Medicina Intensiva 49 (2025) 502142

medicina intensiva

http://www.medintensiva.org/en/

REVIEW ARTICLE

Quality of causality assessment among observational studies in intensive care: A methodological review

Laura Del Campo-Albendea[a,b,c,*], **Ana García De La Santa Viñuela**[d], **Óscar Peñuelas**[e], **José Ignacio Pijoan Zubizarreta**[b,f], **Khalid Saeed Khan**[b,g], **Alfonso Muriel**[a,b,h,◊], **Javier Zamora**[a,b,i,◊]

[a] *Clinical Biostatistics Unit, Hospital Universitario Ramón y Cajal, IRYCIS, Madrid, Spain*
[b] *CIBER of Epidemiology and Public Health, CIBERESP, Madrid, Spain*
[c] *Escuela de Doctorado UAM, Universidad Autónoma de Madrid, Madrid, Spain*
[d] *Preventive Medicine and Public Health Service, Hospital Universitario Ramón y Cajal, IRYCIS, Madrid, Spain*
[e] *Intensive Care Unit, University Hospital of Getafe, Biomedical Research Networking Center for Respiratory Diseases, CIBERES, Madrid, Spain*
[f] *Department of Epidemiology, Cruces University Hospital, Biocruces Bizkaia Health Research Institute, Barakaldo, Spain*
[g] *Department of Preventive Medicine and Public Health, University of Granada, Granada, Spain*
[h] *Nursing and Physiotherapy Department, University of Alcalá, Madrid, Spain*
[i] *Institute of Metabolism and Systems Research, University of Birmingham, Birmingham, United Kingdom*

Received 8 April 2024; accepted 6 November 2024
Available online 5 February 2025

KEYWORDS
Causality;
Methodological review;
Critical care;
Observational studies

Abstract Intensive care units (ICUs) rely in many instances on observational research and often encounter difficulties in establishing cause-and-effect relationships. After conducting a thorough search focused on ICU observational studies, this review analysed the causal language and evaluated the quality of reporting of the methodologies employed. The causal was assessed by analysing the words linking exposure to outcomes in the title and main objective. The quality of the reporting of the key methodological aspects related to causal inference was based on STROBE and ROBINS-I tools. We identified 139 articles, with 87 (63%) and 82 (59%) studies having non-causal language in their title and main objective, respectively. Among the total, 49 (35%) articles directly addressed causality. The review found vague causal language in observational ICU research and highlighted the need for better adherence to reporting guidelines for improved causal analysis and inference.

DOI of original article: https://doi.org/10.1016/j.medin.2025.502142
Abbreviations: ICU, intensive care units; RCT, randomized controlled trials; DAG, directed acyclic graphs; ITS, interrupted time series; IV, instrumental variables; MSM, marginal structural models (MSM); CCU, critical care unit; ATE, Average Treatment Effect; ATT, Average Treatment effect for the Treated; ATU, Average Treatment effect for the Untreated.

* Corresponding author.
E-mail address: lcampo@salud.madrid.org (L. Del Campo-Albendea).
◊ Co-senior authors of the article.

https://doi.org/10.1016/j.medine.2025.502142

PALABRAS CLAVE
Causalidad;
Revisión metodológica;
Cuidados críticos;
Estudios observacionales

Calidad de la evaluación de la causalidad en estudios observacionales en cuidados intensivos: una revisión metodológica

Resumen Las unidades de cuidados intensivos (UCI) dependen en muchas ocasiones de la investigación observacional y, a menudo, encuentran dificultades para establecer relaciones causales. Después de realizar una búsqueda exhaustiva de estudios observacionales en UCI, se analizó el lenguaje causal mediante el análisis de las palabras que vinculan la exposición a los resultados en el título y objetivo principal. La calidad del reporte de los aspectos metodológicos claves relacionados con la inferencia causal se evaluó utilizando las herramientas STROBE y ROBINS-I. Identificamos 139 artículos, con 87 (63%) y 82 (59%) estudios que usaban lenguaje no causal en su título y objetivo principal, respectivamente. De estos, 49 (35%) artículos abordaron directamente causalidad. La revisión encontró un uso vago de lenguaje causal en la investigación observacional en UCI y resaltó la necesidad de mejorar la adherencia a las guías de reporte para mejorar la investigación causal.

Introduction

A fundamental challenge in health research lies in establishing whether there is a genuine cause-and-effect relationship between exposure and outcome.[1] Being stochastic in nature, health research is inexact. Causal inference is best addressed through randomized controlled trials (RCTs), which by virtue of random assignment enable comparisons of groups similar for prognosis at baseline. However, many circumstances render RCTs impractical or unethical.[2,3] For example, in the context of intensive care unit (ICU) research, the design and conduct of RCTs becomes challenging due to ethical considerations such as the difficulty of withholding life-saving interventions,[4] eligibility restrictions due to patient heterogeneity[5,6] and challenges in obtaining informed consent from the critically ill.[7,8]

In ICU research, observational studies provide a feasible alternative to RCTs. In the past decade, there has been a notable shift in observational studies, with increased emphasis on innovative designs and statistical analyses.[9–13] Key modern causal methods include directed acyclic graphs (DAGs),[14,15] propensity score methods,[16,17] inverse probability treatment weighting,[18,19] G-methods,[20,21] interrupted time series (ITS),[22] instrumental variables (IV),[23] and marginal structural models (MSM),[24] amongst others. When RCTs have been replicated using observational data with rigorous methodology, the results have been similar.[2,25–27]

Given the recent improvements in designs and statistics, one might expect greater precision in article language avoiding ambiguous and falsely positive causal inferences. It has been claimed in the past that much of the causal association literature has avoided direct language. It has oscillated between excess caution and exaggeration in suggesting a cause-and-effect relationship.[28] The extent to which there has been ambiguity has not been quantified. The objective of this review was to quantify the utilization of correct causal language in recent observational studies in ICU settings. We also assessed the quality of reporting and the methods employed to address causal relationships.

Methods

This methodological study was prospectively registered at Open Science Forum Registries (DOI: https://doi.org/10.17605/OSF.IO/NZRVT) and is reported following PRISMA 2020 guidelines[29] (checklist provided in e-Table 1).

Search and article selection

Our search focused on observational studies evaluating any procedure in ICU setting, published in peer-reviewed journals indexed in the Ovid Medline database between 2019 and 2022. Our search included Medical Subject Heading (MeSH) terms and keywords for causality adapted OVID-Medline (see e-Table 2). We restricted our search to critical care settings using the following terms ''critical care unit, intensive care unit, critical care facility, intensive treatment unit, emergency unit, critical room, and ICU or CCU''. Language was restricted to English. All reports of observational studies that included the term ''causal'' were eligible for inclusion. We excluded research involving cellular or animal models, as well as all types of non-observational studies. Search results were organised using the Rayyan web application for systematic review management. Two reviewers (LdC, AM) identified potentially eligible articles based on their title and abstract. After the reviewers piloted 200 articles, the observed agreement was greater than 90% and the selection was made by a single reviewer (LdC). The same reviewer selected articles based on the full text. We also conducted a manual search scrutinizing the articles published in the same period in all the 35 journals indexed in the category ''Critical Care Medicine'' of the Journal Citation Reports (JCR).

Data extraction

Data extraction was performed using a standardized pre-piloted form. We extracted verbatim the sentence con-

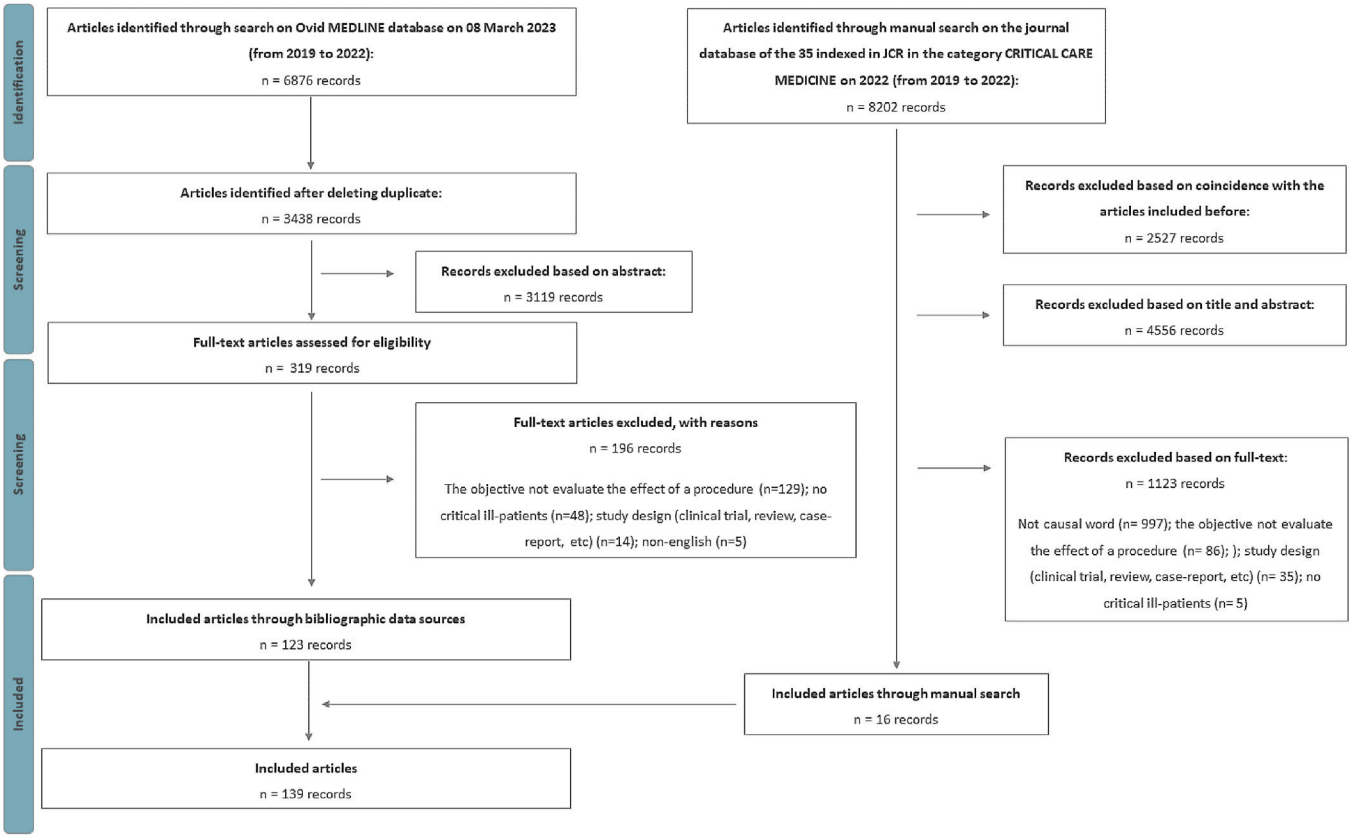

Figure 1 Flow chart of the article selection process for the methodological review of causality assessment among observational studies in intensive care.

taining the word ''causal'' to analyse the causal intention of the authors. The articles were divided into two groups based on the use of this term. One group consisted of articles where the authors used the word ''causal'' or synonyms to indicate an intention to address causality (see e-Box 1) directly. The other group included articles acknowledging that the term ''causal'' could not be appropriately used due to issues related to study design, statistical analysis, etc.

For causal language analysis, we extracted the words linking the exposures to the outcomes from the titles and main objective of the study. We then categorized these linking words according to the definitions of causal language provided by Thapa et al.[30] (see e-Box 2).

For the quality of causality assessment, we extracted data with respect to reporting and methodology. We extracted data on reporting of the subset of key methods, results and discussion items related to causality in the STROBE checklist.[31] To assess the methodological quality of causal inference, we used the relevant items of the ROBINS-I tool,[32] focusing on the dimensions confounding, selection bias, bias due to missing data and bias in the classification of interventions and in the measurement of outcomes. The data extraction form created by combining STROBE and ROBINS-I is presented in e-Table 3. We also extracted information regarding limitations provided in narrative form in the discussion sections of the manuscripts.

Data synthesis

We calculated the percentage for frequency data and mean, standard deviation and range for continuous data for each characteristic of interest. We present the results along with their corresponding 95% confidence intervals. Graphing and statistical analysis were conducted using R version 4.3.1 and Stata 18, respectively. We assessed the accuracy of language usage in title and abstract by comparing articles that directly assessed causality with those that did not, employing a chi-squared test for analysis. In order to assess the quality of reporting and methodology, we only analysed data from studies that directly assessed causality. We focused on these articles because the other group of articles did not explicitly address causal inferences.

Results

Search and article selection

Fig. 1 summarizes the search and the selection processes. Among the 6876 records retrieved, 319 were selected for full-text assessment and 123 articles were included. Based on the manual search, 16 articles were further included. Altogether, data extraction was conducted in 139 articles.

Table 1 Characteristics of included articles the methodological review of causality assessment among observational studies in intensive care.

	N = 139 (%)	[95%CI]
JIF quartile		
Q1	67 (48%)	[39%; 57%]
Q2	34 (25%)	[17%; 32%]
Q3	20 (15%)	[9%; 21%]
Q4	9 (6%)	[3%; 12%]
No indexed	9 (6%)	[3%; 12%]
Median journal impact factor [min; max]	6.9 [0.2; 39.2]	[5.8; 8.0]
Journals with restrict language policy[a]	27 (19%)	[13%; 27%]
World region of corresponding authors		
Northern America	76 (55%)	[46%; 63%]
Europe	36 (26%)	[19%; 34%]
Eastern Asia	20 (14%)	[9%; 21%]
Oceania	5 (4%)	[1%; 8%]
Western Asia	2 (1%)	[0.1%; 5%]
Statistician/epidemiologist in the author list	46 (33%)	[25%; 42%]
Using a reporting guideline	27 (19%)	[13%; 27%]
Study funding		
No external funding	73 (53%)	[44%; 61%]
Non-industry funded	58 (42%)	[33%; 50%]
Industry-funded	6 (4%)	[2%; 9%]
Not clearly stated	2 (1%)	[0.1%; 5%]
Conflict of interest		
No	86 (62%)	[53%; 70%]
Yes	38 (27%)	[20%; 36%]
Not reported in the article	13 (9%)	[5%; 15%]
Not clearly stated	2 (1%)	[0.1%; 5%]
Statistical software		
R	36 (26%)	[19%; 34%]
SPSS	24 (17%)	[11%; 24%]
Stata	22 (16%)	[10%; 23%]
More than one	20 (14%)	[9%; 21%]
SAS	19 (14%)	[8%; 21%]
Not reported	10 (7%)	[4%; 13%]
Other	8 (6%)	[3%; 11%]

[a] Journals that indicate in their guidance to authors to restrict the use of causal language only to randomized controlled trials.

Most of the studies originated from the United States (55.0%) and Europe (25.7%) and were published in first quartile (Q1) category journals (48% of the total). The median impact factor of the journals was 6.9. In 46 articles (33%), the authors included a statistician/epidemiologist and in 27 articles (19%) they used a reporting guideline. More characteristics of the included articles can be found in Table 1.

Analysis of causal language

We identified 49 (35%) articles as directly addressing causality and 90 (65%) non-causal articles. The words linking the exposure to the outcome in the title and the main objective were non-causal in 27 (55%) and 28 (57%) studies in the directly causal articles, respectively. In the case of non-causal articles, non-causal language was used in the titles and objectives of approximately 60 articles (67%) and 54 articles (60%), respectively. The term ''association'' was used in the title in 11 (22%) studies in the group directly addressing causality and in 24 (27%) studies in the group not addressing it (p-value = 0.684). The term ''effect'' and its synonyms were used in 12 (24%) studies in the group directly addressing causality and in 11 (12%) in the group not addressing it (p-value = 0.093). We found a similar frequency of the authors' use of the words linking exposure with outcome in the studies' main objectives for both groups; the term ''association'' was used in 17 (35%) studies in the group directly addressing causality and in 34 (38%) studies in the group not addressing it (p-value = 0.854). The term ''effect'' and its synonyms were used in 10 (20%) studies in the group directly addressing causality and in 19 (21%) studies in the group not addressing it (p-value = 0.922) (Fig. 2).

Out of the 49 articles directly addressing causality, 31 articles (65%) had statistically significant results. Among these 31 articles, 17 (55%) used accurate causal language in their titles, and 13 (42%) did so in their main objectives. Of the 18 articles with non-significant results, 12 articles (71%) used non-causal language in the title, and 10 (59%) did so in the main objectives. In one article of the group

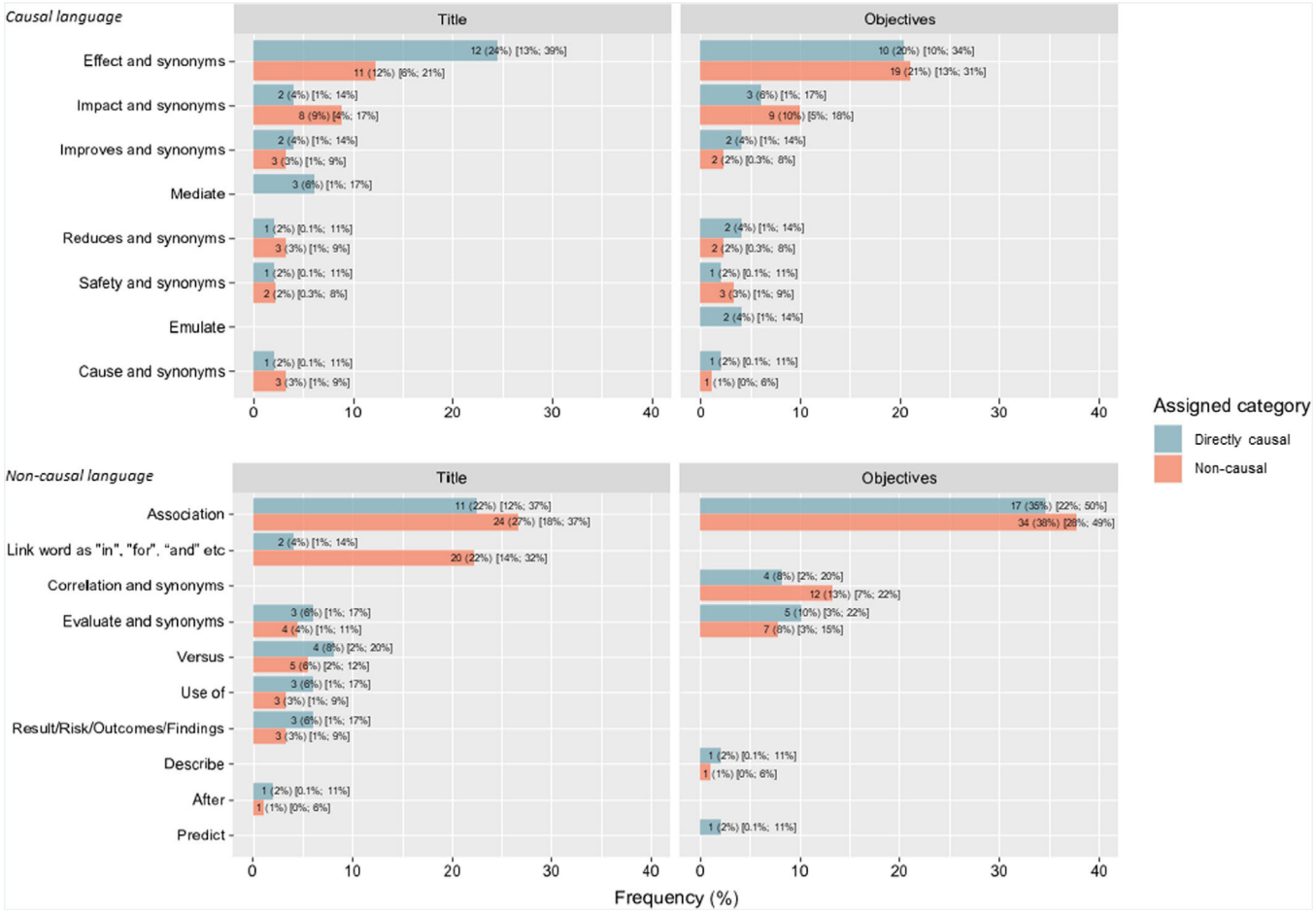

Figure 2 Frequency of linking words used in the title and objective sections of the articles included in the methodological review of causality assessment among observational studies in intensive care.

Table 2 Frequency of words used to link exposure with outcome according to the statistical significance in title and objective sections of the articles included in the methodological review of causality assessment among observational studies in intensive care.

	Title			Objective		
Directly addressing causality	Statistically significant N = 31	No statistically significant N = 17	p-Value	Statistically significant N = 31	No statistically significant N = 17	p-Value
Causal word, n (%) [95%CI]	17 (55%) [36%; 72%]	5 (29%) [10%; 56%]	0.091	13 (42%) [24%; 61%]	7 (41%) [18%; 67%]	0.959
No causal word, n (%) [95%CI]	14 (45%) [27%; 64%]	12 (71%) [44%; 90%]		18 (58%) [39%; 75%]	10 (59%) [33%; 82%]	
Non-causal	Statistically significant N = 60	No statistically significant N = 23	p-Value	Statistically significant N = 60	No statistically significant N = 23	p-Value
Causal word, n (%) [95%CI]	18 (30%) [19%; 43%]	11 (48%) [27%; 69%]	0.127	22 (37%) [25%; 50%]	12 (52%) [31%; 73%]	0.199
No causal word, n (%) [95%CI]	42 (70%) [57%; 81%]	12 (52%) [31%; 73%]		38 (63%) [50%; 75%]	11 (48%) [27%; 69%]	

directly addressing causality, the authors did not present results for the main objective. We found a similar frequency among studies that did not focus on causality. Of the 90 articles, 60 (67%) reported statistically significant findings. In 7, researchers did not provide information on their main outcome. In studies with statistically significant results, authors used causal terminology in the titles of 18 articles (30%) and in the main objectives of 22 articles (37%). In the articles

Table 3 Key causality items according to the instrument developed in the methodological review of causality assessment among observational studies in intensive care.

	True causal N = 49 (%)	[95%CI]
Effort to address potential sources of bias described in methods		
Acknowledge confounding	42 (86%)	[73%; 94%]
Acknowledge unmeasured confounding	8 (16%)	[7%; 30%]
Missing data reported	27 (55%)	[40%; 69%]
Assumptions made	6 (12%)	[5%; 25%]
Bias in classification of interventions and outcomes	27 (55%)	[40%; 69%]
Evaluated reliability	6 (12%)	[5%; 25%]
Statistical methods description		
Selection of covariates	30 (61%)	[46%; 75%]
Based on prior knowledge	13 (26%)	[15%; 41%]
Included a DAG	7 (14%)	[6%; 27%]
p-Value-based	6 (12%)	[5%; 25%]
Alternative approaches	4 (8%)	[2%; 20%]
Adjustment methodology	44 (90%)	[78%; 97%]
Regression adjustment	18 (37%)	[23%; 52%]
Inverse probability of treatment weighting	13 (26%)	[15%; 41%]
Propensity score-based methods	11 (22%)	[12%; 37%]
Other	2 (4%)	[1%; 14%]
Reporting of the numbers of individuals at each stage of study		
Patient's flow-chart	22 (45%)	[31%; 60%]
Give reasons for non-participation at each stage	40 (82%)	[68%; 91%]
Reporting of the characteristics of study participants		
Baseline characteristics table	41 (84%)	[70%; 93%]
Quantification of the sample comparability	34 (69%)	[55%; 82%]
Quantification of the sample comparability after adjustment to control confusion	16 (33%)	[20%; 48%]
Reporting of other analysis done: sensitivity analysis		
Robustness checks with sensitivity analyses	17 (35%)	[22%; 50%]
Reporting of the limitations of the study[a]		
Study design	39 (80%)	[66%; 90%]
Unmeasured confounder	38 (77%)	[63%; 88%]
Data quality	36 (73%)	[59%; 85%]
Short follow-up or limited data collected	26 (53%)	[38%; 67%]
No generalizability	20 (41%)	[27%; 56%]

[a] Percentages could be greater than 100% as the same study could have referred more than one limitation.

with a non-significant result, causal terminology appeared in 11 (48%) of the titles and 12 (52%) of the main objectives. More information is provided in Table 2.

Reporting and quality in the group of studies directly addressing causality

The group directly addressing causality included retrospective cohorts (32 articles, 65%) as well as prospective cohorts (15 articles, 31%). Two articles (4%) were classified as mixed because they included both retrospective and prospective data collection. Among the 49 articles, five studies (10%) were designed as a ''target trial emulation'' by the manuscript authors. Twenty-five (51%) of the studies employed administrative data while 24 (49%) used data collected for research purpose. The terms ''real-world data'' or ''real-world evidence'' were used in 7 (14%) of the studies.

In total, 42 articles (86%) acknowledged confounding as a potential bias and addressed it in the statistical methods section (Table 3). In eight (16%) the researchers explored the unmeasured confounding issue. The presence of missing data was treated by authors in 27 studies (55%). Six articles (12%) made assumptions about the type of missing data before dealing with it. Multiple imputation was used to correct this bias in 9 articles (18%), complete cases were used in 7 articles (14%) and 5 articles (10%) used other types of approaches. In 6 articles (12%), the authors did not specify their approach. The selection of covariates was reported in 30 articles (61%), based on prior knowledge in 13 articles (43%) and relying on p-value-based decisions in 6 articles (12%). The construction of a multivariable regression model to address confounding was used in 18 articles (37%). Modern causal analysis such as propensity score-based methods or inverse probability of treatment weighting methods were employed in 26 articles (53%).

Regarding the results section, 40 articles (82%) gave reasons for excluding patients and 22 articles (45%) pro-

vided a flow chart depicting the number of patients in their respective studies. The table of baseline characteristics was presented in 41 articles (84%), of which 34 studies (69%) reported a quantification measure of similarity between the groups compared. The p-value was used as a pre-adjustment measure of comparability in 23 articles (48%) and the standardized differences in 10 articles (29%). In one article (3%), the authors used a risk ratio as a comparability measure. In 16 articles (33%) the authors reported a post-fitting comparability measure.

Regarding reporting of study limitations, in 39 articles (80%) the authors acknowledged the limitations associated with the characteristics of an observational study. Thirty-eight articles (77%) acknowledged unmeasured confounding as a limitation in their study. The data quality issues, and the short follow-up or limited data collected were also notable concerns highlighted by the authors in 36 (73%) and 26 (53%) articles, respectively. Authors still viewed the lack of generalizability as a limitation in their observational studies in 20 articles (41%).

Discussion

Statement of principal findings

Our systematic evaluation of the use of causal language and its implications among observational studies in the ICU setting revealed the following findings: most of the articles included in our review did not follow to any reporting guidelines during their writing; non-causal terminology was widely used in articles that directly addressed causality and those that did not, regardless of whether the results were statistically significant or not; the key elements for appropriate causal inference in observational studies, such as dealing with missing data and interchangeability and generalizability issues were poorly reported and authors did not give sufficient considerations to methods for addressing the limitations of observational design.

Comparison with other studies

The STROBE statement was published in 2007, yet only 19% of the articles reviewed utilized this guideline for reporting observational studies. This issue is not unique to ICU studies. Generally, adherence to reporting guidelines in other health research areas is also low.[33–35] This raises concerns about the potential under-reporting of the items we identified as crucial for causal analysis.

Aligned with the findings in our review, most articles in the existing literature avoid directly discussing causes and instead use unclear and vague language. Olarte Parra et al.[36] conducted a review of 60 studies published in general medical journals with the aim of assessing the consistency of causal statements. In this review, many of the studies presented their conclusions in terms of associations while subtly incorporating causal messages within their findings. Similarly, Haber et al.[37] conducted a study to quantify the degree of causality implicit in the words linking exposure to outcomes and its consistency with the conclusions about their findings. They found a disconnection between the causality expressed in technical linking language and the research implications. The use of technical language that is not aligned with research implications may distort the interpretation of findings, hinder decision-making, and diminish transparency, impacting the credibility of research.[38,39] To maintain credibility, researchers must prioritize accurate language to convey the true implications of their findings and avoid the ''Schrodinger's causal paradox'' where the authors are cautious with their causal language while continuing to offer causal interpretations.[40]

In the articles of the group directly addressing causality, confounding was considered in most, but less account was taken of unmeasured confounding. The validity of observational research relies heavily on the assumption that all potential confounding factors are adequately measured and accounted for. However, despite methodologies available to assess and quantify the influence of unmeasured confounding on the outcomes,[41–43] only eight articles explicitly addressed the analysis of unmeasured confounding in their methods section.

Another crucial aspect of causal analysis is the handling of missing data. Among the group of studies directly addressing causality, about half of them acknowledged the presence of missing data. However, only in six articles (12%) did the authors make assumptions about the mechanisms behind the missing data. Nevertheless, authors employed various approaches to correct biases due to missing data, with multiple imputation being the most commonly used approach to address this issue. This reporting gap is consistent with findings from other studies, underscoring the insufficient reporting and handling of missing data in longitudinal observational studies. In 2004, Burton et al.[44] published a review of missing data in cancer prognostic studies and found a deficiency in the reporting of missing covariate. After reviewing 100 articles, they found that only 40% of articles provided information about the method used to handle missing covariate data and only 12 articles would have satisfied their proposed guidelines for the reporting of missing data. In a study conducted in 2012, Karahalios et al.[45] reviewed cohort study publications in PubMed. They found that a greater number of articles reported the method employed to address missing data in the analysis. However, many articles still did not report the amount of missing data and the reasons for missingness. Frameworks are available to assist researchers in systematically considering missing data and transparently reporting its potential impact on study outcomes.[46] One crucial step in these frameworks is identifying plausible mechanisms behind missingness, which is one of the least reported aspects.

In 2020, Tennant et al.[15] conducted a review examining the use of DAGs in applied health research and noted their increasing popularity for identifying confounding variables. However, out of the articles reviewed, only seven (14%) utilized DAGs to visually represent the relationship between exposure and outcome variables. Nevertheless, alongside the use of DAGs, there is a tendency to select confounding variables based on prior knowledge rather than relying solely on the results of univariate analysis. This suggests that researchers are more inclined to incorporate established confounding factors into their study design, emphasizing reliance on prior knowledge rather than solely on statistical associations observed in the initial analysis, which aligns with various recommendations.[47] This contrasts with the

predominant approach used in 18 articles (37%), which relies on regression model fitting as the primary approach for control of confounding, despite its suboptimal effectiveness for this purpose.[17,39,48] The widespread use of multivariate regression as a technique for controlling confounding is in the background of limited evaluation of its effectiveness in reducing confounding. Only 16 articles assessed this aspect using metrics such as Standardized Mean Difference (SMD) before and after adjustment.[49]

Although other articles may have concluded that the inability to attribute causality in observational research was rarely mentioned in journals,[50] in our case, the authors are cautious, highlighting the main characteristics of observational studies as major limitations. This recognition is commendable, researchers need to acknowledge the substantial limitations of observational studies. However, authors should incorporate a consideration regarding the efforts undertaken to ensure unbiased results in their discussions, particularly if they employ robust methodologies to alleviate the effects of some the limitations inherent in observational studies. Distinguishing between an insurmountable limitation inherent to observational studies and a limitation arising from uncertainty in results due to inadequate application of current methods is crucial. Researchers should receive formal training in statistics and research methodology, equipping them with the skills needed to conduct analyses that align with best practices.

Strengths and weaknesses of the study

Consistency and precision of language is crucial in observational research. Even in the absence of explicit statements, a causal conclusion is implicit when the language encourages interventions. This demands authors to give special attention to ensure that every word in the title and main objective are well-thought-through. Avoiding causal language inconsistencies is important because readers ought to be able to trust the conclusion reached regarding causality. To this end, we encourage researchers to adhere to reporting guidelines such as STROBE.

Our article has some limitations. One of the primary limitations of conducting a narrative literature review is the inherent subjectivity in synthesizing findings. Additionally, the broad and general nature of our search criteria resulted in a large volume of articles, making it challenging to manage and thoroughly analyse each piece of literature. We have not been able to include in this review the observational studies where the term ''causal'' did not appear in the text. Additional research is required to investigate the extent to which causal claims stated in the text are substantiated by the design and methods applied. This necessitates a more in-depth evaluation of the methods used and whether the articles manage to eliminate potential biases present to meet all the necessary assumptions for drawing causal conclusions. Furthermore, in future studies, it would be interesting to include all those articles that conduct causal statistical analyses, regardless of whether they explicitly use the term ''causal''. We anticipate that our review serves as a step towards a precise systematic evaluation. This will involve assessing the level of causality implied in the language used in observational ICU research and analysing its consistency with the methods employed in study designs and the results obtained in statistical analysis for causal inference.

Conclusion

Language consistency and precision are vital in observational research. Even without explicit causal statements, language suggesting interventions can imply causality, and misinterpretations can impact decision-making. Researchers should balance causal language with careful statistical analysis to enhance the clarity and robustness of their findings. Understanding statistical methods and following established guidelines like STROBE will improve the accuracy of future research and contribute to a clearer and more reliable body of knowledge.

CRediT authorship contribution statement

LDCA, JZ and AM conceived the study. LDCA and AM developed the research protocol and search strategy. LDCA and AM screened the titles, abstracts, and full texts. LDCA and AGS did the data abstraction. AM and JZ verified the study data. LDCA did the statistical analysis. LDCA, JZ and AM cowrote the first draft of the manuscript, and JZ, KK, OP and JIPZ reviewed subsequent drafts. JZ and KK provided major revisions to the first draft and reviewed subsequent drafts. All authors had full access to all the study data throughout the review process. All authors contributed to the final edits and agreed to submit the final manuscript.

Ethics declaration

Not applicable.

Declaration of Generative AI and AI-assisted technologies in the writing process

The authors declare that they have not used any type of generative artificial intelligence to write this manuscript or to create images, graphics, tables, or their corresponding captions.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Declaration of competing interest

None of the authors have conflicts of interest.

Data availability

The data used in this systematic review are derived from publicly available. Readers can access the original articles for detailed information on the data utilized. Access to specific datasets may vary depending on the policies of indi-

vidual publishers. For further inquiries regarding the data used in this study, please contact the corresponding author.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.medine.2025.502142.

References

1. Robins J, Hernán M. Advances in longitudinal data analysis. Boca Raton, FL: Chapman & Hall; 2009. p. 553–99.
2. Coscia Requena C, Muriel A, Peñuelas O. Analysis of causality from observational studies and its application in clinical research in Intensive Care Medicine. Med Intensiva (Engl Ed). 2018;42(4):292–300, http://dx.doi.org/10.1016/j.medine.2018.01.010.
3. Pattison N, Arulkumaran N, Humphreys S, Walsh T. Exploring obstacles to critical care trials in the UK: a qualitative investigation. J Intensive Care Soc. 2017;18(1):36–46, http://dx.doi.org/10.1177/1751143716663749.
4. Ford VJ, Klein HG, Danner RL, Applefeld WN, Wang J, Cortes-Puch I, et al. Controls, comparator arms, and designs for critical care comparative effectiveness research: it's complicated. Clin Trials. 2024;21(1):124–35, http://dx.doi.org/10.1177/17407745231195094.
5. Ridley S, Burchett K, Gunning K, Burns A, Kong A, Wright M, et al. Heterogeneity in intensive care units: fact or fiction? Anaesthesia. 1997:531–7, http://dx.doi.org/10.1111/j.1365-2222.1997.109-az0109.x.
6. Dreyfuss D. Is it better to consent to an RCT or to care? Intensive Care Med. 2005;31:345–55, http://dx.doi.org/10.1007/s00134-004-2493-0.
7. Schweickert W, Hall J. Informed consent in the intensive care unit: ensuring understanding in a complex environment. Curr Opin Crit Care. 2005;11(6):624–8, http://dx.doi.org/10.1097/01.ccx.0000186378.41697.09.
8. Ecarnot F, Quenot JP, Besch G, Piton G. Ethical challenges involved in obtaining consent for research from patients hospitalized in the intensive care unit. Ann Transl Med. 2017;5(4):S41, http://dx.doi.org/10.21037/atm.2017.04.42.
9. Hays M, Andrews M, Wilson R, Callender D, O'Malley P, Douglas K. Reporting quality of randomised controlled trial abstracts among high-impact general medical journals: a review and analysis. BMJ Open. 2016;6, http://dx.doi.org/10.1136/bmjopen-2016-011082.
10. Cartwright N. Are RCTs the gold standard? BioScieties. 2007;2(1):11–20.
11. Port FK. Role of observational studies versus clinical trials in ESRD research. Kidney Int. 2000;57:S3–6, http://dx.doi.org/10.1046/j.1523-1755.2000.07402.x.
12. Rubin D. Estimating causal effects of treatments in randomized and nonrandomized studies. J Educ Psychol. 1974;66(5):688–701, http://dx.doi.org/10.1037/h0037350.
13. Lederer DJ, Bell SC, Branson RD, Chalmers JD, Marshall R, Maslove DM, et al. Control of confounding and reporting of results in causal inference studies. Guidance for authors from editors of respiratory, sleep, and critical care journals. Ann Am Thorac Soc. 2019;16(1):22–8, http://dx.doi.org/10.1513/AnnalsATS.201808-564PS.
14. Pearl J. Causal diagrams for empirical research. Biometrika. 1995;82(4):669–88, http://dx.doi.org/10.2307/2337329.
15. Tennant PWG, Murray EJ, Arnold KF, Berrie L, Fox MP, Gadd SC, et al. Use of directed acyclic graphs (DAGs) to identify confounders in applied health research: review and recommendations International. Int J Epidemiol. 2021;60(2):620–32, http://dx.doi.org/10.1093/ije/dyaa213.
16. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983;70(1):41–55, http://dx.doi.org/10.2307/2335942.
17. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivar Behav Res. 2011;46(3):399–424, http://dx.doi.org/10.1080/00273171.2011.568786.
18. Austin PC, Stuart EA. Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. Stat Med. 2015;34(28):3661–79, http://dx.doi.org/10.1002/sim.6607.
19. Cole SR, Hernan MA. Constructing inverse probability weights for marginal structural models. Am J Epidemiol. 2008;168(6):656–64, http://dx.doi.org/10.1093/aje/kwn164.
20. Vansteelandt S, Joffe M. Structural nested models and G-estimation: the partially realized promise. Stat Sci. 2014;29(4):707–31, http://dx.doi.org/10.1214/14-STS493.
21. Edwards JK, McGrath LJ, Buckley JP, Schubauer-Berigan MK, Cole SR, Richardson DB. Occupational radon exposure and lung cancer mortality: estimating intervention effects using the parametric g-formula. Epidemiology. 2014;25(6):829–34, http://dx.doi.org/10.1097/EDE.0000000000000164.
22. Kontopantelis E, Doran T, Springate D, Buchan I, Reeves D. Regression-based quasi-experimental approach when randomization is not an option: interrupted time series analysis. BMJ. 2014;350:1–4, http://dx.doi.org/10.1136/bmj.h2750.
23. Baiocchi M, Cheng J, Small DS. Instrumental variable methods for causal inference: instrumental variable methods for causal inference. Stat Med. 2014;33(13):2297–340, http://dx.doi.org/10.1002/sim.6128.
24. Robins J, Hernán M, Brumback B. Marginal structural models and causal inference in epidemiology. Epidemiology. 2000;11(5):550–60, http://dx.doi.org/10.1097/00001648-200009000-00011.
25. Kitsios G, Dahabreh I, Callahan S, Paulus J, Campagna A, Dargin J. Can we trust observational studies using propensity scores in the critical care literature? A systematic comparison with randomized clinical trials. Crit Care Med. 2015;43(9):1870–9, http://dx.doi.org/10.1097/CCM.0000000000001135.
26. Anglemyer A, Horvath H, Bero L. Healthcare outcomes assessed with observational study designs compared with those assessed in randomized trials. Cochrane Database Syst Rev. 2014;4:MR000034, http://dx.doi.org/10.1002/14651858.MR000034.pub2.
27. Kuss O, Legler T, Börgermann J. Treatments effects from randomized trials and propensity score analyses were similar in similar populations in an example from cardiac surgery. J Clin Epidemiol. 2011;64(10):1076–84, http://dx.doi.org/10.1016/j.jclinepi.2011.01.005.
28. Hernán MA. The C-word: scientific euphemisms do not improve causal inference from observational data. Am J Public Health. 2018;108(5):616–9, http://dx.doi.org/10.2105/AJPH.2018.304337.
29. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021:n71, http://dx.doi.org/10.1136/bmj.n71.
30. Thapa DK, Visentin DC, Hunt GE, Watson R, Cleary M. Being honest with causal language in writing for publication. J Adv Nurs. 2020;76(6):1285–8, http://dx.doi.org/10.1111/jan.14311.
31. Von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational stud-

ies. Int J Surg. 2014;12(12):1495–9, http://dx.doi.org/10.1016/j.jclinepi.2007.11.008.
32. Sterne JA, Hernán MA, Reeves BC, Savović J, Berkman ND, Viswanathan M, et al. ROBINS-I: a tool for assessing risk of bias in non-randomized studies of interventions. BMJ. 2016;(i4919):355, http://dx.doi.org/10.1136/bmj.i4919.
33. Papathanasiou AA, Zintzaras E. Assessing the quality of reporting of observational studies in cancer. Ann Epidemiol. 2010;20(1):67–73, http://dx.doi.org/10.1016/j.annepidem.2009.09.007.
34. Ziemann S, Paetzolt I, Grüßer L, Coburn M, Rossaint R, Kowark A. Poor reporting quality of observational clinical studies comparing treatments of COVID-19 — a retrospective cross-sectional study. BMC Med Res Methodol. 2022;22(1):23, http://dx.doi.org/10.1186/s12874-021-01501-9.
35. Sarkis-Onofre R, Cenci MS, Demarco FF, Lynch CD, Fleming PS, Pereira-Cenci T, et al. Use of guidelines to improve the quality and transparency of reporting oral health research. J Dent. 2015;43(4):397–404, http://dx.doi.org/10.1016/j.jdent.2015.01.006.
36. Olarte Parra C, Bertizzolo L, Schroter S, Dechartres A, Goetghebeur E. Consistency of causal claims in observational studies: a review of papers published in a general medical journal. BMJ Open. 2021;11(5):e043339, http://dx.doi.org/10.1136/bmjopen-2020-043339.
37. Haber NA, Wieten SE, Rohrer JM, Arah OA, Tennant PWG, Stuart EA, et al. Causal and associational language in observational health research: a systematic evaluation. Am J Epidemiol. 2022;191(12):2084–97, http://dx.doi.org/10.1093/aje/kwac137.
38. Collins G, Le Manach Y. Comparing treatment effects between propensity scores and randomized controlled trials: improving conduct and reporting. Eur Heart J. 2004;33:1867–9, http://dx.doi.org/10.1093/eurheartj/ehs186.
39. Martens EP, Pestman WR, De Boer A, Belitser SV, Klungel OH. Systematic differences in treatment effect estimates between propensity score methods and logistic regression. Int J Epidemiol. 2008;37(5):1142–7, http://dx.doi.org/10.1093/ije/dyn079.
40. Tennant PWG, Murray EJ. The quest for timely insights into COVID-19 should not come at the cost of scientific rigor. Epidemiology. 2021;32(1), http://dx.doi.org/10.1097/EDE.0000000000001258, e2–e2.
41. Zhang X, Stamey JD, Mathur MB. Assessing the impact of unmeasured confounders for credible and reliable real-world evidence. Pharmacoepidemiol Drug Saf. 2020;29(10):1219–27, http://dx.doi.org/10.1002/pds.5117.
42. Huang R, Xu R, Dulai PS. Sensitivity analysis of treatment effect to unmeasured confounding in observational studies with survival and competing risks outcomes. Stat Med. 2020;39(24):3397–411, http://dx.doi.org/10.1002/sim.8672.
43. Gaster T, Eggertsen CM, Støvring H, Ehrenstein V, Petersen I. Quantifying the impact of unmeasured confounding in observational studies with the E value. BMJ Med. 2023;2(1):e000366, http://dx.doi.org/10.1136/bmjmed-2022-000366.
44. Burton A, Altman DG. Missing covariate data within cancer prognostic studies: a review of current reporting and proposed guidelines. Br J Cancer. 2004;91(1):4–8, http://dx.doi.org/10.1038/sj.bjc.6601907.
45. Karahalios A, Baglietto L, Carlin JB, English DR, Simpson JA. A review of the reporting and handling of missing data in cohort studies with repeated assessment of exposure measures. BMC Med Res Methodol. 2012;12(1):96, http://dx.doi.org/10.1186/1471-2288-12-96.
46. Lee KJ, Tilling KM, Cornish RP, Little RJA, Bell ML, Goetghebeur E, et al. Framework for the treatment and reporting of missing data in observational studies: the treatment and reporting of missing data in observational studies framework. J Clin Epidemiol. 2021;134:79–88, http://dx.doi.org/10.1016/j.jclinepi.2021.01.008.
47. VanderWeele TJ, Shpitser I. A new criterion for confounder selection. Biometrics. 2011;67(4):1406–13, http://dx.doi.org/10.1111/j.1541-0420.2011.01619.x.
48. Cepeda MS. Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. Am J Epidemiol. 2003;158(3):280–7, http://dx.doi.org/10.1093/aje/kwg115.
49. Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. Stat Med. 2009;28(25):3083–107, http://dx.doi.org/10.1002/sim.3697.
50. Wang MTM, Bolland MJ, Grey A. Reporting of limitations of observational research. JAMA Intern Med. 2015;175(9):1571–2, http://dx.doi.org/10.1001/jamainternmed.2015.2147.