INPATIENT DATABASES

Johannes M. Schwenke, Thale P. Brown, Corina S. Ruegg, Alain Amstutz, …

1 Introduction

Target trial emulation (TTE) uses observational data together with robust methods to answer causal questions about the effectiveness and safety of interventions, where randomized controlled trials would not be feasible (Matthews et al. 2022). In case of a new pandemic, TTE could be used to quickly expand upon the evidence generated by randomized controlled trials (RCT) and identify promising interventions for further investigation.

High quality TTE require detailed data on treandefinedtments, outcomes, and potential confounding factors (Hernán, Wang, and Leaf 2022). However, a systematic overview of databases suitable for conduction TTE is currently lacking, which could hinder a rapid response capabilities in case of a new pandemic.

We therefore conducted a scoping review to identify and characterize databases potentially suitable for TTE studies during public health emergencies. This review aims to provide researchers and public health decision-makers with a comprehensive overview of available data resources for rapid evidence generation during future pandemics.

2 Methods

Search Strategy We employed two complementary search strategies to identify potentially suitable databases for target trial emulations in infectious respiratory diseases.

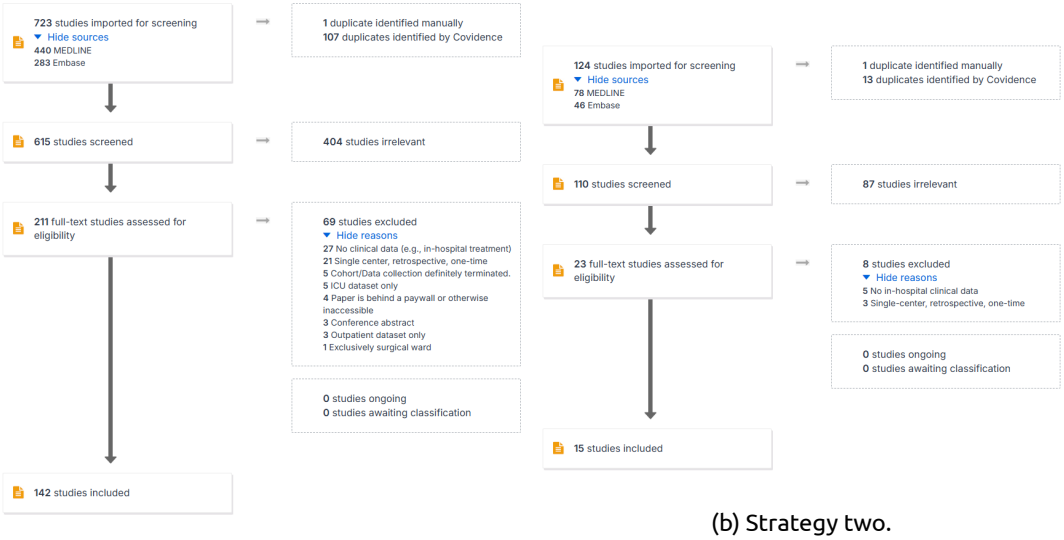
The first strategy identified databases indirectly through published comparative effectiveness studies that employed causal inference methods in inpatient settings. We searched MEDLINE (via Ovid) and Embase using a search string adapted from Smit et al. (Smit et al. 2022), combining terms related to causal inference methods, comparative effectiveness research, observational studies, and respiratory infections in hospital settings (full search strategy available in Section 5).

The second strategy aimed to directly identify suitable databases. We adapted the search string developed by Sauer et al. (Sauer et al. 2022) to identify healthcare databases capable of supporting clinical research in infectious diseases. This search was also conducted in MEDLINE (via Ovid) and Embase (full search strategy available in Section 5).

The systematic search was conducted from 2023-11-22 to 2024-11-25. Title and abstract screening, full-text review, and data extraction were all performed independently by two reviewers using covidence. Any disagreements at any stage were resolved through discussion with a third reviewer when necessary.

3 Results

Through search strategy 1 and two we identified 142 and 15 publications respectively (see Figure 1a and Figure 1b). We identified 70 databases underlying these publications (see Table 1).



(a) Strategy one.

(b) Strategy two.

Figure 1: PRSMA flowcharts for search strategy one and two.

Unfortunately, the full text screening wasn't very specific. During extraction it became apparent that some databases contained exclusively outpatient or bio banking data.[1] These databases (n = 9) were excluded from this report.

Summary statistics of the remaining databases.

- Judging from the publication only 32 (46%) of the databases are likely still collecting data
  - For 32 databases it was not clear whether data were still being collected
  - For 6 databases it was clear that data collection had stopped
- It was not possible to reliably extract how the underlying data of each publication could be accesses for further research purposes
  - Only 2 (3%) clearly stated that data was publicly available

Table 1: The top 5 databases ordered by number of studies identified using their data.

name	country	dataType	public	ongoing	n
Department of Veterans Health Administration (VHA) healthcare system	USA	Hospital data (electronic health record)	No	Yes	16

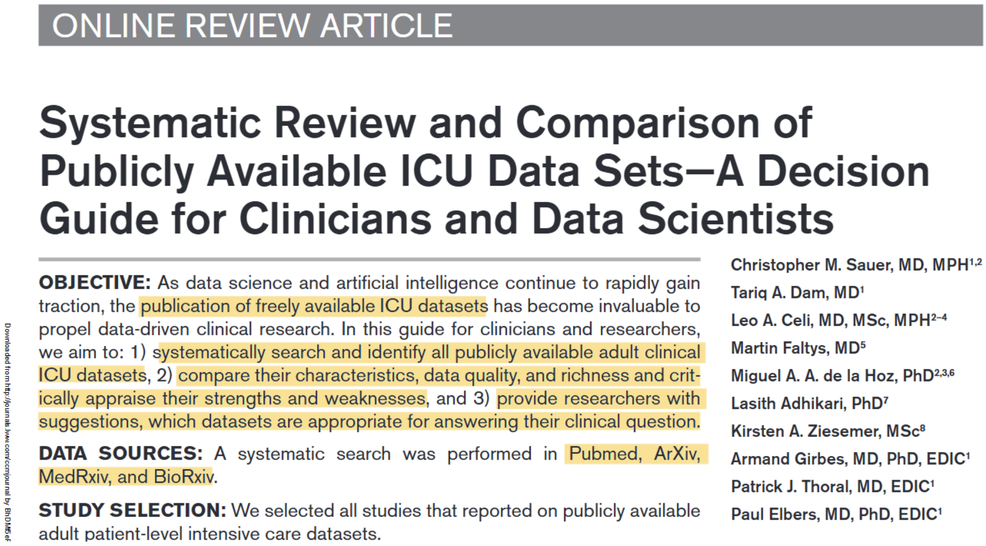
[1]The excluded databases are : iCTCF dataset, Hospital Episode Statistics, National (Nationwide) Inpatient Sample (NIS), CXR8, Clinical Practice Research Datalink (CPRD), Health and Economic Modelling of AMR in Australia (HEMAA) population-level simulation model , Avon Longitudinal Study of Parents and Children (ALSPAC), Public Health England, Covid 19 HGI.

name	country	dataType	public	ongoing	n
DPC Japan	Japan	Hospital data (electronic health record)	No	Yes	9
Premier Healthcare Database	USA	Insurance/claims data (prescription data)	No	Yes	6
Kaiser Permanente Southern California (KPSC)	USA	Hospital data (electronic health record)	No	Yes	5
AP-HP Datawarehouse	France	Hospital data (electronic health record)	No	Yes	4

A more complete overview of the found databases can be found here.

4 What next?

Analogous approach to Sauer et al. ?


ONLINE REVIEW ARTICLE

Systematic Review and Comparison of Publicly Available ICU Data Sets—A Decision Guide for Clinicians and Data Scientists

OBJECTIVE: As data science and artificial intelligence continue to rapidly gain traction, the publication of freely available ICU datasets has become invaluable to propel data-driven clinical research. In this guide for clinicians and researchers, we aim to: 1) systematically search and identify all publicly available adult clinical ICU datasets, 2) compare their characteristics, data quality, and richness and critically appraise their strengths and weaknesses, and 3) provide researchers with suggestions, which datasets are appropriate for answering their clinical question.

DATA SOURCES: A systematic search was performed in Pubmed, ArXiv, MedRxiv, and BioRxiv.

STUDY SELECTION: We selected all studies that reported on publicly available adult patient-level intensive care datasets.

Christopher M. Sauer, MD, MPH[1,2]
Tariq A. Dam, MD[1]
Leo A. Celi, MD, MSc, MPH[2–4]
Martin Faltys, MD[5]
Miguel A. A. de la Hoz, PhD[2,3,6]
Lasith Adhikari, PhD[7]
Kirsten A. Ziesemer, MSc[8]
Armand Girbes, MD, PhD, EDIC[1]
Patrick J. Thoral, MD, EDIC[1]
Paul Elbers, MD, PhD, EDIC[1]

Figure 3: Sauer, Christopher M., Tariq A. Dam, Leo A. Celi, Martin Faltys, Miguel A. A. de la Hoz, Lasith Adhikari, Kirsten A. Ziesemer, Armand Girbes, Patrick J. Thoral, and Paul Elbers. 2022. "Systematic Review and Comparison of Publicly Available ICU Data Sets-A Decision Guide for Clinicians and Data Scientists." Critical Care Medicine 50 (6): e581–88.

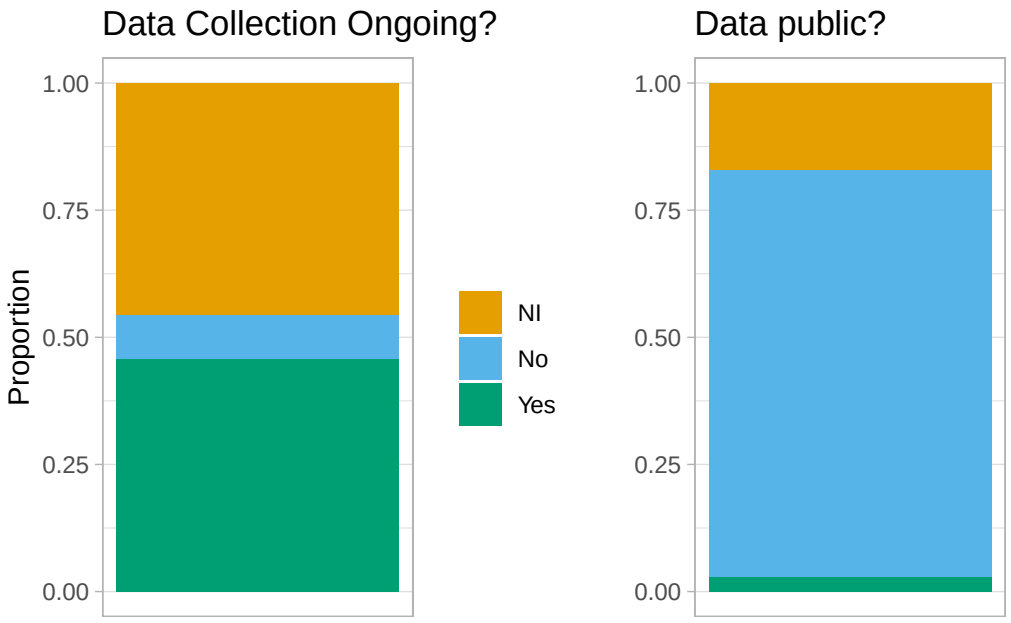


Figure 2

- Requested access to databases, all necessary legal and ethical approvals
- Defined a priori list of database elements to document and extract
- → Descriptive analyses

5 Appendix A

5 Search string for strategy 1

5.1.1 Medline

((caus adj3 (inferen* or model*)) or ((causal or average-treatment* or individuali*-treatment* or personali*-treatment*) adj (effect* or method*)) or time-vary*-confound* or g-computation* or g-estimation* or g-formula* or doubly-robust-estimation* or counterfactual* or (inverse-probabilit* adj3 (weight* or estimat*)) or ((marginal-structur* or structural-nest* or causal-effect* or causal-graphic* or causal-inferen* or semi-paramet* or semiparamet* or fully-paramet*) adj3 (method* or model*)) or TAR-Net or (Treatment*-Agnost* adj3 Representat* adj3 Network*) or double-machine-learning or anchor*-regress* or x- learner* or t-learner* or s-learner* or q-learning or q-network or reinforcement*-learn* or ((policy or value) adj iteration*) or temporal-differen* or actor-critic* or (Markov adj3 decision adj3 process*)).ab,ti. or (RL or IRL).ti.*

AND

(exp "Respiratory Tract Infections"/)

AND

((exp Hospitals/) or hospital or "secondary care")*

5.1.1.1 Hits

- 22/11/2023: 1164 hits
- 22/11/2023: With ((exp Hospitals/) or hospital*.ti,ab,kw. or "secondary care".ti,ab,kw.): 422 hits
- 25/11/2023: With ((exp Hospitals/) or hospital* or "secondary care"): 437 hits

5.1.2 Embase

((caus adj3 (inferen* or model*)) or ((causal or average-treatment* or individuali*-treatment* or personali*-treatment*) adj (effect* or method*)) or time-vary*-confound* or g-computation* or g-estimation* or g-formula* or doubly-robust-estimation* or counterfactual* or (inverse-probabilit* adj3 (weight* or estimat*)) or ((marginal-structur* or structural-nest* or causal-effect* or causal-graphic* or causal-inferen* or semi-paramet* or semiparamet* or fully-paramet*) adj3 (method* or model*)) or TAR-Net or (Treatment*-Agnost* adj3 Representat* adj3 Network*) or double-machine-learning or anchor*-regress* or x- learner* or t-learner* or s-learner* or q-learning or q-network or reinforcement*-learn* or ((policy or value) adj iteration*) or temporal-differen* or actor-critic* or (Markov adj3 decision adj3 process*)).ab,ti. or (RL or IRL).ti.*

AND

(exp "respiratory tract infection"/)

AND

((exp hospital/) or hospital or "secondary care")*

5.1.2.1 Hits

- 25/11/2023: 284 hits

5 Search string for strategy 2

5.2.1 Medline

("Data Warehousing"/) OR ("datawarehous".ti,ab,kw.) OR ("Database Management Systems"/) OR ("dataset*".ti,ab,kw.) OR ("data set*".ti,ab,kw.) OR ("database*".ti,ab,kw.)*

AND

(("publicly available" OR "free of charge" OR "freely accessible" OR "publicly accessible").ti,ab,kw.)

AND

(exp "Respiratory Tract Infections"/)

AND

((exp Hospitals/) or hospital or "secondary care")*

5.2.1.1 Hits

- 22/11/2023: 557 hits
- 22/11/2023: With ((exp Hospitals/) or hospital*.ti,ab,kw. or "secondary care".ti,ab,kw.): 74 hits
- 25/11/2023: With ((exp Hospitals/) or hospital* or "secondary care"): 78 hits

5.2.2 Embase

("data warehouse"/) OR ("datawarehous".ti,ab,kw.) OR ("database management system*"/) OR ("dataset*".ti,ab,kw.) OR ("data set*".ti,ab,kw.) OR ("database*".ti,ab,kw.)*

AND

(("publicly available" OR "free of charge" OR "freely accessible" OR "publicly accessible").ti,ab,kw.)

AND

(exp "respiratory tract infection"/)

AND

((exp hospital/) or hospital or "secondary care")*

5.2.2.1 Hits

- 25/11/2023: 46 hits

6 References

Hernán, Miguel A., Wei Wang, and David E. Leaf. 2022. "Target Trial Emulation." *JAMA* 328 (24): 2446. https://doi.org/10.1001/jama.2022.21383.

Matthews, Anthony A, Goodarz Danaei, Nazrul Islam, and Tobias Kurth. 2022. "Target Trial Emulation: Applying Principles of Randomised Trials to Observational Studies." *BMJ*, August, e071108. https://doi.org/10.1136/bmj-2022-071108.

Sauer, Christopher M., Tariq A. Dam, Leo A. Celi, Martin Faltys, Miguel A. A. de la Hoz, Lasith Adhikari, Kirsten A. Ziesemer, Armand Girbes, Patrick J. Thoral, and Paul Elbers. 2022. "Systematic Review and Comparison of Publicly Available ICU Data Sets—A Decision Guide for Clinicians and Data Scientists." *Critical Care Medicine* 50 (6): e581–88. https://doi.org/10.1097/ccm.0000000000005517.

Smit, J. M., J. H. Krijthe, J. van Bommel, J. A. Labrecque, M. Komorowski, D. A. M. P. J. Gommers, M. J. T. Reinders, and M. E. van Genderen. 2022. "Causal Inference Using Observational Intensive Care Unit Data: A Systematic Review and Recommendations for Future Practice." http://dx.doi.org/10.1101/2022.10.29.22281684.