DATABASES FOR TARGET TRIAL EMULATION

A short report on the identification and characterization of databases suitable for target trial emulation in infectious respiratory diseases.

1 Summary

This report summarizes our work to identify databases potentially suitable for target trial emulation (TTE) in case of a new pandemic. Through systematic literature searches, we identified 79 databases, of which 70 were included in our final analysis after excluding those with exclusively outpatient or biobanking data.

Key findings:

- Only 46% (32/70) of databases were clearly still actively collecting data
- For 46% (32/70) of databases, current collection status is unclear
- Only 3% (2/70) clearly indicated public data availability
- Access procedures for research purposes were poorly documented in publications

Therefore we believe the next step is a systematic survey of database administrators to assess current status, access possibilities, and data sharing capabilities.

2 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 [1]. 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 treatment, outcomes, and potential confounding factors [2]. However, a systematic overview of databases suitable for conducting TTE is currently lacking, which could hinder a rapid response 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, with a focus on respiratory diseases.

3 Methods

We used two complementary search strategies to identify potentially suitable databases for TTE 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. [3], combining terms related to causal inference methods, comparative effectiveness research, observational studies, and respiratory infections in hospital settings (full search strategy available in Section 6).

The second strategy aimed to directly identify suitable databases. We adapted the search string developed by Sauer et al. [4] to identify healthcare databases capable of supporting TTE. This search was also conducted in MEDLINE (via Ovid) and Embase (full search strategy available in Section 6).

The systematic search was conducted from 2023-11-22 to 2023-11-25. 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.

Information on databases identified through discussions with experts or discovered through other means was separately extracted.

4 Results

Through search strategies one and two we identified 142 and 15 publications, respectively (see Figure 1a and Figure 1b). We identified 79 databases underlying these publications.

(a) Strategy one. (b) Strategy two.

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

Through discussions with experts and non-systematic searcher another **XXX** databases were identified.

Unfortunately, the full text screening was not 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.

[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.

- It was not possible to reliably extract how the underlying data of each publication could be accessed for further research purposes.
  - Only 2 (3%) clearly stated that data were publicly available.

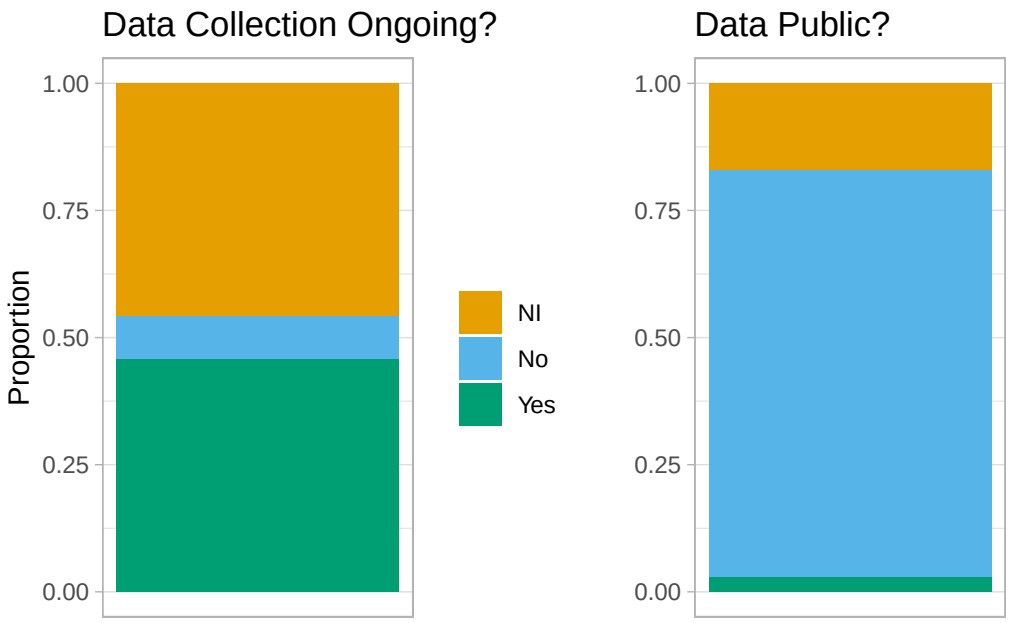

Figure 2: Proportion of databases with ongoing data collection and public data availability.

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

5 What next?

We were not able to sufficiently characterize databases and their access policies through the information provided in the publications. To address the information gaps identified in our review, we will conduct a systematic survey of database administrators and study authors. The survey aims to:

1. Verify current database status and prospective data collection
2. Document access procedures and requirements for research collaborations
3. Assess willingness to share detailed database characteristics

5 Next steps:

1. Survey Design and Implementation
   - Design survey questions focusing on database status, access possibilities, and specific data characteristics
   - Select appropriate survey platform balancing accessibility and professional appearance
2. Professional Communication Strategy
   - Develop template emails for initial contact
   - Create professional presentation materials explaining the project's significance highlighting it's connection to the european comission

- Follow-up protocol for non-responders (?)

3. Survey Distribution

    - Compile contact information for database administrators and study authors
    - Implement tracking system for contact attempts and responses
    - Plan for systematic documentation of survey responses

The detailed survey design document can be found here.

Progress on these steps will be tracked and updated in this document as we move forward with the survey implementation.

6 Appendix A

6 Search string for strategy 1

6.1.1 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")*

6.1.1.1 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

6.1.2 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")*

6.1.2.1 Hits

- 25/11/2023: 284 hits

6 Search string for strategy 2

6.2.1 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")*

6.2.1.1 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

6.2.2 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")*

6.2.2.1 Hits

- 25/11/2023: 46 hits

7 References

[1] Matthews AA, Danaei G, Islam N, Kurth T. Target trial emulation: applying principles of randomised trials to observational studies. BMJ 2022:e071108. https://doi.org/10.1136/bmj-2022-071108.

[2] Hernán MA, Wang W, Leaf DE. Target Trial Emulation. JAMA 2022;328:2446. https://doi.org/10.1001/jama.2022.21383.

[3] Smit JM, Krijthe JH, Bommel J van, Labrecque JA, Komorowski M, Gommers DAMPJ, et al. Causal inference using observational intensive care unit data: A systematic review and recommendations for future practice 2022.

[4] Sauer CM, Dam TA, Celi LA, Faltys M, Hoz MAA de la, Adhikari L, et al. Systematic Review and Comparison of Publicly Available ICU Data Sets—A Decision Guide for Clinicians and Data Scientists. Critical Care Medicine 2022;50:e581–8. https://doi.org/10.1097/ccm.0000000000005517.