TURKISH JOURNAL OF EMERGENCY MEDICINE

Invited Review Article

Access this article online

Quick Response Code:

Website: https://journals.lww.com/TJEM

DOI: 10.4103/tjem.tjem_214_26

Retrospective studies in emergency medicine: A practical guide to study design, bias, confounding, and interpretation

Suha Turkmen[1,2]*, Sinan Pasli[3]

[1]Department of Emergency Medicine, Hamad Medical Corporation, Doha, Qatar, [2]Department of Emergency Medicine, Qatar University, Doha, Qatar, [3]Department of Emergency Medicine, School of Medicine, Karadeniz Technical University, Trabzon, Türkiye

**Corresponding author*

Abstract:

Retrospective observational studies are a cornerstone of emergency medicine research, offering practical ways to address important clinical questions using real-world data. Yet, despite their widespread use, these studies are often misunderstood. Associations may be interpreted too confidently as causal effects, while important issues such as bias, confounding, and limitations of statistical adjustment may be overlooked or oversimplified. This practical guide provides a clinically grounded framework for understanding how retrospective studies should be designed, interpreted, and critically appraised. It reviews core methodological principles, including protocol development, sampling strategies, data abstraction, common sources of bias, confounding by indication, and responsible interpretation of observational findings. Frequent misconceptions, such as equating statistical significance with clinical importance or assuming adjustment can fully resolve confounding, are discussed through pragmatic examples relevant to emergency care. Rather than focusing on complex statistical techniques, this review emphasizes conceptual clarity, transparency, and thoughtful interpretation. It highlights practical principles that can improve validity and reproducibility while helping readers recognize both the strengths and inherent limitations of retrospective research. Emerging analytical approaches and evolving discussions around reproducibility and reporting standards are also considered. By linking foundational epidemiologic principles with the realities of emergency medicine practice, this article aims to support clinicians, trainees, reviewers, and early career investigators in designing stronger studies, interpreting findings more critically, and using observational evidence more thoughtfully in clinical decision-making and future research.

Keywords:

Bias, causal inference, confounding, critical appraisal, emergency medicine, observational research, retrospective studies, study design

Submitted: 30-04-2026
Revised: 01-06-2026
Accepted: 06-06-2026
Published: 09-07-2026

ORCID:
ST: 0000-0002-0557-6071
SP: 0000-0003-0052-2258

Address for correspondence:
Prof. Suha Turkmen, Department of Emergency Medicine, Qatar University, Doha, Qatar.
E-mail: drsuhaturkmen@hotmail.com

Introduction

Retrospective observational studies account for a large and steadily increasing proportion of research in emergency medicine. Their appeal is largely practical. Compared with prospective designs, they can often be completed more quickly, require fewer resources, and take advantage of routinely collected clinical data from diverse patient populations.[1] In a clinical environment defined by urgency, unpredictability, and logistical constraints, these designs frequently represent the most realistic way to address clinically relevant questions.[2,3]

How to cite this article: Turkmen S, Pasli S. Retrospective studies in emergency medicine: A practical guide to study design, bias, confounding, and interpretation. Turk J Emerg Med 2026;26:183-8.

In addition, this approach allows for the examination of rare events or situations that cannot be randomized for ethical reasons.[4] At the same time, retrospective studies are commonly misunderstood. They may be dismissed outright because they lack randomization, or conversely, overinterpreted when statistically significant associations are mistaken for causal relationships. Both reactions reflect an incomplete appreciation of what retrospective designs can reasonably achieve. The absence of controlled exposure allocation limits causal inference, but it does not render findings meaningless. Problems arise when conclusions extend beyond what the design can legitimately support. This article focuses on key conceptual principles for study design, data collection, peer review, and interpretation rather than exhaustive statistical methods.

A defining challenge of retrospective research is its vulnerability to bias and confounding. Selection bias, information bias, and confounding by indication are inherent risks when analyzing data that were not originally collected for research purposes. Analytical strategies such as multivariable regression or propensity-based methods can reduce, but never fully eliminate, these threats. Overreliance on statistical adjustment has, in some cases, created unwarranted confidence in results that remain fundamentally observational.

This article provides a practical user's guide to retrospective observational studies in emergency medicine. Rather than introducing new statistical methodologies, this article seeks to bridge the gap between established epidemiological principles and the unique realities of emergency medicine, where incomplete data, rapid decision-making, and clinician-driven interventions introduce distinct methodological challenges. The goal is to provide a practical, clinically grounded framework for designing, interpreting, and critically appraising retrospective research.

Why Retrospective Studies Matter in Emergency Medicine

Within the hierarchy of evidence, retrospective observational studies occupy a distinct and important role. Randomized controlled trials remain the preferred method for establishing causality, yet they are not always feasible, timely, or ethically appropriate in emergency care. Patients often present with acute or unstable conditions, interventions may need to be delivered immediately, and randomization or treatment withholding may be impractical.[3] Retrospective analyses therefore provide an essential alternative. Large electronic health record systems, administrative datasets, and institutional registries allow investigators to examine thousands of encounters efficiently. These data sources are particularly valuable for studying rare outcomes, variations in clinical practice, and real-world treatment patterns because they capture large and heterogeneous patient populations that are often underrepresented in randomized trials.[5] For many questions in emergency medicine, retrospective studies serve as the foundation for hypothesis generation and for informing the design of future prospective trials.[6] The value of retrospective research, however, depends on a clear understanding of its limitations. Appreciating what these studies can and cannot demonstrate is essential for their appropriate use in clinical decision-making, guideline development, and policy discussions. Although large electronic health record systems, administrative datasets, and registries can increase statistical power and allow rare outcomes to be studied, larger sample sizes do not by themselves improve validity. For example, if diagnostic codes inconsistently capture sepsis severity across hospitals, or if symptom onset is poorly documented in a substantial proportion of records, increasing the number of observations may amplify systematic misclassification rather than reduce error. Careful protocol development, predefined variable definitions, and transparent handling of missing data are therefore essential before large datasets can generate clinically meaningful evidence.

Fundamental Methodology of Retrospective Research

For many years, retrospective studies were viewed as relatively simple projects, often undertaken for rapid publication or academic requirements. This perception shifted following the landmark work by Gilbert and Lowenstein, which demonstrated that a substantial proportion of published chart reviews failed to meet basic methodological standards. Their work reframed chart review as a scientific method requiring rigor and systematic planning comparable to prospective research.[7]

Research protocol and formulation of the research question

High-quality retrospective research begins with a clearly defined research question and a written protocol developed before data extraction.[8] Approaches that involve searching datasets for statistically significant associations without a predefined question – the so-called "fishing expedition" – are inconsistent with scientific standards and substantially increase the risk of false-positive findings.[9] The research question determines the study population, eligibility criteria,

and variable definitions. Precise operational definitions are particularly important. Ambiguity in terms such as "hypotension," "fever," or "clinical improvement" is a frequent source of error in retrospective work. Defining these variables in advance and compiling a standardized data dictionary reduce misclassification and improve reproducibility.[4]

Sampling strategies and inclusion and exclusion criteria

Sample selection directly affects external validity. Convenience sampling, or the selective inclusion of readily available records, can introduce significant selection bias. Exclusion criteria should be reported with the same transparency as inclusion criteria.[7,10] Selecting an appropriate sample that represents the research population and conducting the sampling sequentially or randomly increase reliability. Importantly, investigators should describe and, where possible, compare patients excluded because of missing or incomplete data. If excluded patients are systematically sicker or more unstable, results may be biased toward better outcomes. In such cases, mortality or complication rates may be underestimated. Clear reporting of these issues allows readers to judge the potential impact on study conclusions.

For instance, a study evaluating outcomes in admitted trauma patients may underestimate mortality if patients who die in the emergency department before admission are excluded. Transparent reporting of inclusion and exclusion criteria is therefore critical.

Data collection and observer management

In retrospective studies, data abstraction functions as the primary measurement instrument. As with any instrument, it requires calibration. Data collectors should receive structured training, use standardized abstraction forms, and operate according to predefined rules.[10] Data collection should, whenever feasible, be conducted by individuals who are unaware of the study hypothesis and the outcome measures being evaluated. Awareness of the research question can unconsciously influence how ambiguous clinical notes are interpreted. This form of expectation bias is common and often underappreciated in retrospective research. Whenever more than one abstractor is involved, investigators should pilot test the abstraction form and report inter-rater agreement (e.g., using percent agreement or kappa statistics) for key variables.[4] Explicit rules for resolving discrepancies between abstractors further improve the reliability of data collection.

For instance, knowledge of the study objective may influence how ambiguous documentation is interpreted, leading to expectation bias.

Common Misconceptions about Retrospective Observational Research

Misconception 1: Association implies causation

The most persistent misconception is the assumption that observed associations reflect causal effects. In retrospective studies, treatment allocation is driven by clinical judgment, disease severity, comorbidities, and system factors rather than randomization. As a result, exposed and unexposed groups often differ in ways that are incompletely measured. These differences can produce misleading associations. Interventions preferentially used in sicker patients may appear harmful, while treatments reserved for more stable patients may seem beneficial. Without full adjustment for underlying risk, such findings should not be interpreted as causal. For example, the existence of a relationship between X and Y does not mean that X directly causes Y; another variable may also be creating this relationship.[11] Accordingly, results from retrospective studies should be framed as associations unless supported by exceptionally strong methodological safeguards and corroborating evidence from other designs. Causal language should be used sparingly and with explicit justification.

For instance, vasopressor use may appear associated with increased mortality, but this often reflects underlying illness severity rather than a harmful treatment effect.

Misconception 2: Statistical significance equals clinical importance

A frequent error in interpreting retrospective studies is the assumption that statistical significance implies clinical relevance. In large datasets, very small differences may achieve statistical significance despite having minimal practical importance. *P* values therefore cannot be interpreted in isolation. Effect size, confidence intervals, and absolute risk differences provide a more meaningful basis for clinical interpretation. As emphasized in the American Statistical Association statement, scientific conclusions should never hinge on a single statistical threshold.[12,13] Clinicians and researchers should consider whether the observed magnitude of effect is likely to influence decision-making, resource utilization, or patient outcomes. A statistically significant finding with a trivial absolute effect may be clinically irrelevant, whereas a nonsignificant result with a large point estimate and wide confidence interval may still warrant attention and further study. Even if the *P* value of a study is significant, the magnitude of the difference and its clinical importance must be evaluated.[14]

Misconception 3: Multivariable adjustment eliminates confounding

Multivariable regression, propensity score matching, and related techniques are essential tools in observational

research, but they do not eliminate confounding. Adjustment can only account for variables that are accurately measured and included in the model. Multiple regression models can only control for defined confounders; they are ineffective for unmeasured or unknown factors.[1] Unmeasured or poorly captured factors, such as frailty, clinician intuition, or institutional practice patterns, remain unaddressed. In emergency medicine, where treatment decisions are often guided by subtle clinical cues and time pressure, residual confounding is the rule rather than the exception. Adjusted estimates should therefore be interpreted as conditional associations rather than unbiased causal effects.

Bias and Confounding in Retrospective Studies

Bias refers to a consistent distortion that causes study results to differ from the true effect and is a major concern for the validity of retrospective observational studies. Because data are collected for clinical or administrative purposes rather than research, the potential for bias is embedded in patient selection, exposure classification, outcome ascertainment, and follow-up procedures.

Selection bias

Selection bias arises when the included study population does not adequately represent the target population. In emergency medicine, a classic example is admission bias, also referred to as Berkson's paradox. When hospital admission depends on both the exposure and the outcome, the associations observed in admitted patients may not represent the true relationship in the source population.[15] Limiting analyses to admitted patients, intensive care unit (ICU) transfers, or individuals with complete datasets can inadvertently exclude patients who die early, are rapidly discharged, or are transferred to other institutions. These exclusions may systematically remove higher-risk patients and lead to underestimation of mortality or complication rates. Once introduced, selection bias cannot be fully corrected by statistical adjustment, underscoring the importance of transparent cohort definitions and reporting.[16,17] For instance, restricting analysis to ICU admissions may exclude patients who die early, leading to underestimation of mortality.

Information bias

Information bias arises from systematic inaccuracies in how exposures, outcomes, or covariates are assessed, recorded, or classified. Retrospective studies depend heavily on electronic health records, billing codes, and clinical documentation that vary in accuracy and completeness.[18] Misclassification may dilute true associations or generate spurious ones, particularly when errors differ between comparison groups. Large database studies frequently rely on International Classification of Diseases (ICD) coding to identify diagnoses and outcomes. However, multiple studies have demonstrated limited sensitivity and specificity of diagnostic codes, especially for complex or stigmatized conditions such as suicide attempts.[19] Analyses based solely on structured codes may therefore underestimate incidence and mischaracterize comorbidity profiles. In contrast, information derived from clinician notes often reflects the clinical reality more accurately, albeit at the cost of increased abstraction complexity.[20] Where feasible, investigators should report any validation of coding algorithms, cite existing validation studies, and describe how potential misclassification may influence the direction and magnitude of their findings. For instance, reliance on ICD coding may misclassify conditions such as sepsis or suicide attempts.

Confounding and confounding by indication

Confounding arises when an external factor is related to both the exposure and the outcome, leading to a distorted estimate of their association.[1] In studies of treatment effectiveness, confounding by indication is particularly problematic. Treatments are selected based on clinician assessment of disease severity, prognosis, and perceived benefit rather than being assigned randomly.[21] In emergency medicine, many determinants of treatment choices, such as frailty, subtle physiologic trends, or clinician experience, are not fully captured in routine datasets. Even extensive adjustment cannot eliminate these unmeasured influences. As a result, apparent treatment effects in retrospective studies may reflect underlying clinical decision-making rather than true causal relationships. Hess and Boyko emphasized that residual confounding can be reduced using techniques such as propensity score matching, but that uncertainty will remain.[1,22] For instance, patients receiving aggressive interventions may appear to have worse outcomes due to higher baseline severity.

Principles for Responsible Interpretation and Reporting

Interpretation is often the stage at which otherwise rigorous retrospective studies lose credibility. Overinterpretation, causal language, and selective emphasis on statistically significant results can mislead readers and overstate the implications of the findings.

Use association-focused language

Results should be framed explicitly as associations unless strong justification for causal inference exists. Phrasing such as "was associated with" or "was more frequently observed among" appropriately reflects the limitations of the design. If causal interpretations

are proposed, they should be supported by robust sensitivity analyses and consistency with external evidence.

Emphasize effect size, precision, and clinical context

Effect estimates and their confidence intervals convey far more information than *P* values alone.[23] Absolute risk differences and measures of precision help clinicians judge whether an observed association is likely to be clinically meaningful. Findings with wide confidence intervals should be interpreted cautiously, even when point estimates appear compelling.

Avoid selective reporting and overinterpretation

Balanced reporting of both significant and nonsignificant results is essential. Secondary and exploratory analyses should be clearly identified as such. Overemphasis on borderline *P* values or "trends toward significance" risks overstating uncertain findings.

Acknowledge residual confounding and uncertainty

Residual confounding is unavoidable in retrospective observational research. Explicit discussion of unmeasured confounders and data limitations reflects methodological maturity rather than weakness. Authors should consider the likely direction of bias and situate their findings within the broader body of evidence.

Aligning conclusions with design and objectives

Conclusions must remain consistent with the study design and analytical approach. Retrospective studies are well suited for describing patterns of care and generating hypotheses, but they rarely provide definitive evidence for changes in clinical practice. Modest, design-consistent conclusions enhance credibility and facilitate appropriate application of the findings.

Data security and reporting standards

Researchers must adhere to ethical and legal standards for data protection, including appropriate de-identification or use of limited datasets. Reporting should follow established guidelines such as the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) and Reporting of Studies Conducted Using Observational Routinely Collected Health Data (RECORD) to ensure transparency and reproducibility.[24-26] Interpretation is often where otherwise well-designed retrospective studies falter. Overstated conclusions, causal language, and selective emphasis on significant findings can undermine credibility. Using association-focused language, emphasizing effect size and uncertainty, and aligning conclusions with study design help ensure that findings are applied appropriately.

Reproducibility and evolving statistical paradigms

Interest in the reproducibility of observational research has grown considerably in recent years. Practices such as selective reporting, data-driven analyses, and flexible modeling strategies can produce findings that are difficult to replicate. At the same time, there is ongoing debate about the limitations of *P* values and whether alternative approaches, including Bayesian methods, may offer more informative frameworks for interpretation. Reproducibility concerns have also affected observational research, particularly when multiple analytic decisions can influence results. Issues such as selective reporting and model flexibility deserve attention, especially in large database studies. Ongoing discussions about alternatives or complements to conventional *P* value-based inference reflect this broader concern.

Emerging Analytical Approaches

Recent methodological developments have expanded the tools available for retrospective research, including machine learning methods, natural language processing, and modern causal inference techniques. These approaches may improve prediction, support richer use of routinely collected data, and in some settings strengthen adjustment for confounding. Nevertheless, they do not eliminate core limitations related to bias, missing data, or study design. Their value depends less on novelty than on whether they are applied within sound epidemiologic principles.

Practical Checklist for Retrospective Studies

To support implementation, the following checklist summarizes key principles:

1. Clearly defined research question and protocol
2. Prespecified inclusion and exclusion criteria
3. Standardized data abstraction forms
4. Blinded data collection where feasible
5. Assessment of inter-rater reliability
6. Transparent handling of missing data
7. Identification of key confounders
8. Emphasis on effect size and confidence intervals
9. Avoidance of causal language
10. Adherence to the STROBE/RECORD guidelines.

Conclusion

Retrospective observational studies play a central role in emergency medicine research by offering access to large, diverse patient populations and real-world clinical data. When designed carefully and interpreted with appropriate caution, they can provide valuable insights that complement evidence from prospective and randomized studies. Their usefulness, however, depends

on recognizing their limitations. Misinterpretation of associations, overreliance on statistical significance, and excessive confidence in adjustment methods can all undermine validity. By emphasizing transparency, effect size, uncertainty, and design-appropriate conclusions, researchers can strengthen both the credibility and clinical relevance of their findings.

Author contribution statement

ST: Conceptualization, Supervision, Writing – original draft, and Writing – review and editing; SP: Conceptualization, Methodology, and Writing – original draft. Each author has approved the submitted version. Each author has agreed both to be personally accountable for the author's own contributions and to ensure that questions related to the accuracy or integrity of any part of the work, even ones in which the author was not personally involved, are appropriately investigated, resolved, and documented in the literature.

Conflicts of interest

None Declared.

Funding

None.

References

1. Hess DR. Observational studies. Respir Care 2023;68:1585-97.
2. Cimino J, Braun C. Clinical research in prehospital care: Current and future challenges. Clin Pract 2023;13:1266-85.
3. Morrison CA, Horwitz IB, Carrick MM. Ethical and legal issues in emergency research: Barriers to conducting prospective randomized trials in an emergency setting. J Surg Res 2009;157:115-22.
4. Kaji AH, Schriger D, Green S. Looking through the retrospectoscope: Reducing bias in emergency medicine chart review studies. Ann Emerg Med 2014;64:292-8.
5. Ohaiba MM, Anamazobi EG, Okobi OE, Aguda K, Chukwu VU. Trends and patterns in emergency department visits: A comprehensive analysis of adult data from the national center for health statistics (NCHS) database. Cureus 2024;16:e66059.
6. Dziadkowiec O, Durbin J, Muralidharan VJ, Novak M, Cornett B. Improving the quality and design of retrospective clinical outcome studies that utilize electronic health records. HCA Healthc J Med 2020;1:131-8.
7. Gilbert EH, Lowenstein SR, Koziol-McLain J, Barta DC, Steiner J. Chart reviews in emergency medicine research: Where are the methods? Ann Emerg Med 1996;27:305-8.
8. De Sanctis V, Soliman AT, Daar S, Tzoulis P, Fiscina B, Kattamis C, *et al*. Retrospective observational studies: Lights and shadows for medical writers. Acta Biomed 2022;93:e2022319.
9. Worster A, Bledsoe RD, Cleve P, Fernandes CM, Upadhye S, Eva K. Reassessing the methods of medical record review studies in emergency medicine research. Ann Emerg Med 2005;45:448-51.
10. Vassar M, Holzmann M. The retrospective chart review: Important methodological considerations. J Educ Eval Health Prof 2013;10:12.
11. Sargeant JM, O'Connor AM, Renter DG, Ruple A. What question are we trying to answer? Embracing causal inference. Front Vet Sci 2024;11:1402981.
12. Yaddanapudi LN. The American statistical association statement on *P* values explained. J Anaesthesiol Clin Pharmacol 2016;32:421-3.
13. Wasserstein RL, Lazar NA. The ASA statement on *P* values: Context, process, and purpose. Am Stat 2016;70:129-33.
14. Sharma H. Statistical significance or clinical significance? A researcher's dilemma for appropriate interpretation of research results. Saudi J Anaesth 2021;15:431.
15. Geneletti S, Richardson S, Best N. Adjusting for selection bias in retrospective, case-control studies. Biostatistics 2009;10:17-31.
16. Vena JE, Sultz HA, Carlo GL, Fiedler RC, Barnes RE. Sources of bias in retrospective cohort mortality studies: A note on treatment of subjects lost to follow-up. J Occup Med 1987;29:256-61.
17. del Junco DJ, Fox EE, Camp EA, Rahbar MH, Holcomb JB, PROMMTT Study Group. Seven deadly sins in trauma outcomes research: An epidemiologic post mortem for major causes of bias. J Trauma Acute Care Surg 2013;75:S97-103.
18. Ramirez-Santana M. Limitations and Biases in Cohort Studies [Internet]. Cohort Studies in Health Sciences. InTech; 2018. Available from: https://doi.org/10.5772/intechopen.74324. [Last accessed on 2026 Jun 25].
19. Xu R, Bode L, Geva A, Mandl KD, McMurry AJ. Accuracy of ICD-10 codes for suicidal ideation and action in pediatric emergency department encounters. Pediatr Emerg Care 2025;41:378-82.
20. Grasso MA, Dezman ZD, Jerrard DA. Coding disparity and specificity during emergency department visits after transitioning to the tenth version of the International Classification of Diseases. AMIA Annu Symp Proc 2022;2022:495-501.
21. Kyriacou DN, Lewis RJ. Confounding by indication in clinical research. JAMA 2016;316:1818-9.
22. Boyko EJ. Observational research-opportunities and limitations. J Diabetes Complications 2013;27:642-8.
23. Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, *et al*. Statistical tests, *P* values, confidence intervals, and power: A guide to misinterpretations. Eur J Epidemiol 2016;31:337-50.
24. Junod V, Elger B. Retrospective research: What are the ethical and legal requirements? Swiss Med Wkly 2010;140:w13041.
25. Benchimol EI, Smeeth L, Guttmann A, Harron K, Moher D, Petersen I, *et al*. The REporting of studies conducted using observational routinely-collected health data (RECORD) statement. PLoS Med 2015;12:e1001885.
26. Vandenbroucke JP, von Elm E, Altman DG, Gøtzsche PC, Mulrow CD, Pocock SJ, *et al*. Strengthening the reporting of observational studies in epidemiology (STROBE): Explanation and elaboration. PLoS Med 2007;4:e297.