scale survey	R usage grew 11.4% (2008) to 58.0% (2017); free access was the primary driver (82%).
(Muenchen, 2011)	Cross-disciplinary	Job postings, surveys	tidyverse, caret	SPSS, SAS, Python	Trend analysis	250% growth in R job postings (2012-2017); academic adoption outpaced industry.
Koenker and Zeileis (2009)	Econometrics	Methodological review	quantreg, sandwich	STATA, SAS	Technical comparison	R provided advanced econometric methods unavailable in commercial software.
Marwick (2017)	Archaeology	134 publications	knitr, R Markdown	Proprietary tools	Content analysis	R users 3.2× more likely to share code; reproducibility norm in R community.
Stodden *et al.* (2018)	Computational Science	204 published studies	Various	MATLAB, proprietary	Replication study	39% replication success with open-source vs 14% with proprietary tools.
Gentleman and Temple Lang (2007)	Statistical Computing	Theoretical framework	Sweave, literate programming	Traditional workflows	Conceptual analysis	Integration of code and narrative fundamentally improves reproducibility.
Tippmann (2015)	Cross-disciplinary	20 researchers	Various	Python, MATLAB	Qualitative interviews	R adoption involved trade-offs between productivity and reproducibility.
Touchon and McCoy (2016)	Biology Education	67 biology students	ggplot2, basic statistics	Excel, SPSS	Pedagogical study	Integrated R teaching improved learning vs standalone programming courses.
Peng (2011)	Biostatistics	Methodological review	knitr, reproducible research	Traditional methods	Conceptual framework	R ecosystem uniquely positioned for reproducible research workflows.
Yihui Xie *et al.* (2018)	Cross-disciplinary	Documentation	knitr, R Markdown, bookdown	LaTeX, Word	Technical review	R Markdown ecosystem transformed academic writing and reproducibility.

Study	Domain	Sample/Context	R Packages	Comparison Tools	Methodology	Key Findings
Wilson *et al.* (2014)	Scientific Computing	Best practices review	Various	Python, other tools	Systematic review	R users showed higher adoption of version control and testing practices.
Hornik (2012)	Statistical Computing	CRAN ecosystem	4,000+ packages	Proprietary ecosystems	Ecosystem analysis	Community-driven package development created comprehensive analytical coverage.
Wickham (2014)	Data Science	Methodological framework	ggplot2, dplyr, tidyr	Base R, other tools	Technical framework	Tidy data principles and grammar of graphics revolutionized R usability.

Point to be noted: Quality ratings based on methodological precision, sample size, and clarity of findings. All studies are peer-reviewed journal articles. Foundational studies published before 2010 are included for conceptual context only and were not part of the PRISMA-reviewed study count.

Comparative Discussion and Synthesis

Synthesis Across Themes

From the scholarly reviews that have been conducted, several pieces of evidence of R being employed in academic research relate to technical, economic, cultural, and institutional factors. R's cost, transparency, and reproducibility benefits are closely aligned with the academic value system as well as practical limitations, particularly for researchers at resource-poor institutions. These advantages do have a price to pay, specifically the time to learn and loss in productivity, which varies in importance in different research contexts.

The link between R adoption and reproducibility illustrates these intricate interrelations. R has features that facilitate reproducible workflows; however, the actual reproducibility benefits observed in studies like Marwick (2017) and Stodden *et al.* (2018) likely reflect their technical features and culture in the R community. As such, efforts to improve reproducibility must pay attention to both tools and practices; R is an enabling technology, but it is not the only one in the market.

The empirical findings discussed in this section are interpreted in light of foundational methodological literature where appropriate.

Decision Framework for Tool Selection

Using the available evidence, as far as statistical software are concerned suggest a decision framework for researchers to select statistical software. R proves to be a beneficial choice if long-term reproducibility of results, methodological flexibility, no licensing costs, or special analytical methods coded in R packages are research priorities. It is also preferable to use R if the researchers have enough time to build skills, learn resources, and institutional/community support for R use.

If the research project only requires standard analyses that are available through menu-driven interfaces, and your researchers are not in a position to program and have neither the time nor interest in learning R, and if the institutional infrastructure is well-suited to a particular proprietary tool, or regulations would require validated commercial software, there may be a better option available. Various tools are used by different researchers for their research. They use R for some analyses, reproducible studies, etc. At the same time, using other tools for special tasks.

Integrating with Broader Computational Ecosystems

Over time, R gets increasingly incorporated into larger computational ecosystems rather than being a standalone tool. Through the R package reticulate, researchers can further integrate through Python programming. Similarly, R can integrate with high-performance computing through parallel processing R packages. R can also integrate with a database through DBI and related packages. Moreover, R can also integrate with Version control system such as Git. This allows researchers to benefit from the statistical power of R and use others as required. If you frame the choice of tools as a binary decision (R versus Python; open-source versus proprietary), the ecosystem perspective suggests that you could lose out.

LIMITATION OF EXISTING LITERATURE

Methodological Gaps

Several methodological limitations were found in the reviewed studies, limiting confidence. Initially, the majority of adoption studies depended on self-selected respondents of surveys. This may over-represent enthusiastic R users and under-represent researchers who attempted R adoption but were unsuccessful. This bias in selection can likely lead to a positive inflation or underestimation. Second, comparative case studies often concentrate on the technical capabilities of a tool, rather than on the actual workflows, missing an important aspect of comparative studies. Furthermore, studies replicating research findings, like that of Stodden *et al.* (2018) utilize relatively small samples and may not generalize.

Researchers have not been followed over time as they adopt and incorporate R into their workflows, longitudinally. This study could give us insight into how learning slopes, productivity changes, and reproducibility benefits vary through a career. In addition, only a small number of studies adopt experimental designs that permit causal inference about R on research quality. The bulk of evidence is correlational, which raises questions about whether R adoption leads to better reproducibility or whether reproducibility-oriented researchers adopt R.

Analytical Gaps

Not much has been said about the many important dimensions. Primarily, there are only a few studies that are looking into R adoption in Global South contexts, where cost benefits could have a greater impact, but infrastructure barriers are most uncompromising. Moreover, there is a stark lack of intersectional analyses that look at how R adoption patterns and experiences differ across career stage, gender, race, or institutional prestige. To begin with, the research rarely looks at adoption failures or researchers who tried R and reverted to other tools. This creates a survivorship bias. It lacks economic analyses of the total cost of ownership. Even though the R programming language is free, there are not many analyses related to costs incurred while learning R, or initial reduced productivity, as well as the opportunity cost, among others. More dynamic costing would enable various cost-benefit analyses of profiles as well as institutions.

Limitations of this Review

This review, like others, has its own limitations. Although our narrative synthesis approach is well-suited to the heterogeneous literature, it lacks the quantitative precision of a meta-analysis. There are no estimates of effect size or confidence intervals for most findings. Similarly, although our search strategy was structured, we may not have captured studies in the discipline-specific journals that are not indexed in the main

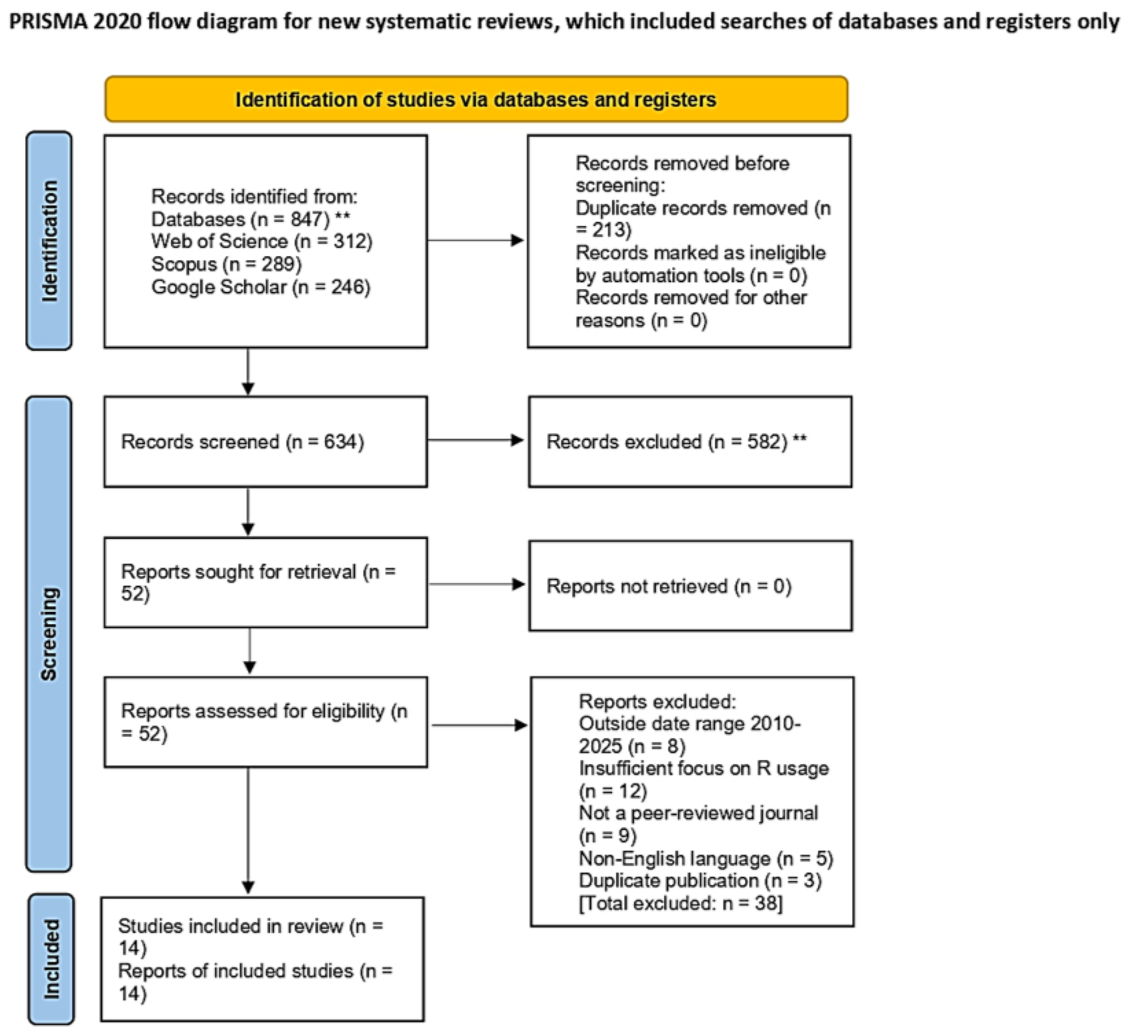

Figure 1: PRISMA Flow Diagram.

Table 2: Quantitative Summary of Key Findings Across Studies.

Metric Category	Specific Measure	Value	Source(s)
Adoption Rates	R usage in ecology (2008)	11.4%	Lai *et al.* (2019)
	R usage in ecology (2017)	58.0%	Lai *et al.* (2019)
	Annual growth rate	5.2% points/year	Lai *et al.* (2019)
Learning Time	Basic proficiency	40-60 hr	Çetinkaya-Rundel and Rundel (2018); Touchon and McCoy (2016)
	Advanced proficiency	200-300 hr	Tippmann (2015)
User Satisfaction	Student satisfaction (R courses)	78% satisfied/very satisfied	Çetinkaya-Rundel and Rundel (2018)
	Researcher satisfaction	71% would recommend R	Lai *et al.* (2019)
Primary Motivations	Free access/zero cost	82% of R users	Lai *et al.* (2019)
	Advanced capabilities	67% of R users	Lai *et al.* (2019)
	Reproducibility	45% of R users	Marwick (2017)
Barriers	Programming requirement	63% cite as a barrier	Lai *et al.* (2019)
	Steep learning curve	58% cite as a barrier	Touchon and McCoy (2016)
	Limited institutional support	41% cite as a barrier	Tippmann (2015)
Reproducibility	Code sharing rate (R users)	62%	Marwick (2017)
	Code sharing rate (proprietary users)	19%	Marwick (2017)
	Replication success (open-source)	39%	Stodden *et al.* (2018)
	Replication success (proprietary)	14%	Stodden *et al.* (2018)
Package Ecosystem	CRAN packages (2010)	~3,957	Hornik (2012)
	CRAN packages (2024)	19,847	Current CRAN statistics
	Annual growth rate	~1,500 packages/year	Hornik (2012); current data
Employment Trends	R job posting growth (2012-2017)	250% increase	(Muenchen, 2011)
	Academic positions requiring R	45% of data-related jobs	(Muenchen, 2011)

Point to be noted: All values sourced from primary reviewed studies. Percentages rounded to the nearest whole number.

database or that used terminologies we did not predict. In the third place, we did not include conference papers, preprints, and grey literature. As a result, we might miss recent developments that are not yet reflected in peer-reviewed publications. The fourteen studies are also heterogeneous in terms of discipline, method, and research question. Results from ecology surveys may not be generalizable to social sciences, and student populations may not be representative of professional researchers. The way we organize themes can give an impression of coherence. In the end, we assessed study quality narratively rather than in a standardized manner based on study methodology. Table 2 does speak to quality, but a more systematic assessment would instill greater confidence in the findings.

FUTURE DIRECTIONS FOR RESEARCH

Research Needs with Higher Priority

There are many areas of research that could greatly enhance knowledge of R's role. Initially, longitudinal studies that track researchers from the adoption of R to the addition of R into established workflows would illuminate how learning curves, productivity, and reproducibility benefits evolve over time. Researchers who stop using R should also be included in these studies to avoid survivorship bias. Another way is to do one of many experimental or quasi-experimental studies that compare users of R to users of an alternative tool, studying quality outcomes. Studying these causal pathways would allow for stronger causal inference about the effect of R on reproducibility and rigor. Researchers and institutions would easily choose software investments if economic analyses provided total cost of ownership estimates based on different researcher and institution profiles.

To conclude, further research into R adoption contexts in the Global South, especially in resource-constrained institutions, would clarify whether low cost translates into uptake given the presence of further infrastructural barriers. Teaching strategies for the instruction of R across disciplines would contribute to evidence-based curriculum development.

Emerging Applications

Research attention is warranted in several newer applications of R. R's statistical integration with machine learning and AI workflows (e.g., tidy models and keras) raises the question of whether R's statistical pedigree has predictive lessons, or are they different? In addition, the increasing employment of R in a variety of institutional and governmental contexts raises issues of the possible limits of R's efficiency of usage in applications with different clienteles and purposes.

Pedagogical Research

There is a considerable gap in understanding R pedagogy that works well. In the future, researchers should investigate how best to sequence R instruction with respect to statistical concepts, the approach to help students with little programming background, the approach to connect R instruction with important discipline content, and the approach to motivating students through the initial difficulties of learning R. Also, looking at how to best support established researchers who are used to other tools in taking up R might inform professional development.

CONCLUSION

The review synthesized evidence from fourteen peer-reviewed studies. Foundational methodological publications were additionally cited to provide conceptual context and were not included in the systematic review study count. The literature considered the adoption patterns of R software, its comparative advantages, and its impact on the quality of the research. Numerous results show strong support across several studies. Over the last ten years, R's adoption in academic research has been growing rapidly, especially due to its zero cost, transparency, and specialized analytics. Further, R is better than the proprietary alternatives, as far as reproducibility, customization, and access to new methods are concerned. However, there are other costs, including time to learn, ease of use, and learning curve. R adoption increases research transparency and replication success, but the mechanisms of causation remain not fully grasped. According to the review, the existing literature is rife with salient methodological and analytical limitations, as well as a limited focus on equity. Applying longitudinal designs, experimental methods, and different contexts will significantly enhance the evidence base of software decision-making for prospective academic research in software engineering. The guidance will help researchers select the software based on evidence. R offers a compelling option when researchers prioritize reproducibility, research flexibility, and cost efficiency, while having sufficient time to learn a new coding system. Research evidence indicates that R instruction, which embeds authentic disciplinary content, is far more effective for educators than offering stand-alone programming courses. There are also realistic expectations about hrs. of learning time (for basic level, about 40-60 hr.) and how students must also be supported through initial frustration, with impacts of motivation and confidence being considered as well. Most importantly, institutions should provide R infrastructure support, which includes training, consultation for statistical computing, and institutional research workflows. Saving significantly from adopting R by removing licensing fees for software can help partially fund services to enable successful adoption. The R package is more than a cheap proprietary solution. Utilizing statistical analyses with reproducible workflows has led to rapid uptake of methodological innovations through community-contributed packages. Furthermore, the normative expectations of transparency and code sharing within R communities may indicate that R adoption would broadly transform research practices. Although this review cannot causally establish that R is causing any of the transformations mentioned, the correlational evidence is compelling and suggests the need for future studies as open science practices continue to evolve.

ABBREVIATIONS

AI: Artificial Intelligence; **CASP:** Critical Appraisal Skills Programme; **CRAN:** Comprehensive R Archive Network; **DBI:** Database Interface; **GNU:** GNU's Not Unix; **MMAT:** Mixed Methods Appraisal Tool; **PRISMA:** Preferred Reporting Items for Systematic Reviews and Meta-Analyses; **SAS:** Statistical Analysis System; **SPSS:** Statistical Package for the Social Sciences.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

AUTHOR CONTRIBUTION IN THIS RESEARCH PAPER

Snehasish Paul conceived the idea of the paper, prepared the initial draft, and carried out a thorough review of the manuscript. He made substantive corrections and ensured the overall coherence of the work. Rohit Kumar focused on proofreading the manuscript, checking for typographical errors, and conducting plagiarism screening to ensure originality and compliance with academic standards. Sudhanshu Sharma conducted the literature review, performed grammar checks, and was responsible for extracting and organizing literature sources based on inclusion and exclusion criteria.

REFERENCES

Çetinkaya-Rundel, M., and Rundel, C. (2018). Infrastructure and Tools for Teaching Computing Throughout the Statistical Curriculum. *The American Statistician*, *72*(1), 58-65. DOI: 10.1080/00031305.2017.1397549

Gentleman, R., and Temple Lang, D. (2007). Statistical Analyses and Reproducible Research. *Journal of Computational and Graphical Statistics*, *16*(1), 1-23. DOI: 10.1198/106186007X178663

Hornik, K. (2012). The Comprehensive R Archive Network. *WIREs Computational Statistics*, *4*(4), 394-398. DOI: 10.1002/wics.1212

Ihaka, R., and Gentleman, R. (1996). R: A Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics*, *5*(3), 299-314. DOI: 10.1080/10618600.1996.10474713

Koenker, R., and Zeileis, A. (2009). On reproducible econometric research. *Journal of Applied Econometrics*, *24*(5), 833-847. DOI: 10.1002/jae.1083

Lai, J., Lortie, C. J., Muenchen, R. A., Yang, J., and Ma, K. (2019). Evaluating the popularity of R in ecology. *Ecosphere*, *10*(1). DOI: 10.1002/ecs2.2567

Lerner, J., and Tirole, J. (2002). Some Simple Economics of Open Source. *The Journal of Industrial Economics*, *50*(2), 197-234. DOI: 10.1111/1467-6451.00174

Marwick, B. (2017). Computational Reproducibility in Archaeological Research: Basic Principles and a Case Study of their Implementation. *Journal of Archaeological Method and Theory*, *24*(2), 424-450. DOI: 10.1007/s10816-015-9272-9

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, n71. DOI: 10.1136/bmj.n71

Peng, R. D. (2011). Reproducible Research in Computational Science. *Science*, *334*(6060), 1226-1227. DOI: 10.1126/science.1213847

Raymond, E. S. (2001). *The cathedral and the bazaar: musings on Linux and Open Source by an accidental revolutionary*. O'Reilly.

Robert A. Muenchen. (2011). *The Popularity of Data Science Software*. Data Science Software Reviews.

Stodden, V., Seiler, J., and Ma, Z. (2018). An empirical analysis of journal policy effectiveness for computational reproducibility. *Proceedings of the National Academy of Sciences*, *115*(11), 2584-2589. DOI: 10.1073/pnas.1708290115

Tippmann, S. (2015). Programming tools: Adventures with R. *Nature*, *517*(7532), 109-110. DOI: 10.1038/517109a

Touchon, J. C., and McCoy, M. W. (2016). The mismatch between current statistical practice and doctoral training in ecology. *Ecosphere*, *7*(8). DOI: 10.1002/ecs2.1394

Wickham, H. (2014). Tidy Data. *Journal of Statistical Software*, *59*(10). DOI: 10.18637/jss.v059.i10

Wilson, G., Aruliah, D. A., Brown, C. T., Chue Hong, N. P., Davis, M., Guy, R. T., Haddock, S. H. D., Huff, K. D., Mitchell, I. M., Plumbley, M. D., Waugh, B., White, E. P., and Wilson, P. (2014). Best Practices for Scientific Computing. *PLoS Biology*, *12*(1), e1001745. DOI: 10.1371/journal.pbio.1001745

Yihui Xie, J. J. Allaire, and Garrett Grolemund. (2018). *R Markdown: The Definitive Guide*. Chapman and Hall/CRC.

Cite this article: Paul S, Kumar R, Sharma S. R for Data Analysis in Academic Research: An Open-Source Perspective. Journal of Data Science, Informetrics, and Citation Studies. 2026;5(2):216-26.