Public Availability of Data Management and Analysis Scripts of Studies Conducted on Open-Access Intensive Care Databases

Coline Andries, Paul Quindroit, Benjamin Popoff, Ikram Oudarrour, Antoine Lamer

Digital Health and Informatics Innovations for Sustainable Health Care Systems
J. Mantas et al. (Eds.)

doi:10.3233/SHTI240429

Public Availability of Data Management and Analysis Scripts of Studies Conducted on Open-Access Intensive Care Databases

Coline ANDRIES[a,b,1], Paul QUINDROIT[c], Benjamin POPOFF[d,e], Ikram OUDARROUR[a,b] and Antoine LAMER[a,b,c,e]
[a]*Univ. Lille, UFR3S - Ingénierie et Management de la Santé, 59000 Lille, France*
[b]*Fédération régionale de recherche en psychiatrie et santé mentale (F2RSM Psy), Hauts-de-France, Saint-André-Lez-Lille, France*
[c]*Univ. Lille, CHU Lille, ULR 2694-METRICS: Évaluation des Technologies de santé et des Pratiques médicales, Lille, France*
[d]*CHU Rouen, Department of Anesthesiology and Critical Care, F-76000 Rouen, France*
[e]*InterHop, Saint-Marcan, France*
ORCiD ID: Coline ANDRIES https://orcid.org/0009-0001-0090-6262

Abstract. Intensive care units (ICUs) provide care for critical patients at high risk of morbidity and mortality, and require continuous monitoring of clinical, biological and, imaging parameters. Collaborative ventures have enabled the emergence of large open access databases for the secondary use of Electronic Health Records (EHRs). The objective of this work was to evaluate the availability of scripts and datasets in publications based on ICU open-access databases. We included 910 original articles based on four ICU open-access databases (Amsterdam University Medical Centers Database, eICU Collaborative Research Database, High time resolution ICU dataset, and Medical Information Mart for Intensive Care). The majority of the studies did not provide their data management scripts (n=839, 92.9%), neither the analysis script (n=843, 93.4%) in the article. Attempts to contact the 845 corresponding authors in question resulted in 89.11% (n=753) of our e-mail requests going unanswered over a two-month period. We received 51 automated messages (55.43%) indicating that emails have not been delivered, while 6 messages (6.52%) redirected to alternative email addresses. Only 20 corresponding authors (18.18%) answered, finally providing the requested materials. Despite scientific journals recommendations to share materials, our study unveils the absence of crucial components for the replication of studies by other research teams.

Keywords. Data reuse, Reproducibility, MIMIC, Intensive care, Open-access

1. Introduction

A key element of reproducible research is the availability of sufficient information to allow other researchers to replicate or extend published results. In addition to new analyses and dissemination, this would have beneficial effects on science and, in turn, on

[1] Corresponding Author: Coline Andries, 211 Rue du Général Leclerc, 59350 Saint-André-lez-Lille, France; E-mail: coline.andries@protonmail.com.

patient care. For data-enabled research, recommendations from funding institutions and journals have become common [1]. Achieving this requires the thorough documentation of data sources, provision of the data itself, the code generating the results, and any supplementary information necessary for understanding the research [2]. In 2016, the International Committee of Medical Journal Editors (ICMJE) also required that publications of a clinical trial contain a data sharing statement [3]. These statements must document whether deidentified data and what type of data will be shared; whether additional documents will be available (study protocol, statistical analysis plan, etc); when the data will become available and for how long; by what access criteria data will be shared (including with whom, for what types of analyses, and by what mechanism).

Intensive care units (ICUs) provide care for critical patients at high risk of morbidity and mortality. These patients, due to their severity, require continuous monitoring and surveillance of clinical, biological and imaging parameters. This generates a large amount of data providing opportunities for data reuse and research. Since the 2000s, open-access databases on critical care have emerged [4]. The best-known example is the Medical Information Mart for Intensive Care (MIMIC) database, which integrates anonymized, comprehensive clinical data from more than 50,000 intensive care admissions from Beth Israel Deaconess Medical Center in Boston, Massachusetts [5]. These open databases have led to the production of numerous research works.

The objective of this work was to assess the reproducibility of published studies conducted using open-access ICU databases, by examining the availability of data management and analysis scripts within the articles.

2. Methods

We included, the original articles based on 4 intensive care open databases: Amsterdam University Medical Centers Database (AmsterdamUMC), eICU Collaborative Research Database (eICUCRD), High time resolution ICU dataset (HiRID), Medical Information Mart for Intensive Care (MIMIC)). The literature research was carried out from 3 databases (PubMed – Medline, Embase, Web of science) from the creation of these databases to August 1st, 2022. The queries were defined as part of a scoping review about open-access databases in intensive care [6].

Articles were excluded if they were only available in abstract form, not freely accessible, duplicates, not written in English, or not original research. If an article met the inclusion criteria, the next step involved determining the specific database mentioned within its Materials section. We also sought information on the version of the database utilized. To verify this information, we thoroughly examined the database's reference section, the study's main body, and any data availability statements provided, in search of either a version number or a link directing to the specific database version used. Furthermore, we extracted information regarding the presence of both the data management and the analysis scripts. We identified the presence of these scripts if they were included in the Supplementary Materials, their linkage to a git repository, or their accessibility on a team's website. Conversely, if the scripts were provided upon request or were not mentioned without data availability statement on demand, we noted their non-availability. Last, we collected data on the country of origin of the first author and retrieved the email address of the corresponding author.

We contacted corresponding authors via email starting from February 5, 2024. We specified that our aim is to reproduce and assess the reproducibility of various studies.

We asked if it would be possible to share the dataset used after applying inclusion criteria and data management procedures. Furthermore, we requested the authors to provide the scripts used for both generating these datasets, as well as the scripts used to compute the final statistical analyses of the study.

Responses were categorized into several categories: "In waiting" for ongoing discussions, "Non-delivered" for emails that bounced back due to non-existent email addresses, "Reply with materials" for authors who replied with the requested materials, "Reply without materials" for responses lacking the requested materials (for instance, if the person with the script was unreachable or no longer affiliated), "Automatic e-mail with redirection to a new address" for automated responses redirecting to another email address, "Non-applicable" for errors occurred during the screening process, and "Error" for situations were materials were initially overlooked but later found to be included in the article.

All material used for this study is available on our git repository [7].

3. Results

3.1. Study selection and characteristics of studies

We identified 1,474 articles through database searching. After assessing them against the inclusion and exclusion criteria, 571 articles were excluded and leaving us with 903 articles. The primary causes for exclusion included works in abstract format only (n = 336), free access to the full article (n = 229), duplicate retrievals (n = 4), non-original research article (n = 9), or non-English language (n = 1).

Among the included articles, 630 utilized the MIMIC-III database (64%), 154 employed MIMIC-IV (15.6%), 135 utilized eICU-CRD (13.7%), 60 relied on MIMIC-II (6.09%), and 6 utilized AmsterdamUMC (0.61%). The studies included in the analysis originated from a diverse range of countries. China stood out as the most prevalent contributor, with 600 articles (66.45%), followed by the United States with 165 articles (18.27%), and Canada with 15 studies (1.66%). A significant number of studies also originated from Germany (1.11%, n = 10), the United Kingdom (1.33%, n = 12), and various other countries such as Australia, Italy, Spain, and Taiwan, each contributing around 0.78% (n = 7) of the studies or less.

3.2. Data and script availability

The majority of the studies did not make their data management script available (n=839, 92.9%), with 672 studies (74.4%) lacking the script entirely and 167 articles (18.5%) proposing to deliver the scripts on demand. A small percentage, 7.09% (n=64), did offer a data management script via a Git repository, a link to a team's website (6.09%, n=55), or 1% (n=9) as supplementary material.

In terms of statistical analysis scripts, a substantial majority, 93.4% (n=843), lacked accessibility, with 74.9% (n=676) being totally absent and 18.5% (n=167) being available on demand. A minor proportion (6.64, n=60), 5.98% (n=54), offered their statistical analysis script via a Git or a link to a team's website, or 0.66% (n=6) as supplementary material.

Out of the total studies analyzed, 435 studies (48.17%), did not furnish scripts for data management, data analysis, or database version information. Conversely, 407

studies (45.07%) provided either one of the aforementioned components. Only 39 studies (4.32%), provided with two components, and 22 studies (2.44%) included of all three components (Figure 1).

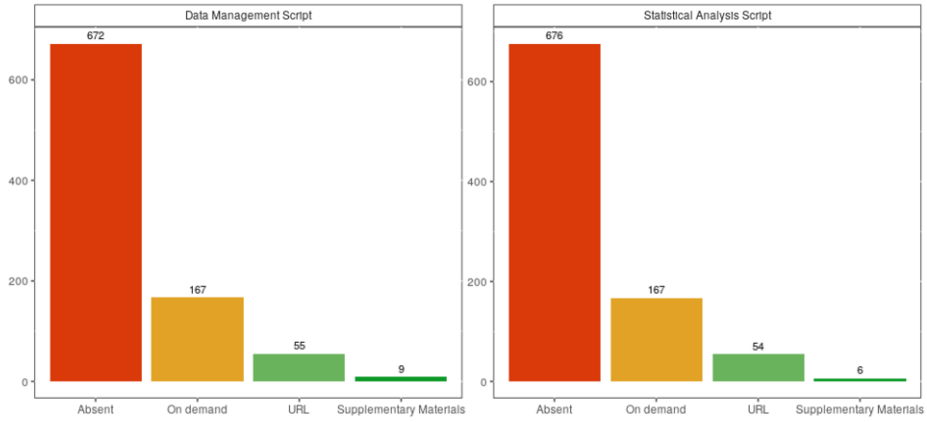

Figure 1. Availability of data management and statistical analysis scripts

3.3. Reply rates to emails

Of the 845 which did not share neither the data management script nor analysis script, the majority did not answer to our e-mail requests after a span of two months (89.11%, n=753). Amongst all replies (n=92), we received 51 automated messages (55.43%) indicating that emails have not been delivered, while 6 messages (6.52%) redirected to alternative email addresses. Only 14 corresponding authors (13.04%) answered by providing the requested materials. Furthermore, 11 conversations (12.20%) commenced to date, typically involving looping another team member for material provision or requesting further details. Ten answers (10.87%) cited reasons such as team member turnover or the study being outdated, hindering material provision.

4. Discussion

Despite scientific journals recommendations to share materials, our study highlights a significant gap in the availability of essential resources required for other research teams to replicate studies. Although the corresponding authors were contacted to obtain the missing material, a large proportion did not respond, even after a period of two months.

Of particular note, 50 responses received were generated automatically, highlighting issues such as invalid email addresses. This observation underscores the challenges posed by outdated or inaccurate contact information, hindering effective communication and collaboration among researchers. Moreover, some authors provided incorrect email addresses (consisting solely of numbers), demonstrating that the publishers did not verify them. The frequent occurrence of automated responses, often associated with student or professional email addresses, highlights the need for improved strategies to ensure the long-term availability of accurate contact details within the scientific community.

Our results are consistent with similar studies [8]. It has already been reported that not all journals explicitly request data sharing, and it is more often the journals with large

impact factors that do so [9]. However, there are still initiatives in place to promote research reproducibility [10].

5. Conclusions

Despite journal recommendations to share materials, our study reveals a notable lack of essential resources necessary for other research teams to replicate studies. Establishing standardized practices for data sharing and communication is essential to enhance scientific integrity and promote the reliability of research findings. Efforts to ensure the long-term availability of accurate contact details require a collective commitment. Thus, it is the responsibility of editors to ensure the availability of scripts and the veracity of the corresponding author's e-mail address, before the article final publication.

References

[1] Hrynaszkiewicz I, Altman DG. Towards agreement on best practice for publishing raw clinical trial data. Trials. 2009 Mar 18;10:17.
[2] Hanson B, Sugden A, Alberts B. Making Data Maximally Available. Science. 2011 Feb 11;331(6018):649–649.
[3] Taichman DB, Sahni P, Pinborg A, Peiperl L, Laine C, James A, et al. Data sharing statements for clinical trials. BMJ. 2017 Jun 5;357:j2372.
[4] Sauer CM, Dam TA, Celi LA, Faltys M, de la Hoz MAA, Adhikari L, et al. Systematic Review and Comparison of Publicly Available ICU Data Sets-A Decision Guide for Clinicians and Data Scientists. Crit Care Med. 2022 Jun 1;50(6):e581–8.
[5] Popoff B. Contribution of open access databases to intensive care medicine research: a scoping review. 2022 Jul 18 [cited 2024 Apr 1]; Available from: https://osf.io/kugaz/
[6] Johnson AEW, Pollard TJ, Shen L, Lehman L wei H, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data. 2016 May 24;3(1):1–9.
[7] GitLab [Internet]. 2024 [cited 2024 Jun 6]. Public Availability of Data Management and Analysis Scripts of Studies Conducted on Open-Access Intensive Care Databases. Available from: https://gitlab.com/coline_andries/mie_datareuse
[8] Rowhani-Farid A, Barnett AG. Has open data arrived at the British Medical Journal (BMJ)? An observational study. BMJ Open. 2016 Oct 13;6(10):e011784.
[9] Stodden V, Guo P, Ma Z. Toward Reproducible Computational Research: An Empirical Analysis of Data and Code Policy Adoption by Journals. PLoS One. 2013;8(6):e67111.
[10] Kretser A, Murphy D, Bertuzzi S, Abraham T, Allison DB, Boor KJ, et al. Scientific Integrity Principles and Best Practices: Recommendations from a Scientific Integrity Consortium. Sci Eng Ethics. 2019;25(2):327–55.