BMJ Health & Care Informatics

Navigating the machine learning pipeline: a scoping review of inpatient delirium prediction models

Tom Strating,[1] Leila Shafiee Hanjani,[1] Ida Tornvall,[1] Ruth Hubbard,[1] Ian A. Scott [1,2]

To cite: Strating T, Shafiee Hanjani L, Tornvall I, *et al*. Navigating the machine learning pipeline: a scoping review of inpatient delirium prediction models. *BMJ Health Care Inform* 2023;**30**:e100767. doi:10.1136/bmjhci-2023-100767

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/bmjhci-2023-100767).

Received 21 March 2023
Accepted 12 June 2023

[1]Centre for Health Services Research, The University of Queensland Faculty of Medicine, Brisbane, Queensland, Australia
[2]Internal Medicine and Clinical Epidemiology, Princess Alexandra Hospital, Woolloongabba, Queensland, Australia

Correspondence to
Professor Ian A. Scott;
ian.scott@health.qld.gov.au

ABSTRACT

Objectives Early identification of inpatients at risk of developing delirium and implementing preventive measures could avoid up to 40% of delirium cases. Machine learning (ML)-based prediction models may enable risk stratification and targeted intervention, but establishing their current evolutionary status requires a scoping review of recent literature.

Methods We searched ten databases up to June 2022 for studies of ML-based delirium prediction models. Eligible criteria comprised: use of at least one ML prediction method in an adult hospital inpatient population; published in English; reporting at least one performance measure (area under receiver-operator curve (AUROC), sensitivity, specificity, positive or negative predictive value). Included models were categorised by their stage of maturation and assessed for performance, utility and user acceptance in clinical practice.

Results Among 921 screened studies, 39 met eligibility criteria. In-silico performance was consistently high (median AUROC: 0.85); however, only six articles (15.4%) reported external validation, revealing degraded performance (median AUROC: 0.75). Three studies (7.7%) of models deployed within clinical workflows reported high accuracy (median AUROC: 0.92) and high user acceptance.

Discussion ML models have potential to identify inpatients at risk of developing delirium before symptom onset. However, few models were externally validated and even fewer underwent prospective evaluation in clinical settings.

Conclusion This review confirms a rapidly growing body of research into using ML for predicting delirium risk in hospital settings. Our findings offer insights for both developers and clinicians into strengths and limitations of current ML delirium prediction applications aiming to support but not usurp clinician decision-making.

INTRODUCTION

Delirium is a common but underdiagnosed state of disturbed attention and cognition that afflicts one in four older hospital inpatients.[1] It is independently associated with a longer length of hospital stay, mortality, accelerated cognitive decline[2] and new-onset dementia.[1] Since older people are particularly vulnerable to severe illness from COVID-19 infection, delirium emerged as a frequent acute geriatric syndrome during the pandemic.[3] Predicting who is likely to develop delirium before symptom onset may facilitate the targeted implementation of preventive strategies that can avoid up to 40% of cases.[4]

Risk stratification models enable clinicians to identify patients at high risk of an adverse event and intervene where appropriate.[5] The advent of wearables, genomics, and dynamic datasets within electronic health records (EHRs) provides big data to which machine learning (ML) can be applied to individualise clinical risk prediction.[6] ML is a subset of artificial intelligence that uses advanced computer programmes to learn patterns and associations within large datasets and develop models (or algorithms), which can then be applied to new data in rapidly producing predictions or classifications, including diagnoses.[7] Across developed nations, more than 150 ML applications are approved for use in routine clinical practice, and this number is projected to rise exponentially over the coming years.[6 8]

The key stages of the ML pipeline that models must traverse, from initial in-silico (computer-based) development to real-world deployment, comprise the following[6] (figure 1): (1) data collection; (2) data preparation; (3) feature selection and engineering; (4) model training; (5) model validation, both internal and external; (6) deployment of the model within a working application; and (7) post-deployment monitoring and optimisation of the application. During the development phase (stages 1–3), researchers collect, clean and transform data into computable formats and select relevant features as model inputs. The model is then iteratively improved through several training cycles against static, retrospective datasets (stage 4). In stage 5, the model undergoes two processes of validation: internal validation for accuracy and reproducibility against

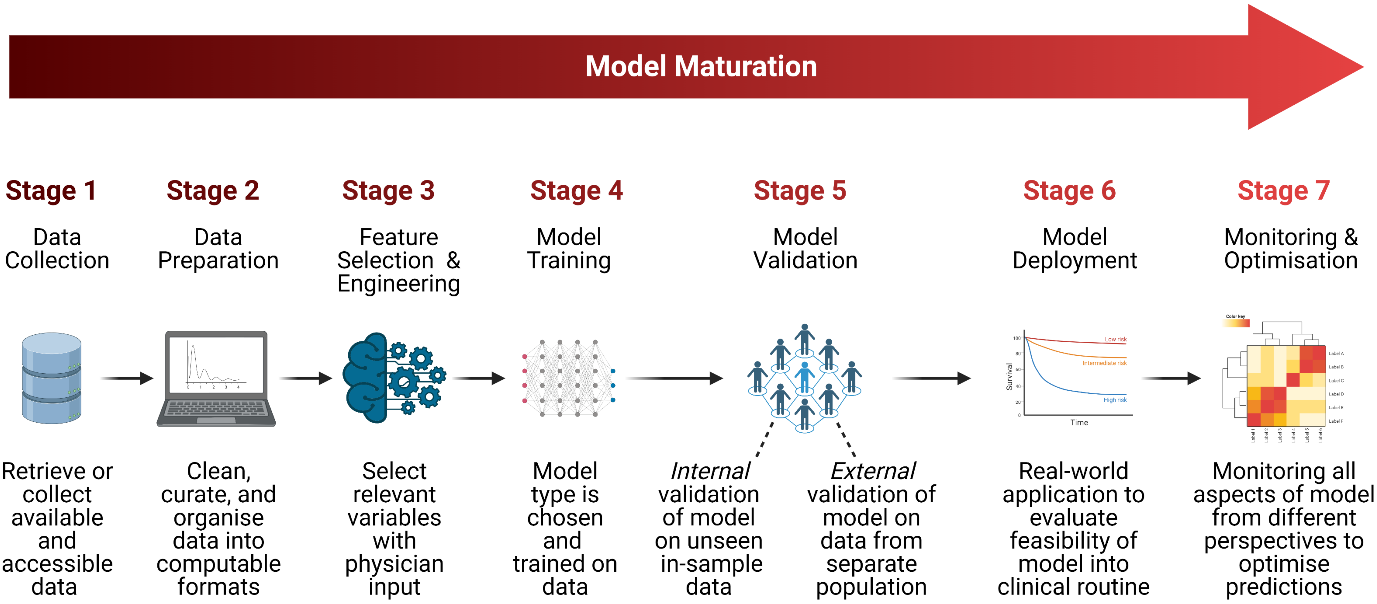

Model Maturation
Stage 1
Data Collection
Stage 2
Data Preparation
Stage 3
Feature Selection & Engineering
Stage 4
Model Training
Stage 5
Model Validation
Stage 6
Model Deployment
Stage 7
Monitoring & Optimisation
Retrieve or collect available and accessible data
Clean, curate, and organise data into computable formats
Select relevant variables with physician input
Model type is chosen and trained on data
Internal validation of model on unseen in-sample data
External validation of model on data from separate population
Real-world application to evaluate feasibility of model into clinical routine
Monitoring all aspects of model from different perspectives to optimise predictions

Figure 1 Machine learning pipeline.

a random sample from the original training dataset ('hold out' sample); and external validation, whereby researchers validate the model on a new external dataset set derived from previously unencountered patients using the same performance metrics. In stage 6, the model is subject to prospective validation using live (or near-live) dynamic data in a form reflecting its future real-world deployment, integrated into a prototype application, and evaluated for its feasibility in clinical workflows. Then, it is assessed for its clinical utility within clinical trials, which compares application-guided patient care and outcomes with the current standard of care. Finally, stage 7 entails monitoring the effectiveness and safety of the model over its life cycle using surveillance data.

ML models have enormous potential in facilitating more accurate risk stratification, preventive intervention and avoidance of incident delirium, but external validation, prospective evaluation and clinical adoption remain limited,[6] and analysis of the clinical impact of deployed models on patient care is rarely performed.[9 10] Previous systematic reviews of delirium prediction models have been limited to in-silico models focusing on performance metrics using static retrospective data,[11 12] and the studies within these reviews are limited to those published before 2019. The objectives of this review were to: (1) provide a more contemporary overview of research on all ML delirium prediction models designed for use in the inpatient setting; (2) characterise them according to their stage of development, validation and deployment; and (3) assess the extent to which their performance and utility in clinical practice have been evaluated.

METHODS

This review follows the Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews guidelines[13] and is registered within the Open Science Framework (OSF) database (*osf.io/8r5cd*). A scoping review methodology was selected as it allows us to map the broad and emerging ML evidence base in a flexible but systematic manner.[14]

Literature search

The search strategy was developed by two authors (TS, LSH) and reviewed by a third author (IAS) and a librarian. We searched PubMed, EMBASE, IEEE Xplore, Scopus, Web of Science, CINAHL, PsycInfo, Cochrane, OSF pre-prints and the aiforhealth.app machine learning research dashboard between inception and 14 June 2022, using a mixture of medical subject headings (MeSH) and keywords related to delirium and ML (for the exact search terms, see online supplemental appendix 1). Additional studies were identified by perusing the reference lists of retrieved articles.

Study selection

Retrieved studies were imported into EndNote 20 and screened for relevance and duplicates in Covidence. Two reviewers (TS, IT) independently screened the titles and abstracts, and two authors (TS, LSH) reviewed the full-text articles. Disagreements between screening authors were resolved by discussion or settled by a third reviewer (IAS). We considered full-length original studies published in peer-reviewed journals, pre-prints and conference proceedings. Eligible studies had to fulfil all the following criteria: use of at least one ML method that predicts delirium; applied to an adult hospital inpatient population; published in English; and reporting at least one of the following performance measures (area under the receiver-operator curve (AUROC), sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV). Studies were excluded if they were: editorials, position statements, letters to the editor, conference abstracts or press releases; conducted in non-hospital settings; or did not report any model performance metrics.

Data extraction and synthesis

One reviewer (TS) independently performed data extraction using a preagreed form designed in Covidence. The following data items were extracted: title; author; publication year; country (where data were collected); study aim and design; clinical setting; population characteristics; ML modelling method(s); reference standard used to diagnose delirium; frequency of delirium; data source and type; evolutionary stage and respective sample size; model performance measures (comprising, where reported, AUROC, sensitivity, specificity, PPV and NPV, Brier score, calibration plot concordance), primary outcome measures; comparison to standard care; principal discharge diagnosis; and length of stay. Qualitative information on user acceptance of deployed models was also recorded where reported.

We defined a model as being in the 'development and internal validation' stage if the dataset used for validating the model came from the same patient population as the training dataset. An 'external validation' study was where the model was validated using a dataset from a population temporally or geographically separate from that used to provide the original training data. Finally, we labelled a study as having a 'deployment'-level study was where the was evaluated in a routine clinical setting.

Corresponding authors were contacted for studies that did not report the reference standard used to define delirium in their dataset. Two authors (LSH, IT) cross-checked the data extracted for a random sample of 25% (n=10) of studies, and disagreement was managed through discussion.

A narrative approach was taken to synthesise the data extracted from the selected studies, including tabular and graphical representations, summarising the number of studies in each stage, year and country published, performance metrics, algorithm type, data type and stage of development. Descriptive statistics for continuous variables comprised mean and SD and median and IQR for normally and non-normally distributed data, respectively. All analyses and visualisations were done within R.[15] As this was a scoping review, no attempt was made to assess the quality of individual study design or methods.

RESULTS

The search strategy identified a total of 921 records; after duplicate removal and title and abstract screening, 114 full-text studies were retrieved, of which 39[16–54] met the selection criteria for inclusion in the final analysis (figure 2).

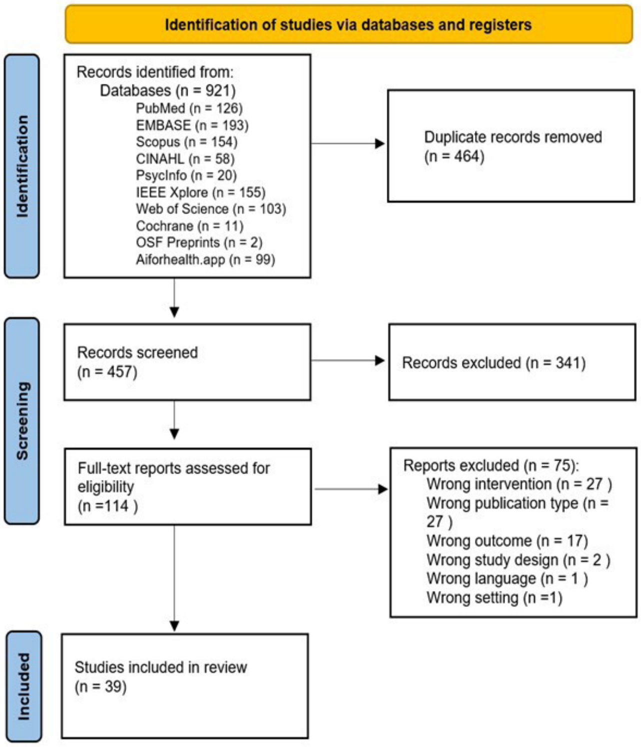

Identification of studies via databases and registers
Identification
Screening
Included
Records identified from:
Databases (n = 921)
PubMed (n = 126)
EMBASE (n = 193)
Scopus (n = 154)
CINAHL (n = 58)
PsycInfo (n = 20)
IEEE Xplore (n = 155)
Web of Science (n = 103)
Cochrane (n = 11)
OSF Preprints (n = 2)
Aiforhealth.app (n = 99)
Duplicate records removed (n = 464)
Records screened (n = 457)
Records excluded (n = 341)
Full-text reports assessed for eligibility (n =114)
Reports excluded (n = 75):
Wrong intervention (n = 27)
Wrong publication type (n = 27)
Wrong outcome (n = 17)
Wrong study design (n = 2)
Wrong language (n = 1)
Wrong setting (n =1)
Studies included in review (n = 39)

Figure 2 Preferred Reporting Items for Systematic Reviews and Meta-Analyses flow chart.

Study characteristics

Study characteristics are summarised in online supplemental table 1. Studies originated from the USA (n=12),[17,19–23,25,41,43,50,51,54] Austria (n=9),[24,28–31,33,39,47,48] China (n=6),[26,32,35,49,52,53] Germany (n=3),[37,45,46] South Korea (n=3),[27,40,44] Canada (n=3),[30,36,38] Brazil (n=1),[16] Japan (n=1),[34] Spain (n=1)[18] and one study was labelled as international.[42] Over the 6-year distribution of publications to June 2022, most studies were published in 2021 (n=10) and the first half of 2022 (n=12), indicating considerable growth in research in this area since the publication of previous reviews of studies published up to 2019.[11,12] Study design comprised retrospective cohort study (n=25), prospective cohort studies (n=9); secondary analyses of trial data (n=2), prospective pilot study (n=2) and a retrospective case-control study (n=1). Studies mostly used data from EHRs alone to develop their models (n=21), with the remainder including specified clinical assessments (eg, nursing assessment, n=8), compiled clinical databases (eg, data repository or open-access database, n=6), data from a clinical quality improvement registry (n=1), data from both EHRs and clinical assessments (n=1), data from EHRs and a clinical database (n=1) and data solely from electrocardiographs (n=1).

The median (IQR) sample size of training datasets was 2389 (IQR: 371–27,377) participants, of whom, when reported as a percentage, a median of 20% (IQR: 20%–25%) was used as a 'hold-out' sample for internal validation. External validation and deployment studies had a median of 4765 (IQR: 2429–11 355) and 5887 (IQR: 3456–10 975) participants, respectively. The age of participants ranged from a mean of 54.4–84.4 years. Hospital inpatients were treated in surgical wards (n=14), medical wards (n=10), intensive care units (ICU) (n=7) or a combination of all three settings (n=8). The reported reference standards for verifying delirium cases in the training dataset comprised the confusion assessment method for the Intensive Care Unit (CAM-ICU) (n=10), International Classification of Diseases codes (n=14), the CAM (n=7) and the Diagnostic Statistical Manual (n=3). Several alternative screening methods, such as the 4 A's Test (n=2), were used infrequently, and three studies reported no information as to what reference standard was used. The prevalence of delirium in training and internal validation datasets ranged from 2.0% to 53.6%, and from 10% to 39% in external validation studies. Delirium prevalence was 1.5%[28] and 31.2%[31] for the two deployment studies that reported data on this outcome. Length of stay ranged from an average of 1.9–13.6 days, but was not reported in 27 (69%) of studies.

Model characteristics

Thirty of thirty-nine publications described the training and internal validation of a delirium model,[17 18 21–26 30 32–41 43 44 46–54] with investigators of 6 of these studies (20%) externally validating their model in a subsequent paper.[16 19 20 27 29 42] Investigators of three studies (10%) implemented and evaluated their model in real-time clinical workflows,[2 8 31 45] but no publications described monitoring or optimising a deployed model.

Figure 3 depicts the numbers of publications that used each type of model across each stage of application maturity. In total, random forest models were the most common (n=11), followed by logistic regression (n=6), gradient boosting (n=5) and artificial neural networks (n=4). Two other papers each described using a decision tree, L1-penalised regression, or natural language processing models, with another seven papers describing different models unique to the study.

Performance metrics of each model at their different stages of validation, when reported, are listed in online supplemental table 2. In the absence of any universal task-agnostic standard, we regarded values of AUROC>0.7, of sensitivity and specificity ≥80%, of PPV ≥30% and NPV ≥90%, of Brier scores <0.20 and calibration plots showing high concordance as being acceptable accuracy thresholds for clinical application. For internal validation, omitting two studies for which the AUROC statistic was not reported,[40 44] the median AUROC for the remaining models was 0.85 (IQR: 0.78–0.90). For external validation

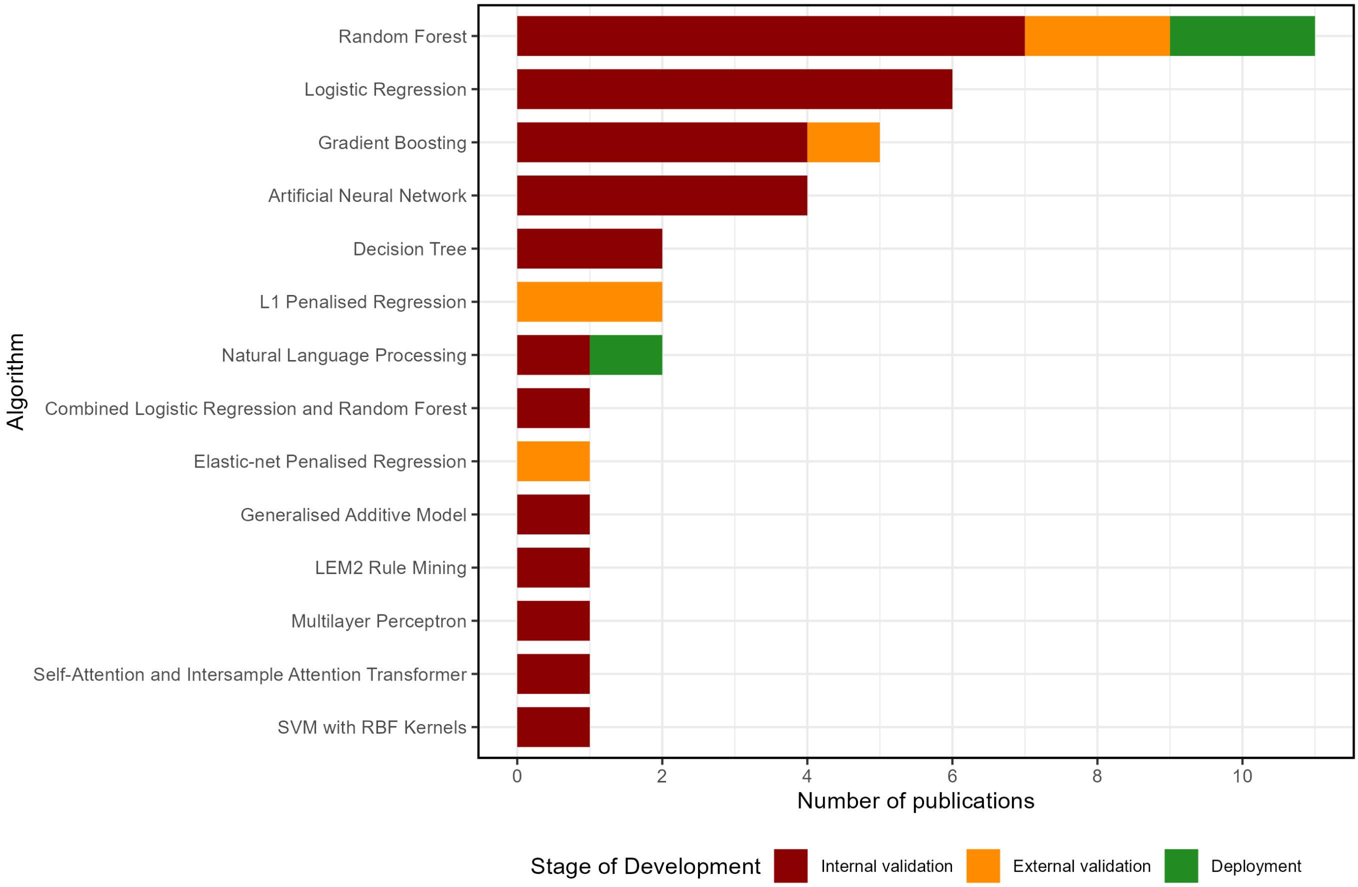


Figure 3 Number of publications by machine learning method. If a study describes multiple models, only the best-performing (area under receiver-operator curve) model is shown. LEM, learning from examples module 2; LR, logistic regression; RBF, radial basis function; RF, random forest; SAINTENS, self-attention and intersample attention transformer; SVM, support vector machine.

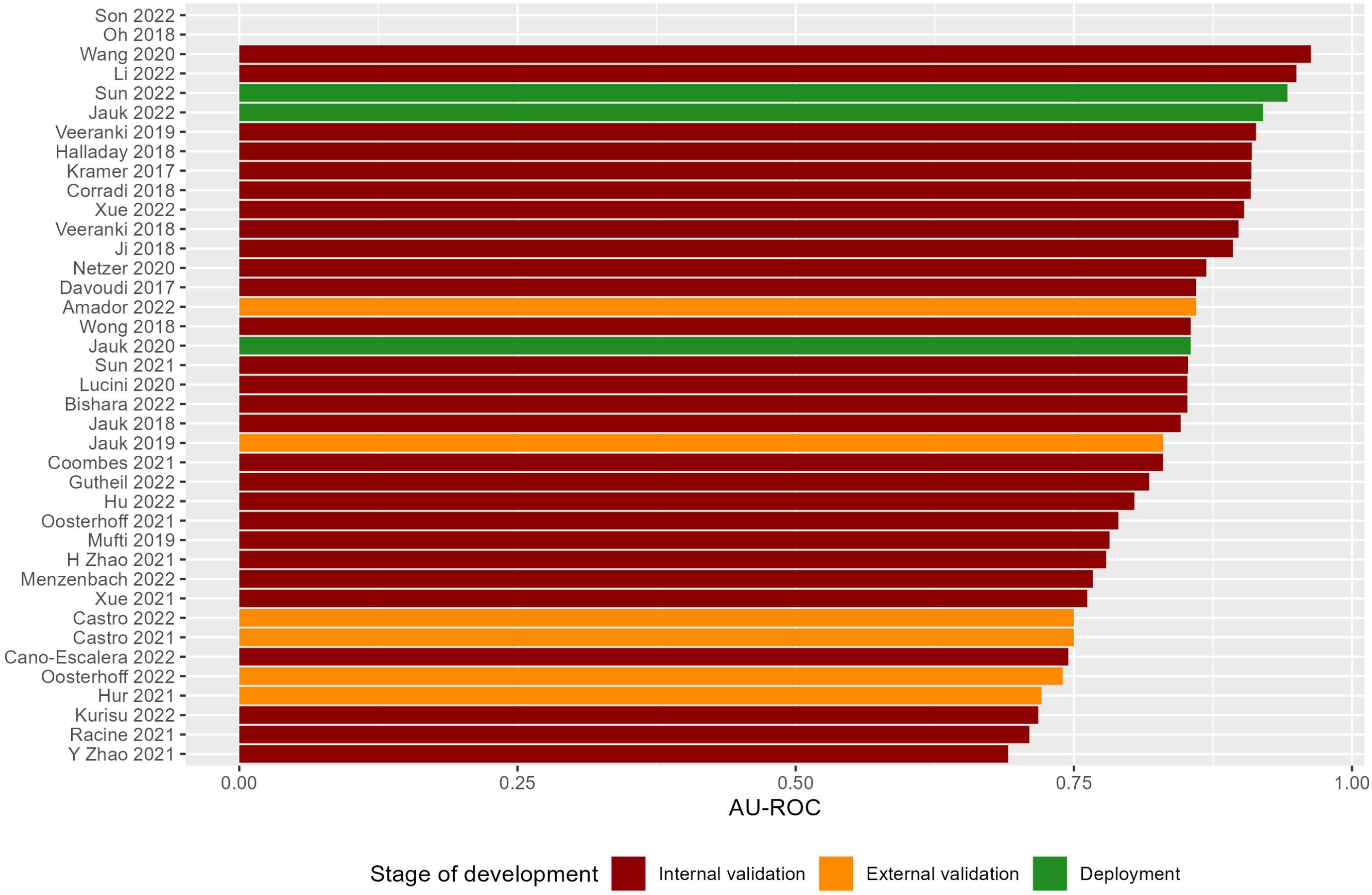


Figure 4 Graphical representation of AUROC performance metrics stratified by stage of development. Son *et al*[44] andOh *et al*[40] did not report AUROC but are included in the analysis as they reported other performance metrics. AUROC, area under receiver-operator curve.

and deployment studies, the reported median AUROC scores were 0.75 (IQR: 0.74–0.81) and 0.92 (IQR: 0.89–0.93), respectively.

Stratified by algorithm type, the median AUROC (models with >1 publication) for training and internal validation studies was highest for random forest models (0.91, IQR: 0.88–0.91). In order of decreasing performance were natural language processing (AUROC: 0.85, IQR: 0.83–0.91); decision trees (AUROC: 0.83, IQR: 0.78–0.89); artificial neural networks (AUROC: 0.81, IQR: 0.76–0.86); gradient boosting (AUROC: 0.81, IQR: 0.77–0.85); artificial neural networks (AUROC: 0.81, IQR: 0.75–0.87) and logistic regression models (AUROC: 0.80, IQR: 0.78–0.82).

In regards to external validation, a gradient boosting algorithm performed best (AUROC: 0.86), followed by random forest models (AUROC: 0.78, IQR: 0.75–0.80) and L1-penalised regression (AUROC: 0.75, IQR: 0.75–0.75). For prospective studies of deployed models, the best performance was observed in one study using natural language processing, with an AUROC score of 0.94,[45] with random forest models achieving a median AUROC score of 0.89 (IQR: 0.87–0.90). The AUROC performance metrics for all models, stratified by stage of maturity, is presented graphically in figure 4.

The median sensitivity and specificity for training studies were 75% (IQR: 64.1%–82.3%) and 82.2% (73.3%–90.4%), respectively. For external validation studies, median sensitivity and specificity dropped to 73% (IQR: 67.5%–81.7%) and 69% (IQR: 48%–72%), respectively. However, in deployment-level studies, median sensitivity and specificity were 87.1% (IQR: 80.6%–93.5%) and 86.4% (IQR: 84.3%–88.5%), respectively. The PPV and NPVs of included ML models were only reported for 10 of 39 studies (26%), which ranged respectively from 5.8% to 91.6% and from 90.6% to 99.5%.

Of the total, only 14 studies (35.9%) reported calibration metrics which showed considerable variation. Using calibration plots, four studies reported poor calibration, an equal number reported reasonable calibration, while the remainder employed alternative calibration methods with variable results (see online supplemental table 1). The Brier score was reported for only five studies (13%) and ranged from 0.14 to 0.22.

Clinical application

Three articles from two investigators subjected their prototype model to prospective validation using live data in a form reflecting its future application to clinical workflows.[28 31 45] Sun *et al*[45] trained three separate models to predict delirium, acute kidney injury and sepsis. They found their delirium model performed slightly worse using live data from three hospitals at admission (AUROC decreased by 3.6%) and when deployed in another participating hospital with data separate to that of the training set, performance dropped by another 0.8% at discharge. Sun *et al* reported user feedback only for the acute kidney injury model.

Jauk *et al*[28] implemented their delirium prediction model in an Austrian hospital system for 7 months and thereafter for an additional month in the trauma surgery department of another affiliated hospital.[31] The prediction model performed somewhat worse on prospective data (AUROC: 0.86) as it did on retrospective data used in the training and internal validation study[33] (AUROC: 0.91). In addition, predictions of the random forest model used in this study correlated strongly with nurses' ratings of delirium risk in a sample of internal medicine patients (correlation coefficient (r)=0.81 for blinded and r=0.62 for non-blinded comparison). In the external validation study, the model achieved an AUROC value above 0.85 across three prediction times (on admission: 0.863; first evening: 0.851; second evening: 0.857). However, when the model was re-trained using local data, the AUROC value exceeded 0.92 for all three prediction times, and correctly predicted all 29 patients who were deemed high risk for delirium by a senior physician (sensitivity=100%, specificity=90.6%). In a qualitative survey, the 13 health professionals involved in the project perceived the ML application as useful and easy to use.

DISCUSSION

This scoping review examined contemporary research around ML models for predicting delirium in adult inpatient settings and identified an additional 22 studies published since late 2019 which was the finish date for previous reviews.[11 12] We have mapped the development and implementation stage and associated performance metrics of these new models according to a six-stage evolutionary ML pipeline. Importantly, we included three novel implementation studies which demonstrated good predictive accuracy and user acceptance, underscoring the potential clinical utility of ML models for delirium prediction.

However, our review reveals several limitations in the existing research that future studies need to address. First, training data in most studies comprised routinely collected data obtained retrospectively from EHRs which, while providing vast quantities of data for training complex models, suffer from inaccuracies and omissions relating to key predictor variables. Only a quarter of studies[18 26 28 29 31 32 35 37 40 44 45] in this review sourced prespecified and prospectively collected data, such that missing or incomplete data relevant to model optimisation, and which could not be remedied using imputation methods, emerged as a critical limitation for many studies. For instance, the EHR-derived models of Zhao *et al*[53] lacked microbiological, radiological and biomarker data relevant to delirium, limiting their predictive accuracy. Similarly, missing information about medication use and frailty indices posed a limiting factor in several other studies.[17 22 49–51] Many studies also did not have access to demographic data of their study population, such as socioeconomic status, gender and race.[26 53] Reliance on data sources with missing data and unrepresentative of target populations weakens model performance and introduces biases, generating models that may exacerbate healthcare inequities.[7]

Second, similar to the findings of previous reviews,[11 12] most models described in our scoping review did not mature past the stage of internal validation. Only six studies validated their model on an external dataset[16 19 20 27 29 42] despite evidence that models that perform well on ‘hold out’ training data usually have lower performance when applied to more noisy datasets from different institutions due to model overfitting.[5]

Third, of all 39 included studies, only those of Jauk *et al*[28 31] and Sun *et al*[45] subjected their models to a prospective evaluation using live data in clinical practice. The extent to which clinicians will adopt a model depends on their trust in its predictive accuracy and utility and the ease with which it can be integrated into clinical workflows.[7] Sun and colleagues[45] demonstrated their deep learning model performed equally well in training and prospective validation studies.[29] In a subsequent case study, the authors demonstrated an instance where their application correctly predicted postoperative delirium in a patient with a negative preoperative CAM-ICU, demonstrating its clinical utility in a surgical ward.[55] In addition, they found ML applications could be particularly useful for the early detection of delirium in wards where delirium screening is often not performed and delirium is underdiagnosed.[1]

Similarly, Jauk and colleagues[28] analysed 5530 predictions over 7 months of deployment, finding their model performance was reliable and attracted high satisfaction ratings by a senior physician. In a later qualitative study, the 47 nurses and physicians associated with the project rated the delirium prediction model as useful, easy to use and interpretable without increasing workload.[56] These favourable findings were replicated in a follow-up study where the random forest model was implemented in a separate hospital network.[31] However, cross-hospital evaluations underscored the need to re-train the model with local data to mitigate declines in performance when applied to new clinical settings.[31 45] However, neither of these models has been subjected to clinical trials to establish impacts on patient care or outcomes.

Our review has some limitations. As our study was a scoping exercise, and in the absence of an agreed risk of bias assessment tool for ML prediction studies, we chose not to critically appraise the quality of individual studies. For similar reasons, and given the heterogeneity of the data source, model type and performance metrics reported in included studies, quantitative meta-analysis was not performed.

CONCLUSION

Prediction models derived using ML methods can potentially identify individuals at risk of developing delirium before symptom onset to whom preventive strategies can be targeted, which may, in turn, reduce incident delirium and improve patient outcomes. This scoping review identified all publications describing ML-based delirium prediction models over the last 5 years, evaluated their stage in the ML evolution pipeline, and assessed their performance and utility. Relatively few were subject to external validation, which, when performed, showed degraded model performance. In addition, while few studies underwent prospective evaluation in real-world clinical settings, performance and user acceptance seemed promising in those that did. However, given the limitations of current delirium prediction models, they should not be seen as substitutes for expert clinician judgement.

Acknowledgements We thank librarian Nicole Rayner for helping to construct the search strategy.

Contributors All authors contributed to the conception and design of the work. TS, LSH and IT contributed to database searching. TS, LSH, IT and IAS contributed to full-text screening. TS, LSH, IT and IAS contributed to data extraction and analysis. TS and IAS drafted the paper. All authors commented on and helped revise iterative drafts. IAS is the guarantor who accepts full responsibility for the finished article and had access to all data from literature searches. All authors gave final approval for the manuscript to be published and agreed to be accountable for all aspects of the work.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement All data relevant to the study are included in the article or uploaded as supplementary information.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

ORCID iD

Ian A. Scott http://orcid.org/0000-0002-7596-0837

REFERENCES

1 Richardson SJ, Davis DHJ, Stephan BCM, *et al*. Recurrent delirium over 12 months predicts dementia: results of the delirium and cognitive impact in dementia (DECIDE) study. *Age and Ageing* 2021;50:914–20.

2 Han JH, Shintani A, Eden S, *et al*. Delirium in the emergency department: an independent predictor of death within 6 months. *Ann Emerg Med* 2010;56:244–52.

3 Inouye SK. The importance of delirium and delirium prevention in older adults during lockdowns. *JAMA* 2021;325:1779–80.

4 Inouye SK, Westendorp RGJ, Saczynski JS. Delirium in elderly people. *Lancet* 2014;383:911–22.

5 Haendel MA, Chute CG, Robinson PN. Classification, ontology, and precision medicine. *N Engl J Med* 2018;379:1452–62.

6 Scott IA. Demystifying machine learning: a primer for physicians. *Intern Med J* 2021;51:1388–400.

7 Scott IA, Carter S, Coiera E. Clinician checklist for assessing suitability of machine learning applications in healthcare. *BMJ Health Care Inform* 2021;28:e100251.

8 Muehlematter UJ, Daniore P, Vokinger KN. Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015-20): a comparative analysis. *Lancet Digit Health* 2021;3:e195–203.

9 Scott I, Cook D, Coiera E. Evidence-based medicine and machine learning: a partnership with a common purpose. *BMJ Evid Based Med* 2021;26:290–4.

10 Goldstein BA, Navar AM, Pencina MJ, *et al*. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *J Am Med Inform Assoc* 2017;24:198–208.

11 Chua SJ, Wrigley S, Hair C, *et al*. Prediction of delirium using data mining: a systematic review. *J Clin Neurosci* 2021;91:288–98.

12 Ruppert MM, Lipori J, Patel S, *et al*. ICU delirium-prediction models: a systematic review. *Crit Care Explor* 2020;2:e0296.

13 Tricco AC, Lillie E, Zarin W, *et al*. PRISMA extension for scoping reviews (PRISMA-SCR): checklist and explanation. *Ann Intern Med* 2018;169:467–73.

14 Munn Z, Peters MDJ, Stern C, *et al*. Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. *BMC Med Res Methodol* 2018;18:143.

15 Team, RC. A language and environment for statistical computing. In: *R foundation for statistical computing*. Vienna, Austria, 2013. Available: http://www.R-project.org

16 Amador T, Saturnino S, Veloso A, *et al*. Early identification of ICU patients at risk of complications: regularization based on robustness and stability of explanations. *Artif Intell Med* 2022;128:102283.

17 Bishara A, Chiu C, Whitlock EL, *et al*. Postoperative delirium prediction using machine learning models and preoperative electronic health record data. *BMC Anesthesiol* 2022;22:8.

18 Cano-Escalera G, Graña M, Irazusta J, *et al*. Risk factors for prediction of delirium at hospital admittance. *Expert Systems* 2022;39:e12698.

19 Castro VM, Hart KL, Sacks CA, *et al*. Longitudinal validation of an electronic health record delirium prediction model applied at admission in COVID-19 patients. *Gen Hosp Psychiatry* 2022;74:9–17.

20 Castro VM, Sacks CA, Perlis RH, *et al*. Development and external validation of a delirium prediction model for hospitalized patients with Coronavirus disease 2019. *J Acad Consult Liaison Psychiatry* 2021;62:298–308.

21 Coombes CE, Coombes KR, Fareed N. A novel model to label delirium in an intensive care unit from clinician actions. *BMC Med Inform Decis Mak* 2021;21:97.

22 Corradi JP, Thompson S, Mather JF, *et al*. Prediction of incident delirium using a random forest Classifier. *J Med Syst* 2018;42:261.

23 Davoudi A, Ebadi A, Rashidi P, *et al*. Delirium prediction using machine learning models on preoperative electronic health records data. Proc IEEE Int Symp Bioinformatics Bioeng 2017;2017:568–73.

24 Gutheil J, Donsa K. SAINTENS: self-attention and Intersample attention transformer for digital biomarker development using tabular healthcare real world data. *Stud Health Technol Inform* 2022;293:212–20.

25 Halladay CW, Sillner AY, Rudolph JL. Performance of electronic prediction rules for prevalent delirium at hospital admission. *JAMA Netw Open* 2018;1:e181405.

26 Hu X-Y, Liu H, Zhao X, *et al*. Automated machine learning-based model predicts postoperative delirium using readily extractable perioperative collected electronic data. *CNS Neurosci Ther* 2022;28:608–18.

27 Hur S, Ko RE, Yoo J, *et al*. A machine learning-based algorithm for the prediction of intensive care unit delirium (PRIDE): retrospective study. *JMIR Med Inform* 2021;9:e23401.

28 Jauk S, Kramer D, Großauer B, *et al*. Risk prediction of delirium in hospitalized patients using machine learning: an implementation and prospective evaluation study. *J Am Med Inform Assoc* 2020;27:1383–92.

29 Jauk S, Kramer D, Quehenberger F, *et al*. Information adapted machine learning models for prediction in clinical workflow. *Stud Health Technol Inform* 2019;260:65–72.

30 Jauk S, Kramer D, Schulz S, *et al*. Evaluating the impact of incorrect diabetes coding on the performance of multivariable prediction models. *Stud Health Technol Inform* 2018;251:249–52.

31 Jauk S, Veeranki SPK, Kramer D, *et al*. External validation of a machine learning based delirium prediction software in clinical routine. *Stud Health Technol Inform* 2022;293:93–100.

32 Ji M, Xing S, Yang Y. Pathophysiological factors of delirium among critically ill elders after non-cardiac surgery based on artificial neural networks: a pilot study. *Anaesthesia, Pain and Intensive Care* 2018;22:424–30.

33 Kramer D, Veeranki S, Hayn D, *et al*. Development and validation of a multivariable prediction model for the occurrence of delirium in hospitalized gerontopsychiatry and internal medicine patients. *Stud Health Technol Inform* 2017;236:32–9.

34 Kurisu K, Inada S, Maeda I, *et al*. A decision tree prediction model for a short-term outcome of delirium in patients with advanced cancer receiving pharmacological interventions: a secondary analysis of a multicenter and prospective observational study (phase-R). *Palliat Support Care* 2022;20:153–8.

35 Li Q, Zhao Y, Chen Y, *et al*. Developing a machine learning model to identify delirium risk in geriatric internal medicine inpatients. *Eur Geriatr Med* 2022;13:173–83.

36 Lucini FR, Fiest KM, Stelfox HT, *et al*. Delirium prediction in the intensive care unit: a temporal approach. *Annu Int Conf IEEE Eng Med Biol Soc* 2020;2020:5527–30.

37 Menzenbach J, Kirfel A, Guttenthaler V, *et al*. Pre-operative prediction of postoperative delirium by appropriate screening (PROPDESC) development and validation of a pragmatic POD risk screening score based on routine preoperative data. *J Clin Anesth* 2022;78:110684.

38 Mufti HN, Hirsch GM, Abidi SR, *et al*. Exploiting machine learning models and methods for the prediction of agitated delirium after cardiac surgery: models development and validation study. *JMIR Med Inform* 2019;7:e14993.

39 Netzer M, Hackl WO, Schaller M, *et al*. Evaluating performance and Interpretability of machine learning methods for predicting delirium in gerontopsychiatric patients. *Stud Health Technol Inform* 2020;271:121–8.

40 Oh J, Cho D, Park J, *et al*. Prediction and early detection of delirium in the intensive care unit by using heart rate variability and machine learning. *Physiol Meas* 2018;39:035004.

41 Oosterhoff JHF, Karhade AV, Oberai T, *et al*. Prediction of postoperative delirium in geriatric hip fracture patients: a clinical prediction model using machine learning models. *Geriatr Orthop Surg Rehabil* 2021;12:21514593211062277.

42 Oosterhoff JHF, Oberai T, Karhade AV, *et al*. Does the SORG orthopaedic research group hip fracture delirium algorithm perform well on an independent intercontinental cohort of patients with hip fractures who are 60 years or older *Clin Orthop Relat Res* 2022;480:2205–13.

43 Racine AM, Tommet D, D'Aquila ML, *et al*. Machine learning to develop and internally validate a predictive model for post-operative delirium in a prospective, observational clinical cohort study of older surgical patients. *J Gen Intern Med* 2021;36:265–73.

44 Son CS, Kang WS, Lee JH, *et al*. Machine learning to identify psychomotor behaviors of delirium for patients in long-term care facility. *IEEE J Biomed Health Inform* 2022;26:1802–14.

45 Sun H, Depraetere K, Meesseman L, *et al*. Machine learning-based prediction models for different clinical risks in different hospitals: evaluation of live performance. *J Med Internet Res* 2022;24:e34295.

46 Sun H, Depraetere K, Meesseman L, *et al*. A Scalable approach for developing clinical risk prediction applications in different hospitals. *J Biomed Inform* 2021;118:103783.

47 Veeranki SPK, Hayn D, Jauk S, *et al*. An improvised classification model for predicting delirium. *Stud Health Technol Inform* 2019;264:1566–7.

48 Veeranki SPK, Hayn D, Kramer D, *et al*. Effect of nursing assessment on predictive delirium models in hospitalised patients. *Stud Health Technol Inform* 2018;248:124–31.

49 Wang Y, Lei L, Ji M, *et al*. Predicting postoperative delirium after Microvascular decompression surgery with machine learning. *J Clin Anesth* 2020;66:109896.

50 Wong A, Young AT, Liang AS, *et al*. Development and validation of an electronic health record-based machine learning model to estimate delirium risk in newly hospitalized patients without known cognitive impairment. *JAMA Netw Open* 2018;1:e181018.

51 Xue B, Li D, Lu C, *et al*. Use of machine learning to develop and evaluate models using preoperative and intraoperative data to identify risks of postoperative complications. *JAMA Netw Open* 2021;4:e212240.

52 Xue X, Chen W, Chen X. A novel Radiomics-based machine learning framework for prediction of acute kidney injury-related delirium in patients who underwent cardiovascular surgery. *Comput Math Methods Med* 2022;2022:4242069.

53 Zhao H, You J, Peng Y, *et al*. Machine learning algorithm using electronic chart-derived data to predict delirium after elderly hip fracture surgeries: a retrospective case-control study. *Front Surg* 2021;8:634629.

54 Zhao Y, Luo Y. Unsupervised learning to Subphenotype delirium patients from electronic health records. Zhao Y, Luo Y, eds. 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); Houston, TX, USA.BIBM, 2021

55 Fliegenschmidt J, Hulde N, Preising MG, *et al*. Artificial intelligence predicts delirium following cardiac surgery: a case study. *J Clin Anesth* 2021;75:110473.

56 Jauk S, Kramer D, Avian A, *et al*. Technology acceptance of a machine learning algorithm predicting delirium in a clinical setting: a mixed-methods study. *J Med Syst* 2021;45:52.

BMJ Health & Care Informatics

Adherence of randomised controlled trials using artificial intelligence in ophthalmology to CONSORT-AI guidelines: a systematic review and critical appraisal

Niveditha Pattathil ,[1] Jonathan Z L Zhao,[2] Olapeju Sam-Oyerinde,[3] Tina Felfeli [4,5]

To cite: Pattathil N, Zhao JZL, Sam-Oyerinde O, *et al*. Adherence of randomised controlled trials using artificial intelligence in ophthalmology to CONSORT-AI guidelines: a systematic review and critical appraisal. *BMJ Health Care Inform* 2023;**30**:e100757. doi:10.1136/bmjhci-2023-100757

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/bmjhci-2023-100757).

NP and JZLZ are joint first authors.

Received 01 March 2023
Accepted 15 June 2023

For numbered affiliations see end of article.

Correspondence to
Dr Tina Felfeli;
tina.felfeli@mail.utoronto.ca

ABSTRACT

Purpose Many efforts have been made to explore the potential of deep learning and artificial intelligence (AI) in disciplines such as medicine, including ophthalmology. This systematic review aims to evaluate the reporting quality of randomised controlled trials (RCTs) that evaluate AI technologies applied to ophthalmology.

Methods A comprehensive search of three relevant databases (EMBASE, Medline, Cochrane) from 1 January 2010 to 5 February 2022 was conducted. The reporting quality of these papers was scored using the Consolidated Standards of Reporting Trials-Artificial Intelligence (CONSORT-AI) checklist and further risk of bias was assessed using the RoB-2 tool.

Results The initial search yielded 2973 citations from which 5 articles satisfied the inclusion/exclusion criteria. These articles featured AI technologies applied to diabetic retinopathy screening, ophthalmologic education, fungal keratitis detection and paediatric cataract diagnosis. None of the articles reported all items in the CONSORT-AI checklist. The overall mean CONSORT-AI score of the included RCTs was 53% (range 37%–78%). The individual scores of the articles were 37% (19/51), 39% (20), 49% (25), 61% (31) and 78% (40). All articles were scored as being moderate risk, or 'some concerns present', regarding potential risk of bias according to the RoB-2 tool.

Conclusion A small number of RCTs have been published to date on the applications of AI in ophthalmology and vision science. Adherence to the 2020 CONSORT-AI reporting guidelines is suboptimal with notable reporting items often missed. Greater adherence will help facilitate reproducibility of AI research which can be a stimulus for more AI-based RCTs and clinical applications in ophthalmology.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Within the field of ophthalmology, there is a growing interest in exploring the potential of deep learning and artificial intelligence (AI), however, the level of quality of the randomised controlled trials (RCTs) currently published on the efficacy of AI-driven interventions is not known.

WHAT THIS STUDY ADDS

⇒ This systematic review aimed to characterise the RCTs using AI within the field of ophthalmology and vision science, and to critically appraise the adherence of each included study to the Consolidated Standards of Reporting Trials-Artificial Intelligence (CONSORT-AI) reporting guideline.

⇒ A small number of RCTs have been published to date on the applications of AI in ophthalmology and vision science, and adherence to the 2020 CONSORT-AI reporting guidelines is suboptimal with notable reporting items often missed.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ Further studies should aim for greater adherence to reporting standards to help facilitate reproducibility and generalisability of AI research and clinical applications in ophthalmology.

INTRODUCTION

The growing advent of artificial intelligence (AI) has sparked interest globally across all fields of medicine and healthcare.[1] Within ophthalmology, AI has been used in the analysis of fundus photographs, visual field testing, optical coherence tomography and surgical skill assessment.[1] It has also been applied to improve the efficiency and robustness of detection of conditions including diabetic retinopathy,[2] retinopathy of prematurity,[3] glaucoma,[4] macular oedema[5] and age-related macular degeneration.[6] However, further expansion of AI into clinical practice requires extensive research and development.

Randomised controlled trials (RCTs) are considered the gold standard experimental design for researchers seeking to provide evidence to support the safety and efficacy of a new intervention.[7] However, deficiencies in reporting clarity can interfere with accurate assessment of sources of potential bias arising from inadequacies in methodologies. The Consolidated Standards of Reporting Trials (CONSORT) statement provides the minimum guidelines for reporting randomised trials, and its use has been key in ensuring transparency in the assessment of new interventions. It was originally published in 1996,[8] revised in 2001,[9] and most recently updated in 2010.[10] A 2001 review of 24 RCTs in ophthalmology found that on average, only 33.4 out of 57 descriptors were reported adequately according to the 1996 CONSORT guidelines.[11] A 2014 review that assessed the compliance of 65 ophthalmological RCTs published in 2011 with the 2008 CONSORT extension for non-pharmacological treatment interventions reported a mean CONSORT score of 8.9 out 23 criteria, or 39%.[12]

The Consolidated Standards of Reporting Trials - Artificial Intelligence (CONSORT-AI) extension is the reporting guideline for clinical trials evaluating interventions with an AI component, published in 2020.[13] The extension includes 14 new items considered sufficiently relevant with regard to evaluation of reporting requirements for methodologies in RCTs involving assessment of AI as the intervention.[13] With the recent rise in new initiatives using AI, adherence to reporting guidelines such as CONSORT-AI plays a critical role in guiding and standardising the conduct and reporting of AI-related trials.

This systematic review aimed to characterise the RCTs using AI within the field of ophthalmology and vision science, and to critically appraise the adherence of each included study to the CONSORT-AI reporting guideline.

METHODS

Search strategy

This systematic review was conducted in accordance to the Preferred Reporting Items for a Systematic Reviews and Meta-analyses guidelines. The protocol was prospectively registered in PROSPERO (registration number: CRD42022304021). A comprehensive search of the relevant databases MEDLINE, EMBASE, Cochrane Central Register of Controlled Trials and Cochrane Database of Systematic Reviews was done in consultation with an experienced librarian. All English-language RCTs using AI within the field of ophthalmology and vision science from 1 January 2010 to 5 February 2022 were identified. This restriction in publication date was put in place to capture the most relevant and recent publications in light of the increasing interest in research on AI following the advent and popularisation of the computing technique ‘deep learning’, especially with regard to image analysis.[14] A combination of keywords and Medical Subject Headings related to concepts of RCTs, ophthalmology and AI were used to build the search strategy (online supplemental appendix 1).

Study selection and data extraction

Two authors (NP and JZLZ) independently conducted an initial title-abstract screening followed by full-text screening of all articles. All conflicts were resolved by consensus and in consultation with a third reviewer (TF or OS). The inclusion criteria were: articles that were (1) RCTs, (2) using AI as their main intervention and (3) evaluating the AI for application within any aspects in the field of ophthalmology. Articles were excluded if they were (1) not specific to ophthalmology and/or (2) were not available in English. The authors of articles whose full-text was not available were contacted to request full-text versions directly. Data from the final set of articles included in the review were extracted and recorded in a predetermined datasheet by two authors (NP and JZLZ).

Risk of bias assessment

Risk of bias assessment was completed for each study by two independent reviewers (NP and JZLZ) using the RoB-2 tool.[15] Any conflicts were resolved in consultation with a third reviewer (TF or OS). For each domain, the risk of bias was reported as ‘high’, ‘low’ or ‘some concerns present’.

CONSORT-AI checklist

The final articles were scored independently by two authors (NP and JZLZ) using the CONSORT-AI checklist.[13] Based on previously published methods, articles were scored 1 for an item if all of the components identified in the respective criterion were reported, and 0 if any portions were missing.[12 16–18] There are 51 criteria in the CONSORT-AI checklist. Each item was given equal weight, scoring 1 point each. The resulting mark was termed the ‘CONSORT-AI score.’ The criterion regarding providing an explanation of any interim analyses and/or stopping guidelines if applicable (7b) was not applicable to any of the articles and was therefore scored as ‘0’ for all. After initial scoring, any discrepancies were resolved by consensus. If an agreement could not be reached, a third author (TF or OS) was consulted to make the final decision.

RESULTS

The search strategy yielded a total of 2973 citations (figure 1). Following deduplication and screening, five articles met the inclusion and exclusion criteria. The characteristics of the included articles are summarised in online supplemental table 1. The final articles included in this review looked at the utility of AI in diabetic retinopathy screening,[19 20] ophthalmologic education,[21] detecting fungal keratitis[22] and diagnosing childhood cataracts.[23] Three out of the five included articles were studies conducted in China, and the remaining two were conducted in Mexico and Rwanda. The majority (3/5) of the articles were published in 2021 or 2022,[19 20 22] and the remaining two were published in 2019 and 2020.[21 23]

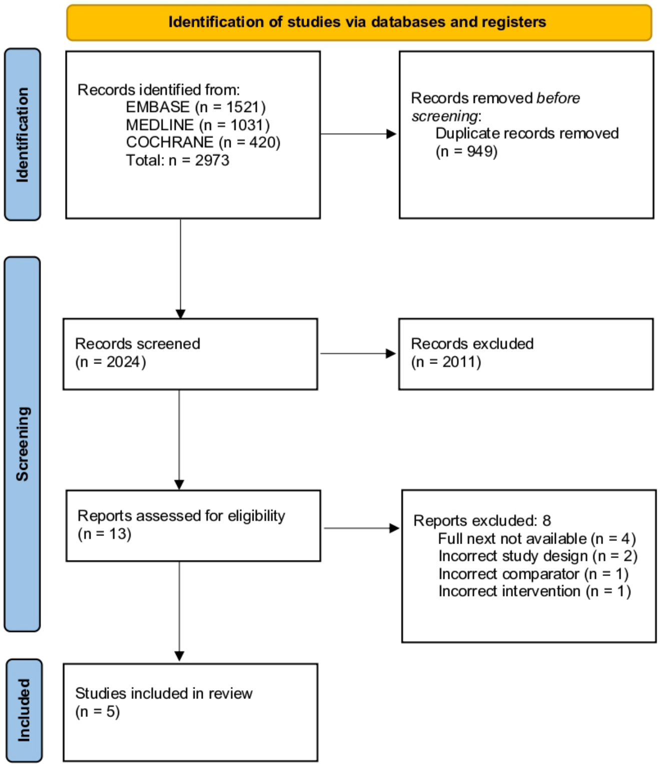

Identification of studies via databases and registers
Identification
Records identified from:
EMBASE (n = 1521)
MEDLINE (n = 1031)
COCHRANE (n = 420)
Total: n = 2973
Records removed before screening:
Duplicate records removed (n = 949)
Screening
Records screened (n = 2024)
Records excluded (n = 2011)
Reports assessed for eligibility (n = 13)
Reports excluded: 8
Full next not available (n = 4)
Incorrect study design (n = 2)
Incorrect comparator (n = 1)
Incorrect intervention (n = 1)
Included
Studies included in review (n = 5)

Figure 1 PRISMA, Preferred Reporting Items for a Systematic Reviews and Meta-analyses flow chart diagram for study identification and selection.[30]

The overall mean CONSORT-AI score of the included RCTs was 53% (range 37%–78%), and the median score was 49%. The individual scores of the articles were 19/51 (37%), 20/51 (39%), 25/51 (49%), 31/51 (61%) and 40/51 (78%). Following the initial round of scoring, there was conflict on 14 items (5.49%). The inter-rater concordance for the CONSORT-AI scoring had a kappa score of 0.89.

The compliance of the included articles to each of the individual CONSORT-AI criteria is shown in online supplemental table 2. None of the articles addressed the following criteria: important changes to methods after trial commencement with reasons (3b), changes to trial outcomes after trial commenced with reasons (6b), information on why the trial ended or was stopped (14b), important harms or unintended effects in each group,[19] analysis of performance errors and how errors were identified (19-i), and where the full trial protocol can be accessed.[24] Only one of the articles addressed the following criteria: information on which version of the AI algorithm was used (5-i), whether there was human–AI interaction in the handling of the input data and what level of expertise was required of users (5-iv), mechanism used to implement random allocation sequence,[9] who generated random allocation sequence, who enrolled participants and who assigned participants to interventions,[10] methods for additional analyses (12b), presentation of both absolute and relative effect sizes (17b), and where and how AI intervention and/or its code can be accessed (25-i). None of the articles reported all of the items in the CONSORT-AI checklist.

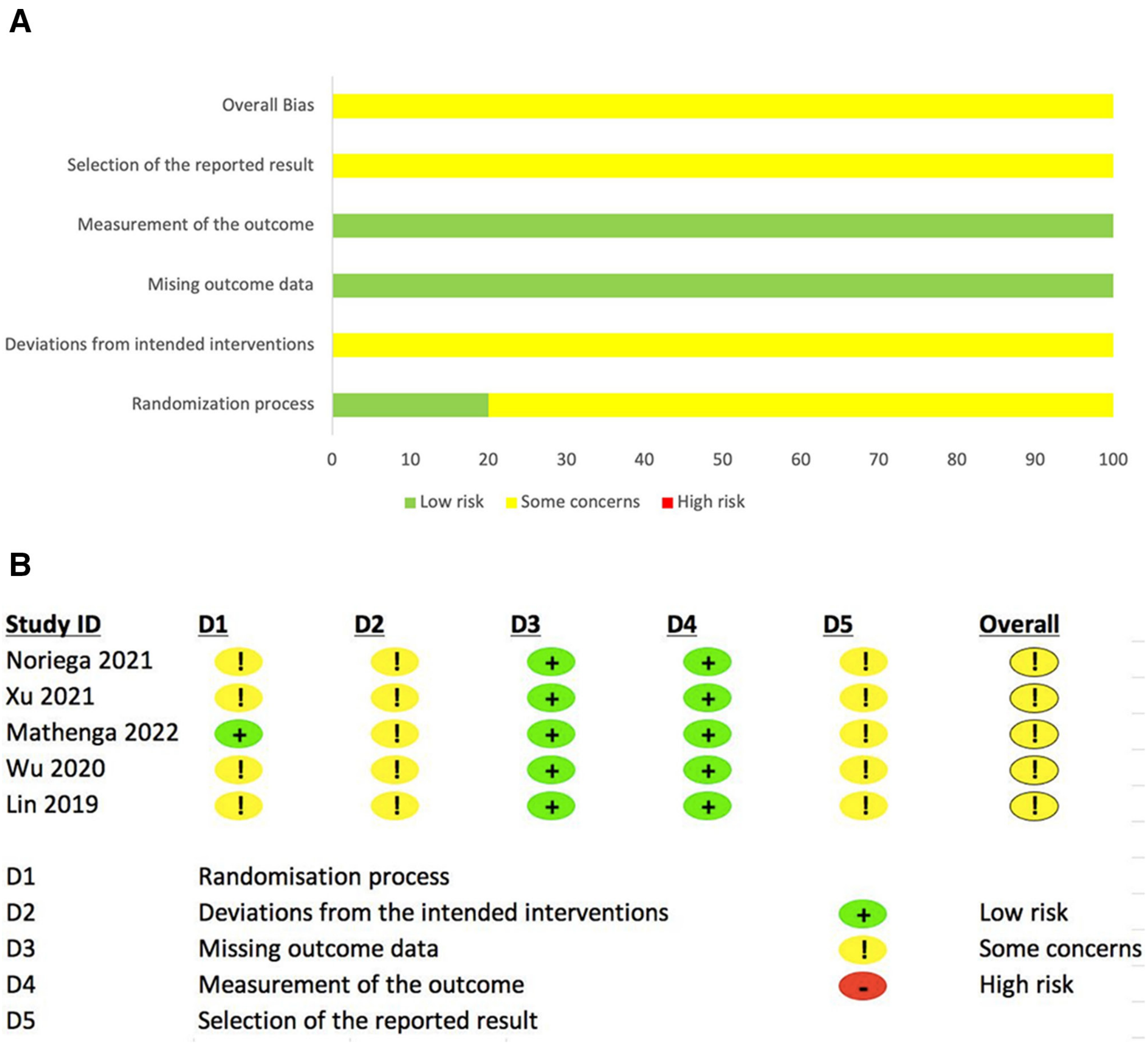
Quality of evidence

The results of the RoB-2 scoring are shown in figure 2. All included articles had an overall moderate risk of bias, with all articles having a score of ‘some concerns present’. All articles scored moderate risk for the domains of ‘selection of the reported result’ and ‘deviations from intended interventions’. All articles were scored as low risk for the domains of ‘measurement of the outcome’ and ‘missing outcome data’. For the domain of ‘randomisation process’, 80% of the articles were moderate risk and the remaining 20% were low risk. None of the articles scored high risk in any domains. The inter-rater concordance for RoB-2 scoring had a kappa score of 0.86.

DISCUSSION

Here, we aimed to evaluate the adherence of RCTs investigating the use of AI within ophthalmology to the guidelines set by CONSORT-AI checklist for reporting standards for RCTs. Our study found a total of five RCTs that evaluated AI applications in ophthalmology. These articles looked at the utility of AI in diabetic retinopathy screening,[19 20] ophthalmologic education,[21] detecting fungal keratitis[22] and diagnosing childhood cataracts.[23] The mean CONSORT-AI score of the articles was 53% (range 37%–78%). None of the articles reported all items in the CONSORT-AI checklist, and all articles were rated as moderate risk, or ‘some concerns present’, through the RoB-2 tool assessment. All articles had moderate risk of bias for the ‘selection of the reported result’ and ‘deviations from intended interventions’ domains, and low risk of bias for ‘measurement of the outcome’ and ‘missing outcome data’ domains. Only one article had low risk of bias for their ‘randomisation process’, with the remainder having moderate risk in this domain.

The mean CONSORT score for our included studies (53%) is higher than mean score of 39% reported in the previous work by Yao *et al* in 2014 which reviewed the quality of reporting guidelines in 64 RCTs focused on ophthalmic surgery.[12] Aside from the difference in the number of reviewed articles, a potential reason for this difference in reported CONSORT-AI scores is that the articles found in our study are relatively new. The CONSORT-AI guidelines were published in 2020, and 3/5 of our articles were published in 2021 or later,[19 20 22]which suggests that awareness of and adherence to reporting guidelines may have increased over time. Many of the items that the identified articles in our review failed to report on were also missed in studies identified by Yao *et al.*[12] These include determining adequate sample size (item 7), concern random allocation sequence generation (item 8) and its implementation (item 10).[13] The low reporting rate of sample size calculation is a critical concern as this information is essential for protocol



Study ID	D1	D2	D3	D4	D5	Overall
Noriega 2021	!	!	+	+	!	!
Xu 2021	!	!	+	+	!	!
Mathenga 2022	+	!	+	+	!	!
Wu 2020	!	!	+	+	!	!
Lin 2019	!	!	+	+	!	!

D1 Randomisation process
D2 Deviations from the intended interventions
D3 Missing outcome data
D4 Measurement of the outcome
D5 Selection of the reported result

+ Low risk
! Some concerns
- High risk

Figure 2 Risk of bias assessment using RoB2 tool for included studies displayed by means of a weighted plot for the distribution of the overall risk of bias within each bias domain (A) and traffic light plot of the risk of bias of each included clinical study (B).

development in all RCTs. There were some items that were commonly missed in Yao *et al* that were not missed in our reviewed articles, such as mentioning the term RCT in title or abstract (item 1),[12] which demonstrates the value in establishment of expected reporting standards by journals and publishing editors.

We observed some common trends in CONSORT-AI and RoB2 assessments in our study. For AI-based RCTs, it is difficult to blind both the physicians and participants to the intervention received, if the participants are humans and not images. For instance, if an RCT is comparing AI-based screening versus human-based screening, the participant may know whether they have been assigned to the AI or to a human at the time the intervention is given. One strategy to blind the participants, as seen in Noriega *et al*[19] and Xu *et al*,[22] is to replace human participants with human-derived data. Additionally, blinding the outcome assessors to the prescribed intervention is an important feature of the study design in RCTs, but in three of the included studies in this review, Noriega *et al*,[19] Xu *et al*[22] and Wu *et al*,[21] did not outline these steps in their methods.

None of our included articles described where to find their initial trial protocol. Only one of the articles, by Lin *et al*, was registered on ClinicalTrials.gov.[23] This is a critical limitation as it could indicate a potential source of bias if analysis decisions were made after outcomes were measured which undermines the credibility of the RCT findings. Although outcome measurements were standard choices (eg, sensitivity and specificity for binary classification model performance), the role of an initial trial protocol cannot be overlooked as it is a key component of pretrial planning and study integrity. Furthermore, no articles other than Mathenge *et al* reported where the AI algorithm codes could be found.[20] This reduces transparency and may impede the reproducibility of the results as well as the progress of applying AI technologies. Siontis *et al* have found that AI RCTs across all healthcare applications, not just ophthalmology, fail to provide the algorithm code for their AI tools.[24]

Criteria 4b (settings and locations where data were collected), 15 (baseline demographics) and 21 (generalisability of trial findings) of the CONSORT-AI checklist were not perfectly adhered to in our five articles. Only three articles reported items 4b,[20 22 23] three articles reported item 15[20 21 23] and two articles reported item 21.[20 22] Although these criteria were not the most frequently missed items, they are of utmost importance clinically, as they concern whether the results of the trial can be reasonably applied to a clinician's patient population. In a 2021 review of the development and validation pathways of AI RCTs, Siontis *et al* found that most AIs are not tested on datasets collected from patient populations outside of where the AI was developed and thus, it may be unsafe to apply these AIs to such populations.[24] In fact, using limited or imbalanced datasets both in development and validation stages may lead to discriminatory AI.[25] Therefore, special attention should be paid to these criteria.

In our review, we also found that the criterion for providing an explanation of any interim analyses and/or stopping guidelines if applicable (7b) was not reported across all articles. It could be argued that all RCTs should at least comment that an interim analysis was not planned, even if it was not applicable to the specific study design. Shahzad *et al* conducted a systematic review that also used CONSORT-AI to review the reporting quality of AI RCTs across all healthcare applications published between January 2015 and December 2021. They also found that item 7b was not reported in more than 85% of the included studies, and scored these items as non-applicable in their grading using CONSORT-AI.[16]

When analysing the appropriateness of analyses and the clarity of the performance assessments for each article, we found that each article chose suitable methods for their individual trials. Noriega *et al*, Xu *et al* and Lin *et al* evaluated the performance of their different comparators by calculating sensitivity and specificity among other metrics.[19 22 23] Xu *et al* and Lin *et al* presented this information in the form of a table.[22 23] Noriega *et al* and Xu *et al* also presented these results visually by plotting sensitivity and specificity of different comparators on a receiver operating curve which represented the performance of the AI alone.[19 22] In Wu *et al*'s investigation of the effectiveness of AI-assisted problem based learning, ophthalmology clerks did a pre-lecture test and post-lecture test after either a traditional lecture or AI-assisted lecture.[21] Improvement in test performance was assessed and compared between the two groups by analysing differences in the pre-lecture and post-lecture test scores using paired t-tests. A main source of bias in their study, not captured in the risk of bias assessment, is the quality of the test questions which were not made available to the readers. It is important to note that all AI-based RCTs identified in this study had no drop-outs, as all participants that enrolled in the RCT yielded valid data for analysis. This is due to the fact that in some cases, images were subjects, in pre-collected databases and registries.

Despite the comprehensive search of the literature, a limited number of RCTs on AI were retrieved in the current study. The small number of RCTs identified prevented our study from conducting any temporal analyses or stratifying our analyses. In comparison, a literature review on the reporting guidelines of RCTs in ophthalmic surgery overall yielded 65 RCTs.[12] There are a couple of reasons that may explain the small number of RCTs investigating the efficacy of AI for ophthalmological applications. First, this small number may be an indication of the novelty of AI within the field of ophthalmology. Another reason may be the high costs and resources associated with RCTs. It is not feasible to conduct an RCT for all of the various AI tools developed for ophthalmology. Siontis *et al* found that the development and validation stages that different AI models go through before being evaluated in RCTs vary widely between papers.[24] The increasing number of standard guidelines for the reporting and quality assessment of AI, including DECIDE-AI,[26] PROBAST-AI,[27] QUADAS-AI,[28] STARD-AI[29] and TRIPOD-AI[27] are suggestive of the shift towards standardised assessment of AI tools. Another step that may aid in better assessment of AI tools in RCTs is determining performance metric thresholds that must be met at each stage of development and validation, although justifying these cutoffs may be difficult and subjective, and does not automatically imply high reliability for the RCT results.

CONCLUSIONS

AI is a growing field within ophthalmology that holds great promise for its applications in wide-reaching areas. Our findings suggest that there are a limited number of RCTs on applications of AI in ophthalmology, and adherence to some aspects of the 2020 CONSORT-AI reporting guidelines is suboptimal. It is essential that future trials provide information on protocol registration, a clear explanation for sample size calculations and details on the method of randomisation (i.e. type of randomisation, how it was implemented, who it was implemented by). Open access to the AI algorithm codes as well as further details about the software and version number used will enhance reproducibility of research efforts. Attention should be paid to blinding participants, physicians and the outcome assessors whenever possible. Finally, it is critical to report information that allows the readers to assess the generalisability of the trial results, such as baseline demographics of patients and settings where the trial data are collected.

It is recommended that future authors, funding organisations, peer-reviewers and others involved in the ophthalmological research process collaborate and place emphasis on adherence and integration of the CONSORT-AI checklist within the RCT development and publication process. This may facilitate the reproducibility of AI research which can in turn be a stimulus for more AI-based RCTs and its clinical application in ophthalmology.

Author affiliations
[1]Queen's University School of Medicine, Faculty of Health Sciences, Kingston, Ontario, Canada
[2]Temerty Faculty of Medicine, University of Toronto, Toronto, Ontario, Canada
[3]Department of Ophthalmology, University College London, London, UK
[4]Department of Ophthalmology and Vision Sciences, University of Toronto, Toronto, Ontario, Canada
[5]University Health Network, Toronto, Ontario, Canada

Twitter Tina Felfeli @TinaFelfeli

Funding This research was in-part funded by Fighting Blindness Canada awarded to Dr. Tina Felfeli.

Competing interests None declared.

Patient consent for publication Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data sharing not applicable as no datasets generated and/or analysed for this study. All data relevant to the study are included in the article or uploaded as online supplemental information.

ORCID iDs
Niveditha Pattathil http://orcid.org/0000-0002-7583-2640
Tina Felfeli http://orcid.org/0000-0002-0927-3086

REFERENCES

1 Ting DSW, Pasquale LR, Peng L, *et al*. Artificial intelligence and deep learning in ophthalmology. *Br J Ophthalmol* 2019;103:167–75.
2 Gargeya R, Leng T. Automated identification of diabetic retinopathy using deep learning. *Ophthalmology* 2017;124:962–9.
3 Brown JM, Campbell JP, Beers A, *et al*. Automated diagnosis of plus disease in retinopathy of prematurity using deep convolutional neural networks. *JAMA Ophthalmol* 2018;136:803–10.
4 Ting DSW, Cheung CY-L, Lim G, *et al*. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *JAMA* 2017;318:2211–23.
5 Lee CS, Tyring AJ, Deruyter NP, *et al*. Deep-learning based, automated segmentation of macular edema in optical coherence tomography. *Biomed Opt Express* 2017;8:3440–8.
6 Keel S, Li Z, Scheetz J, *et al*. Development and validation of a deep-learning algorithm for the detection of neovascular age-related macular degeneration from colour fundus photographs. *Clin Exp Ophthalmol* 2019;47:1009–18.
7 Centre for Evidence-Based Medicine (CEBM). Oxford centre for evidence-based medicine: levels of evidence (March 2009). University of Oxford. Available: https://www.cebm.ox.ac.uk/resources/levels-of-evidence/oxford-centre-for-evidence-based-medicine-levels-of-evidence-march-2009 [Accessed 17 Sep 2022].
8 Begg C, Cho M, Eastwood S, *et al*. Improving the quality of reporting of randomized controlled trials: the CONSORT statement. *JAMA* 1996;276:637–9.
9 Moher D, Schulz KF, Altman DG. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomised trials. *Lancet* 2001;357:1191–4.
10 Schulz KF, Altman DG, Moher D, *et al*. Statement: updated guidelines for reporting parallel group randomized trials. *Ann Intern Med* 2010;152:726–32.
11 Sánchez-Thorin JC, Cortés MC, Montenegro M, *et al*. The quality of reporting of randomized clinical trials published in ophthalmology. *Ophthalmology* 2001;108:410–5.
12 Yao AC, Khajuria A, Camm CF, *et al*. The reporting quality of parallel randomised controlled trials in ophthalmic surgery in 2011: a systematic review. *Eye (Lond)* 2014;28:1341–9.
13 Liu X, Rivera SC, Moher D, *et al*. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *BMJ* 2020;370:m3164.
14 Varghese J. Artificial intelligence in medicine: chances and challenges for wide clinical adoption. *Visc Med* 2020;36:443–9.
15 Sterne JAC, Savović J, Page MJ, *et al*. RoB 2: a revised tool for assessing risk of bias in randomised trials. *BMJ* 2019;366:l4898.
16 Shahzad R, Ayub B, Siddiqui MAR. Quality of reporting of randomised controlled trials of artificial intelligence in healthcare: a systematic review. *BMJ Open* 2022;12:e061519.
17 Wang J, Wu S, Guo Q, *et al*. Investigation and evaluation of randomized controlled trials for interventions involving artificial intelligence. *Intell Med* 2021;1:61–9.
18 Kothari R, Chiu C, Moukheiber M, *et al*. A descriptive appraisal of quality of reporting in a cohort of machine learning studies in anesthesiology. *Anaesth Crit Care Pain Med* 2022;41:101126.
19 Noriega A, Meizner D, Camacho D, *et al*. Screening diabetic retinopathy using an automated retinal image analysis system in independent and assistive use cases in Mexico: randomized controlled trial. *JMIR Form Res* 2021;5:e25290.
20 Mathenge W, Whitestone N, Nkurikiye J, *et al*. Impact of artificial intelligence assessment of diabetic retinopathy on referral service uptake in a low-resource setting: the RAIDERS randomized trial. *Ophthalmol Sci* 2022;2:100168.
21 Wu D, Xiang Y, Wu X, *et al*. Artificial intelligence-tutoring problem-based learning in ophthalmology clerkship. *Ann Transl Med* 2020;8:700.
22 Xu F, Jiang L, He W, *et al*. The clinical value of explainable deep learning for diagnosing fungal Keratitis using in vivo confocal microscopy images. *Front Med* 2021;8:797616.
23 Lin H, Li R, Liu Z, *et al*. Diagnostic efficacy and therapeutic decision-making capacity of an artificial intelligence platform for childhood cataracts in eye clinics: a multicentre randomized controlled trial. *EClinicalMedicine* 2019;9:52–9.
24 Siontis GCM, Sweda R, Noseworthy PA, *et al*. Development and validation pathways of artificial intelligence tools evaluated in randomised clinical trials. *BMJ Health Care Inform* 2021;28:e100466.
25 Leslie D, Mazumder A, Peppin A, *et al*. Does "AI" stand for augmenting inequality in the era of COVID-19 healthcare. *BMJ* 2021;372:n304.
26 DECIDE-AI Steering Group. DECIDE-AI: new reporting guidelines to bridge the development-to-implementation gap in clinical artificial intelligence. *Nat Med* 2021;27:186–7.
27 Collins GS, Dhiman P, Andaur Navarro CL, *et al*. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. *BMJ Open* 2021;11:e048008.
28 Sounderajah V, Ashrafian H, Rose S, *et al*. A quality assessment tool for artificial intelligence-centered diagnostic test accuracy studies: QUADAS-AI. *Nat Med* 2021;27:1663–5.
29 Sounderajah V, Ashrafian H, Golub RM, *et al*. Developing a reporting guideline for artificial intelligence-centred diagnostic test accuracy studies: the STARD-AI protocol. *BMJ Open* 2021;11:e047709.
30 Page MJ, McKenzie JE, Bossuyt PM, *et al*. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* 2021;372:n71.

BMJ Health & Care Informatics

Surgical pit crew: initiative to optimise measurement and accountability for operating room turnover time

Nicole H Goldhaber ,[1] Robin L Schaefer,[2] Roman Martinez,[2] Andrew Graham,[3] Elizabeth Malachowski,[2] Lisa P Rhodes,[2] Ruth S Waterman,[4] Kristin L Mekeel,[1] Brian J Clay,[5] Michael McHale[6]

To cite: Goldhaber NH, Schaefer RL, Martinez R, *et al*. Surgical pit crew: initiative to optimise measurement and accountability for operating room turnover time. *BMJ Health Care Inform* 2023;**30**:e100741. doi:10.1136/bmjhci-2023-100741

Received 29 January 2023
Accepted 12 June 2023

[1]Department of Surgery, University of California San Diego, La Jolla, California, USA
[2]Perioperative & Procedural Services, University of California San Diego, La Jolla, California, USA
[3]Information Services, University of California San Diego, La Jolla, California, USA
[4]Department of Anesthesia, University of California San Diego, La Jolla, California, USA
[5]Biomedical Informatics, UC San Diego, La Jolla, California, USA
[6]Department of OBGYN, University of California San Diego, La Jolla, California, USA

Correspondence to
Dr Nicole H Goldhaber; nhgoldhaber@health.ucsd.edu

ABSTRACT

Background and objectives Turnover time (TOT), defined as the time between surgical cases in the same operating room (OR), is often perceived to be lengthy without clear cause. With the aim of optimising and standardising OR turnover processes and decreasing TOT, we developed an innovative and staff-interactive TOT measurement method.

Methods We divided TOT into task-based segments and created buttons on the electronic health record (EHR) default prelogin screen for appropriate staff workflows to collect more granular data. We created submeasures, including 'clean-up start', 'clean-up complete', 'set-up start' and 'room ready for patient', to calculate environmental services (EVS) response time, EVS cleaning time, room set-up response time, room set-up time and time to room accordingly.

Results Since developing and implementing these workflows, measures have demonstrated excellent staff adoption. Median times of EVS response and cleaning have decreased significantly at our main hospital ORs and ambulatory surgery centre.

Conclusion OR delays are costly to hospital systems. TOT, in particular, has been recognised as a potential dissatisfier and cause of delay in the perioperative environment. Viewing TOT as one finite entity and not a series of necessary tasks by a variety of team members limits the possibility of critical assessment and improvement. By dividing the measurement of TOT into respective segments necessary to transition the room at the completion of one case to the onset of another, valuable insight was gained into the causes associated with turnover delays, which increased awareness and improved accountability of staff members to complete assigned tasks efficiently.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Delay in turnover time (TOT) can cause meaningful disruption to perioperative operations, and TOT serves many as a key quality process measure.

WHAT THIS STUDY ADDS

⇒ By dividing TOT into process-based components, we created a more reliable model for staff accountability and process efficiency.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ This study describes an innovative method for measuring TOT. Using this system, we hope to reduce operating room TOT.

INTRODUCTION

Turnover time, or the time between surgical cases in the same operating room (OR), can often give the impression of being too long without clear reason. Delay in turnover time can cause meaningful disruption to perioperative operations, and as such, turnover time serves as a key quality process measure at countless institutions.[1–4] In addition, turnover delays impact case-start times, which can be a source of patient distrust and dissatisfaction, as additional wait time can create additional stress to patients who are already anxious.[5] Ideally, it is an institutional goal to minimise the time between surgical cases in order to support surgical demand and growth, as well as to improve revenue and profits.

In attempts to shorten turnover time, others have tried implementing mobile applications,[6] designated specialised OR teams,[7] turnover task cards,[8] remote video auditing with real-time feedback,[9] and other published process improvement initiatives.[10–14] However, despite these and other attempts, turnover time continues to be a frustratingly difficult problem to solve at many if not most institutions. To our knowledge, this is the first study at a multisite quaternary academic medical centre to describe a novel method for measuring OR turnover.

The processes that occur during turnover can be complex, variable and at times seemingly nebulous and/or chaotic; however, they all come down to what needs to get accomplished in order to complete and clean up after the previous case (primary process owners: environmental services (EVS)), and then to setup and prepare for the case to follow in the same OR (primary process owners: surgery technicians and nurses). At many institutions, including the study institution, turnover is measured as one block of time—with therefore little specific insight into which process or processes are responsible for turnover delays. As a result, it becomes difficult to hold teams accountable without measurable, granular data to elucidate the specific causes of increased time.

With the aim of optimising and standardising OR turnover processes and potentially improving perioperative efficiency by reducing turnover time, we developed an innovative and staff-interactive method for measuring the time in between surgical cases. By dividing overall turnover time into parts based on processes or workflows, we hypothesised that we could achieve more accurate measurements of turnover time components, hold-specific teams accountable for their processes within each component, and ultimately decrease turnover time delays.

METHODS

This study was conducted at a suburban regional academic health system, which is composed of two acute care hospitals housing 51 ORs. The EHR used at this institution is Epic (Verona, Wisconsin, USA). A process improvement workshop was held to address the topic of OR turnover time, which was attended by end-users from multiple perioperative teams involved in turnover processes.

For our intervention, we divided OR turnover into task-based segments and created clickable buttons on the EHR default prelogin screen for appropriate staff workflows to collect more granular data. We defined 'turnover time' as the time between 'wheels out' of the previous case to 'wheels in' of the case to follow (figure 1)–case tracking events that are usually recorded in the EHR by the circulating nurse in the OR. We created new case tracking events in between these that include 'clean-up start', 'clean-up complete', and 'set-up start', to allow the calculation of EVS response time, EVS cleaning time, room setup response time and room setup time accordingly.

Given not all staff members (including EVS) have access to log into the EHR at our institution, we made the new buttons readily available on the room-specific ;prelogin status board' located on the home screen of all computers in the ORs to be clicked by appropriate staff members at the time of initiation and completion of clean-up and setup processes. Education was provided on the location of the buttons and other fields (for those with EHR access) as well as best practices for standard work for utilisation of the buttons within existing staff workflows.

Timestamp measurements are stored within the EHR for each case, and are extracted on a weekly and/or monthly basis to be displayed on team-based dashboards created in Microsoft Excel, using data from the EHR timestamp measurements, and sent to team managers via email for accountability and to keep track of progress over time. Overall and location-specific data are displayed on the dashboards in the same presentation as the results figures below. Descriptive statistics, including median time and utilisation percentages, were performed using these data to assess the impact of our intervention. This manuscript was structured based on the Standard QUality Improvement Reporting Excellence (SQUIRE) guidelines for quality improvement (QI) studies.

RESULTS

Since developing and implementing these workflow interventions (table 1), the study institution has seen excellent adoption by staff members (figure 2A–C). EVS utilisation of the 'clean-up start' speed button is stable at 90% of cases with subsequent cases to follow at all main OR locations. Utilisation of the 'set-up start' speed button is stable >80% of cases with subsequent cases to follow at all main OR locations. Utilisation of the 'room ready for patient' speed button is stable >85% of cases. Of note, 'room ready for patient' is a speed button that was not newly created for this project, as it existed previously to help designate to the preoperative and anaesthesia teams that the room is ready for the patient to be brought back.

We have also seen the median duration of EVS response times ('wheels out' → 'clean-up start') and cleaning times ('clean-up start' → 'clean-up complete') decrease substantially over time at our main hospital ORs (figure 3A,B). Median response time remains consistently under 4 min (previously greater than 5 min) and cleaning time remains consistently below 10 minutes

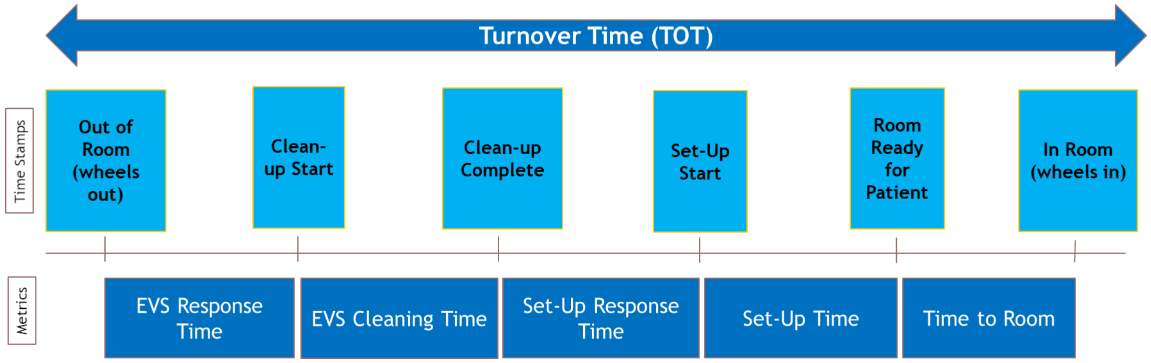


Figure 1 Operating room turnover time measurement timeline. EVS, environmental services.

Table 1 Workshop and intervention timeline

Project milestone	Date
Perioperative efficiency project (PEP) workshop	9 November 2020–16 November 2020
Turnover time (TOT) case tracking event go-live	15 December 2020
PEP dashboard go-live	7 June 2021
TOT workshop	25 October 2021

(previously greater than 15 min). These trends appear to correlate with greater adoption of the EHR functionality intervention.

Median setup response time ('clean-up complete' → 'set-up start'), set-up time ('set-up start' → 'setup complete'), and time to room ('set-up complete' → 'wheels in') have not significantly decreased and display variability across locations as well as over time (figure 4A–C). Overall, turnover time has not significantly decreased over time and remains consistently above target at all locations (figure 5).

DISCUSSION

Delays in the OR can be costly to hospital systems, and one area where known delays occur is during room turnover. One study describes each minute of time running an OR in California hospitals costing approximately US$37 for inpatient settings and US$36 for ambulatory settings.[15] Another study describes a methodology for surgical centres to calculate potential reduction in staffing costs as a result of decreases in OR turnover times.[16] OR turnover can be a seemingly nebulous time between surgical cases. There are certain tasks that need to be accomplished once the previous patient's case was completed to prepare for the case to follow in the same OR. Much like a motor racing pit crew, many different personnel are involved in many distinct, yet potentially overlapping processes during this OR turnover time. While traditionally the measurement of these tasks has been lumped together into one turnover time metric, this study demonstrates an alternative method to help guide efficiency and accountability for the individual tasks that occur during OR turnover.

There can be some variability in tasks during turnover based on the type of operation to be performed, individual staff preferences, patient factors and more.[17 18] Among this variability arises one constant universal theme when discussing OR turnover time: 'Why did it take so long?' Turnover time, in particular, has been recognised as a potential dissatisfier and cause of delay in the perioperative environment.[13] It also has the potential to erode the goals of efficiency and safety within the perioperative environment. A previous study has shown that perceptions of turnover time may be skewed by staff member role and factors perceived as contributing to the time, and suggest for OR managers to reference timestamp data on turnover time length rather than relying on surgeon or anaesthesiologist 'expert judgement'.[19] While there is access to EHR data of the overall length of OR turnover time, the data were not sufficient to answer this universal question, a question that plagues hospitals and surgical centres across the globe.

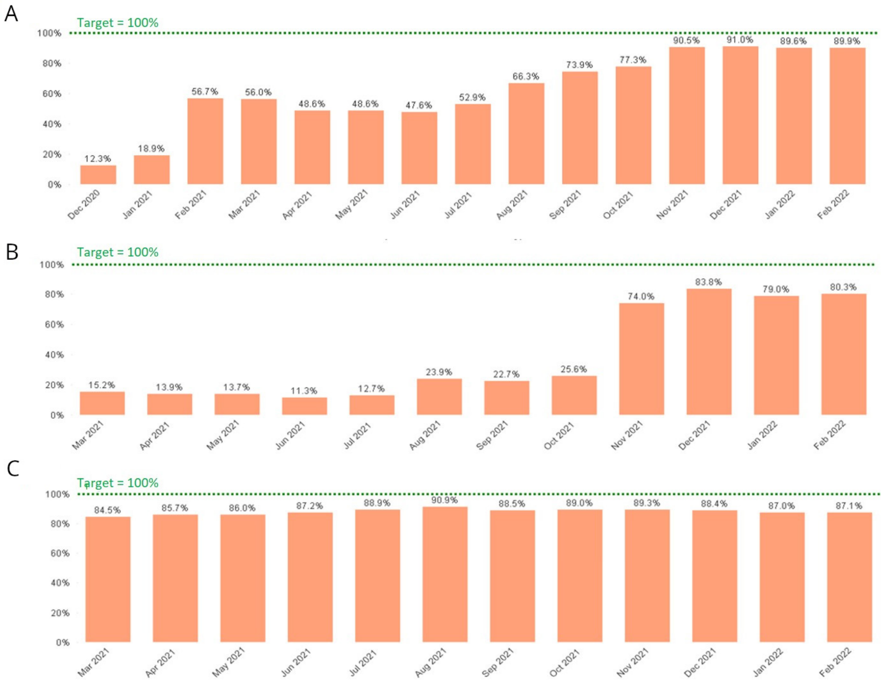


Figure 2 (A) EVS utilisation (percentage of cases) of clean-up start button in the OR over time, (B) utilisation (percentage of cases) of set-up start button in the OR over time, (C) utilisation (percentage of cases) of room ready for patient button in the OR over time. EVS, environmental services' OR, operating room.

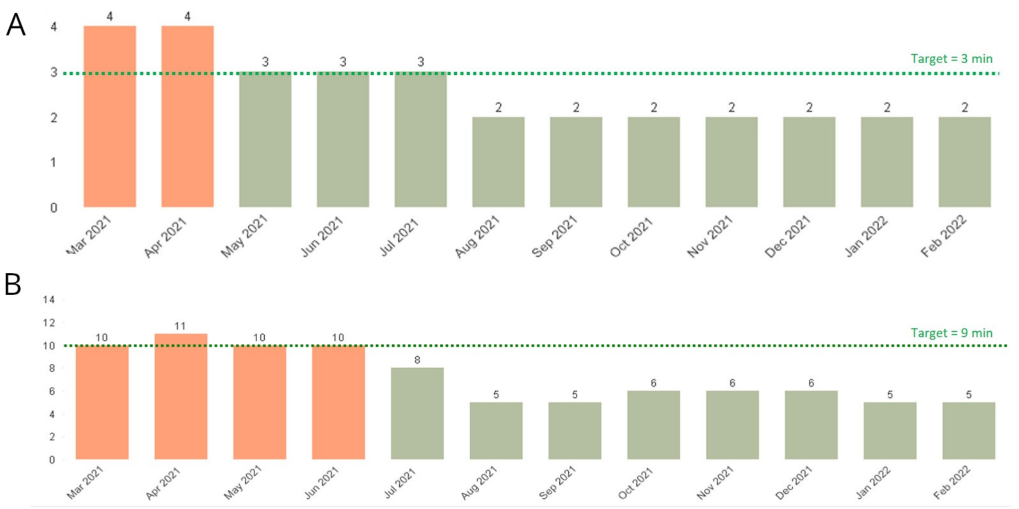


Figure 3 Median (minutes) EVS response time (A) and cleaning time (B) over time. EVS, environmental services.

Viewing turnover time as one finite value or entity and not a series of necessary tasks by a variety of team members limits the possibility of critical assessment and improvement. In this study, in order to provide a more reliable and detailed measurement system to answer this question, OR turnover time was divided into three main phases or components based on the staff and processes that occur accordingly—the respective segments necessary to transition the room at the completion of one case to the onset of another. While that sounds simple, like the flip of a switch, to use to shorten turnover time, in this study we demonstrate this is not so simple and that turnover is comprised of several interconnected and oftentimes interdependent components of a larger whole.

This project demonstrated the successful implementation of a new staff-interactive timestamping system, as demonstrated by the high utilisation rates of the buttons created in the EHR. Both staff engagement and the sharing of performance metrics have been shown to be key to enhancing OR efficiency.[20] With this new process in place at the study institution, staff are now leaving more accurate, room-specific and case-specific, timestamps in the system in order to more precisely pinpoint when components of turnover are being initiated and completed. The time being spent on each of these components can be accurately quantified, assessed and addressed accordingly.

A notable decrease in time for the EVS workflow was observed in this study. It was confirmed with team management that there were no additional significant process changes (eg, new faster drying cleaning solution) that took place during this time that would account for

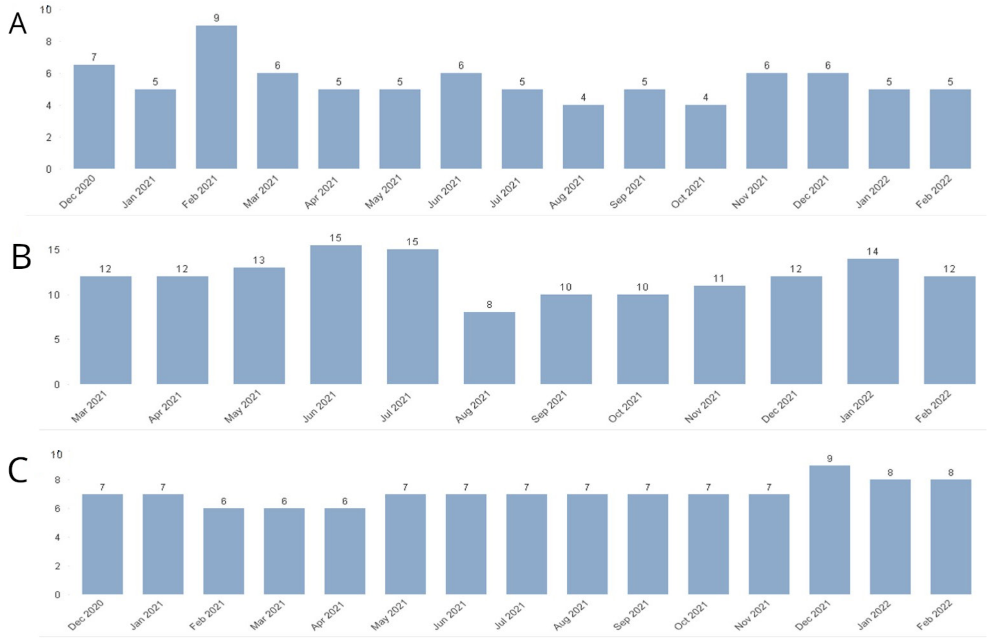


Figure 4 Median (minutes) set-up response time (A), set-up time (B) and time to room (C).

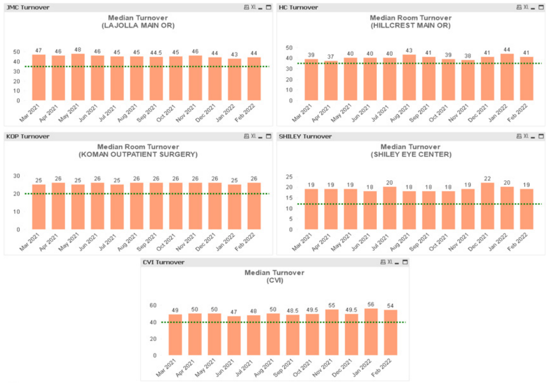


Figure 5 Median turnover time (minutes) by location over time.

the time decrease otherwise. In conjunction with high utilisation rates by EVS, this tells us that our intervention has provided an accurate measurement tool for our EVS teams' efficiency that is successfully holding times below target, an achievement we continue to sustain and celebrate.

While we have not yet seen similar notable decreases in set-up and overall turnover time, significant progress toward this goal has been made now that a more accurate and detailed measurement system is in place. We acknowledge and are limited by the fact that there will always be significant variability in set-up time given different services or different cases require a greater amount and/or greater complexity of equipment that requires set-up time accordingly.

Another limitation identified based on these results is that technology changes alone are not sufficient for process improvement. Focused staff-based and process-based education is required for successful implementation of a new system, as well as for sustainable change to occur. An added benefit of this project is that a transparent system of accountability for staff teams was implemented, providing more awareness of each team or staff member's contribution.

Future directions include further service line-specific analysis of case set-up time to identify potential areas of improvement while accounting for service line-specific variability (eg, amount of case equipment). Additionally, standard work for the teams responsible for bringing the patient to the room once the room is set up or ready will be developed and implemented. This will include a standardised messaging system to the anaesthesia providers assigned to the case so they do not have to wait and multitask while attempting to predict when the room might be ready for the patient to come back. Lastly, this system is reliant on individual human entry of case tracking event data, which could be improved by automation such as through wireless sensors and radio frequency identification device technology.[21 22]

Efficiency-based QI initiatives have been cited to potentially include feelings of pressure to produce a fast result, which may in turn compromise a high-quality result. Other investigations have found no negative impact on patient safety and quality of care resulting from gradual implementation of a methodologically structured efficiency-based QI initiative in a perioperative environment.[23 24] While it is not possible to attribute specific patient safety and outcomes data directly to this particular QI initiative, to our knowledge, there was no significant change in the number of complications since the implementation of this system.

CONCLUSION

By dividing our measurement of turnover time into the processes that need to be completed: 'clean-up' and 'set-up', we were able to gain valuable insight into the components of turnover and turnover delays and increase awareness and accountability of staff members to complete assigned tasks efficiently without compromising quality and patient safety.

Contributors NHG helped plan, conduct and submit the study, and is the guarantor responsible for the overall content. RLS helped plan and conduct the study. RM helped construct the implementation tool. AG helped construct the implementation tool. EM helped plan and conduct the study. LPR helped plan the study. RSW helped plan the study. KLM helped plan the study. BC helped plan and submit the study. MM helped plan and submit the study.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient consent for publication Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available on reasonable request.

ORCID iD

Nicole H Goldhaber http://orcid.org/0000-0002-3847-3634

REFERENCES

1 Wallace L, Muir M, Romano L, *et al*. Assessing operating theatre efficiency: a prospective cohort study to indentify intervention targets to improve efficiency. *ANZ J Surg* 2021;91:2382–8.
2 Reeves JJ, Waterman RS, Spurr KR, *et al*. Efficiency Metrics at an academic Freestanding ambulatory surgery center: analysis of the impact on scheduled end-times. *Anesth Analg* 2021;133:1406–14.
3 Hoffman CR, Green MS, Liu J, *et al*. Using operating room turnover time by anesthesia Trainee level to assess improving systems-based practice milestones. *BMC Med Educ* 2018;18:1–5.
4 Macario A. Are your hospital operating rooms "efficient"? A scoring system with eight performance indicators. *Anesthesiology* 2006;105:237–40.
5 Freeman K, Denham SA. Improving patient satisfaction by addressing same day surgery wait times. *Journal of PeriAnesthesia Nursing* 2008;23:387–93.
6 Uddin M, Allen R, Huynh N, *et al*. Assessing operating room turnover time via the use of mobile application. *Mhealth* 2018;4:12.
7 Avery DM III, Matullo KS. The efficiency of a dedicated staff on operating room turnover time in hand surgery. *The Journal of Hand Surgery* 2014;39:108–10.
8 Souders CP, Catchpole KR, Wood LN, *et al*. Reducing operating room turnover time for Robotic surgery using a motor racing pit stop model. *World J Surg* 2017;41:1943–9.

9 Overdyk FJ, Dowling O, Newman S, *et al*. Remote Video-auditing with real-time feedback in an academic surgical suite improves safety and efficiency Metrics: a clustered randomised study. *BMJ Qual Saf* 2016;25:947–53.

10 Cerfolio RJ, Ferrari-Light D, Ren-Fielding C, *et al*. Improving operating room turnover time in a New York City academic hospital via lean. *Ann Thorac Surg* 2019;107:1011–6.

11 Collar RM, Shuman AG, Feiner S, *et al*. Lean management in academic surgery. *J Am Coll Surg* 2012;214:928–36.

12 Tagge EP, Thirumoorthi AS, Lenart J, *et al*. Improving operating room efficiency in academic children's hospital using lean six Sigma methodology. *Journal of Pediatric Surgery* 2017;52:1040–4.

13 Soliman BAB, Stanton R, Sowter S, *et al*. Improving operating theatre efficiency: an investigation to significantly reduce changeover time. *ANZ J Surg* 2013;83:545–8.

14 Sarpong K, Kamande S, Murray J, *et al*. Consecutive surgeon and anesthesia team improve turnover time in the operating room. *J Med Syst* 2022;46:16.

15 Childers CP, Maggard-Gibbons M. Understanding costs of care in the operating room. *JAMA Surg* 2018;153:e176233.

16 Dexter F, Abouleish AE, Epstein RH, *et al*. Use of operating room information system data to predict the impact of reducing turnover times on staffing costs. *Anesthesia & Analgesia* 2003;97:1119–26.

17 Gottschalk MB, Hinds RM, Muppavarapu RC, *et al*. Factors affecting hand surgeon operating room turnover time. *Hand (New York, N,Y)* 2016;11:489–94.

18 Dexter F, Epstein RH, Marcon E, *et al*. Estimating the incidence of prolonged turnover times and delays by time of day. *Anesthesiology* 2005;102:1242–8.

19 Masursky D, Dexter F, Isaacson SA, *et al*. Nancy A Nussmeier, Surgeons' and Anesthesiologists' perceptions of turnover times. *Anesth Analg* 2011;112:440–4.

20 Cima RR, Brown MJ, Hebl JR, *et al*. Use of lean and six Sigman methodology to improve operating room efficiency in a high-volume tertiary-care academic medical center. *J Am Coll Surg* 2011;213:83–92;

21 Huang AY, Joerger G, Salmon R, *et al*. A robust and non-obtrusive automatic event tracking system for operating room management to improve patient care. *Surg Endosc* 2016;30:3638–45.

22 Marchand-Maillet F, Debes C, Garnier F, *et al*. Accuracy of patient's turnover time prediction using RFID technology in an academic ambulatory surgery center. *J Med Syst* 2015;39:12.

23 Chernov M, Vick A, Ramachandran S, *et al*. Perioperative efficiency vs quality of care - do we always have to choose *Journal of Investigative Surgery* 2020;33:265–70.

24 Inomata T, Mizuno J, Iwagami M, *et al*. The impact of joint Commission International accreditation on time periods in the operating room: A retrospective observational study. *PLoS ONE* 2018;13:e0204301.

BMJ Health & Care Informatics

Validation of US CDC National Death Index mortality data, focusing on differences in race and ethnicity

Monica Ter-Minassian ,[1] Sundeep S Basra ,[1] Eric S Watson,[1] Alphonse J Derus,[2] Michael A Horberg[1]

To cite: Ter-Minassian M, Basra SS, Watson ES, *et al*. Validation of US CDC National Death Index mortality data, focusing on differences in race and ethnicity. *BMJ Health Care Inform* 2023;**30**:e100737. doi:10.1136/bmjhci-2023-100737

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/bmjhci-2023-100737).

Received 11 January 2023
Accepted 24 June 2023

[1]Mid-Atlantic Permanente Research Institute, Mid-Atlantic Permanente Medical Group, Rockville, Maryland, USA
[2]Research Administration, Kaiser Permanente Washington Health Research Institute, Seattle, Washington, USA

Correspondence to
Dr Monica Ter-Minassian; monica.ter-minassian@kp.org

ABSTRACT

Objectives The US Center for Disease Control and Prevention's National Death Index (NDI) is a gold standard for mortality data, yet matching patients to the database depends on accurate and available key identifiers. Our objective was to evaluate NDI data for future healthcare research studies with mortality outcomes.

Methods We used a Kaiser Permanente Mid-Atlantic States' Virtual Data Warehouse (KPMAS-VDW) sourced from the Social Security Administration and electronic health records on members enrolled between 1 January 2005 to 31 December 2017. We submitted data to NDI on 1 036 449 members. We compared results from the NDI best match algorithm to the KPMAS-VDW for vital status and death date. We compared probabilistic scores by sex and race and ethnicity.

Results NDI returned 372 865 (36%) unique possible matches, 663 061 (64%) records not matched to the NDI database and 522 (<1%) rejected records. The NDI algorithm resulted in 38 862 records, presumed dead, with a lower percentage of women, and Asian/Pacific Islander and Hispanic people than presumed alive. There were 27 306 presumed dead members whose death dates matched exactly between the NDI results and VDW, but 1539 did not have an exact match. There were 10 017 additional deaths from NDI results that were not present in the VDW death data.

Conclusions NDI data can substantially improve the overall capture of deaths. However, further quality control measures were needed to ensure the accuracy of the NDI best match algorithm.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ While prior literature describes quality control methods, and enhancements to the National Death Index (NDI) death registry algorithm for identifying true deaths, our study highlighted how race and ethnicity and sex were associated with the ability to match with this death registry for over one million people.

WHAT THIS STUDY ADDS

⇒ An evaluation of the NDI matching algorithm at a large integrated healthcare system highlighted the algorithm strengths and pitfalls. The NDI can provide significantly more mortality data than other US death registry sources, but the matching algorithm missed some deaths determined with other sources. Women, Hispanic and Asian/Pacific Islander populations were more frequent in poorly matched records compared to those well matched to NDI.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ Large studies with mortality outcomes and incomplete follow-up should use multiple mortality databases and will benefit from a similar quality control of data from death registry sources.

INTRODUCTION

The National Death Index (NDI), managed by the National Center for Health Statistics (Hyattsville, Maryland, USA) of the Center for Disease Control and Prevention (CDC; Atlanta, Georgia, USA), is a database of death certificate data from the state vital statistics offices in the USA and territories. NDI is frequently used to ascertain deaths and cause of death in studies where participants may be lost to follow-up.[1] Kaiser Permanente Mid-Atlantic States (KPMAS) is a large multisite healthcare system, which derives death data from databases sourced from the medical record, select state registries, and quarterly updates from the Social Security Administration (SSA) Death Master File. However, more complete information on vital status for members without a valid social security number (SSN) and cause of death on all deceased members was needed for medical research studies.

NDI provides a detailed user's guide to submission and analysis of returned results.[1] However, quality control and testing matching algorithm validity is the responsibility of the submitter. NDI returns a probabilistic score (PS) based on a matching algorithm that evaluates the likelihood of a match over nine demographic variables and their components.[2] Researchers have reported different methods for identifying and verifying a true match to the NDI dataset given varying available data, by comparing it to their large institutional databases.[3–6] Most report that submitting SSN is the best identifier with high sensitivity and specificity.[4] Where SSN is not available or partially matched, researchers have developed algorithms to enhance NDI results or use alternative matching results.

Race and ethnicity may be differentially linked with the availability of SSN, affecting matching with the NDI database, potentially biasing mortality estimates.[7] Hispanic people's records appear to have higher missing SSN and other NDI fields compared with white and black people, possibly leading to inappropriate inferences about survival differences by ethnicity.[7]

Our objective was to produce a comprehensive decedents dataset for KPMAS that could be used for research queries on vital status and cause of death for over 1 million members. We evaluated the NDI PS matching algorithm and developed quality control steps for our racially diverse population.

METHODS

Study population selection

The primary objective of our NDI submission was to find current vital status (by 31 December 2017) and the cause of death of a cohort of KPMAS members that were lost to follow-up, disenrolled or died without record in our electronic health record (EHR) system. We submitted members' data naïve to death status on over one million members enrolled at KPMAS who had a date of last contact (identified by an outpatient encounter or blood pressure measurement) between 1 January 2005 and 31 December 2017. We calculated follow-up time from the year of last contact to 2018, equivalent to the number of years searched per person by NDI.

Patient data were obtained from the Virtual Data Warehouse (VDW) which is a database derived from the KPMAS EHR. We submitted fields required by NDI including first, middle and last names, SSN, birth month, day and year, sex, and state of residence and marital status. We did not submit race, state of birth or father's surname (for women).

To account for significant missing self-reported race and ethnicity data, we used the Bayesian Improved Surname and Geocoding algorithm[8 9] probabilities available in a Kaiser Permanente data repository. We combined the Asian and Pacific Islander (A/PI) populations into one group and combined the American Indian and Alaskan Native (AI/AN) populations with the multiracial population into an other group due to small sample sizes. We tested the association of missing SSN with race and ethnicity using a χ^2 test and by adjusting for sex in a logistic regression analysis.

Evaluation of returned results and additional quality control

NDI returned record level data and summary statistics on people for our submission that matched records in their database. For the matched records, NDI provided information on which fields matched and an overall PS based on weighted matching. KPMAS members with scores above the cut-off were presumed dead because they matched well with someone with a death certificate in the NDI database. KPMAS members with scores below the cut-off were not considered a good match and so presumed alive. However, there is the possibility that some were actually a true match and should not be presumed alive, especially if the PS was borderline and a rerun with additional variables could increase the score. For people with multiple matching records, the PS provided a way to determine the best match. We evaluated each field for percent matched.

Quality assurance

NDI stratified groups of matching fields into classes. In brief, Class 1 requires matching at least eight digits of the 9-digit SSN, and the fields: first name, middle initial, last name, sex, state of birth, birth month, and birth year. Class 2 requires matching at least 7 digits of the SSN and one or more of the fields may not match. For Classes 3 and 4, SSN is unknown but for Class 3, eight or more items match (first name, middle initial, last name, father's surname (for women), birthday, birth month, birth year, sex, race, marital status or state of birth) and for Class 4, fewer than eight of the items match. For Class 5, the SSN exists but does not match and all are considered false matches. The recommended cut-off values for the PS are 44.5, 37.5 and 32.5 for Classes 2, 3 and 4, respectively. The NDI User guide[1] stated that evaluations of the PS algorithm revealed biases in the classification of NDI match status for women and non-white persons, due to changing surnames for women and 'lower reporting of SSNs and incorrect spelling or recording of ethnic names', particularly for Class 4 matches.[1] We compared

Table 1 Characteristics of the population submitted to NDI (n=1 036 449) and the population with scores above the cut-off (presumed dead) (n=38 862) and scores below the cut-off (presumed alive) (n=334 003)

Characteristics	Submitted, n (%)	Above cut-off, n (%)	Below cut-off, n (%)
Sex			
F	527 330 (50.9)	17 213 (44.3)	160 460 (48.0)
M	509 119 (49.1)	21 649 (55.7)	173 543 (52.0)
Population*			
A/PI	93 859 (9.1)	1699 (4.4)	30 384 (9.1)
Black	315 472 (30.5)	15 521 (39.9)	116 749 (35.0)
Hispanic	132 160 (12.8)	2021 (5.2)	43 199 (12.9)
Other	21 546 (2.1)	736 (1.9)	6419 (1.9)
White	470 440 (45.5)	18 767 (48.3)	136 188 (40.8)
Missing/not reported	2972 (0.3)	118 (0.3)	1145 (0.3)

*Imputed; A/PI=Asian or Pacific Islander; Other included American Indian/Alaskan Native and multiple race.
NDI, National Death Index.

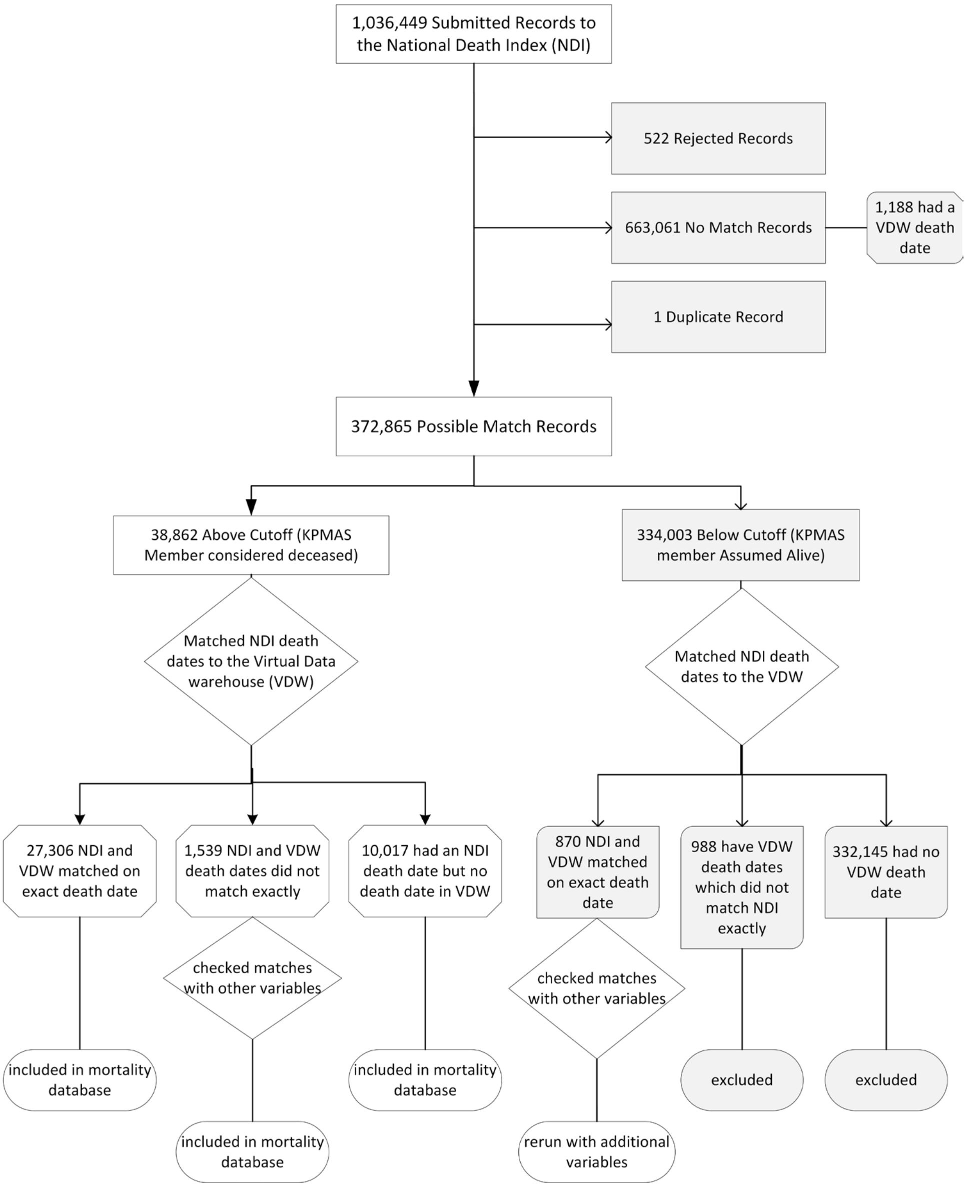

1,036,449 Submitted Records to the National Death Index (NDI)
522 Rejected Records
663,061 No Match Records
1,188 had a VDW death date
1 Duplicate Record
372,865 Possible Match Records
38,862 Above Cutoff (KPMAS Member considered deceased)
334,003 Below Cutoff (KPMAS member Assumed Alive)
Matched NDI death dates to the Virtual Data warehouse (VDW)
Matched NDI death dates to the VDW
27,306 NDI and VDW matched on exact death date
1,539 NDI and VDW death dates did not match exactly
10,017 had an NDI death date but no death date in VDW
870 NDI and VDW matched on exact death date
988 have VDW death dates which did not match NDI exactly
332,145 had no VDW death date
checked matches with other variables
checked matches with other variables
included in mortality database
included in mortality database
excluded
excluded
included in mortality database
rerun with additional variables

Figure 1 KPMAS submission matches the National Death Index and the process of inclusion. KPMAS, Kaiser Permanente Mid-Atlantic States.

percent differences in all above and below cut-off PSs for sex and for imputed race and ethnicity using a χ^2 test. For Class 4 within above and within below cut-off group, we used linear regression analyses to test the associations of PS with imputed race and ethnicity and sex, age at submission to NDI, state of residence and matched birthday, and matched birth year (everyone in Class 4 matched on birth month, so this variable was not included). For comparison, we did the same analyses with self-reported race instead of imputed race.

To determine additional deaths identified by NDI and potential mismatches, we stratified by above and below PS cut-off records and compared NDI results to the KPMAS VDW on vital status, and differences or no difference in death dates. We calculated validation metrics of sensitivity and specificity for the VDW compared with NDI as the gold standard.[10] We used SAS V.9.4 for analyses, where statistical tests were two-sided with a significance level of 0.05.

RESULTS

We submitted 1 036 449 people to NDI. Follow-up time from the date of last contact was 1–13 years with a median of 6 years and average of 6.3 years. However, since we had 69% missing self-reported race and ethnicity (online supplemental table 1), we imputed race and ethnicity resulting in populations that were 9% A/PI, 31% black, 13% Hispanic, 2% AI/AN/multiracial, 45% white and 0.3% not reported due to a missing address (table 1).

Of the 1 033 477 people, with imputed race information, we found 143 452 (13.8%) were missing SSN. Percent missing SSN significantly differed among race and ethnicity (χ^2, p<0.001), and a race–sex-adjusted logistic

Table 2 Exact match on submitted variables for records with scores above the cut-off (presumed dead)

Variable	Class 2 (n=35 441)	Class 3 (n=9)	Class 4 (n=3412)
First name	34 088 (96.2%)	9 (100%)	3337 (97.8%)
Only the first initials and names match NYSIIS	416 (1.2%)	0	31 (0.9%)
Middle initial	21 193 (59.8%)	9 (100%)	1893 (55.5%)
Last name	34 550 (97.5%)	9 (100%)	3385 (99.2%)
Names match only NYSIIS phonetic codes	188 (0.5%)	0	27 (0.8%)
Birth month	35 249 (99.5%)	9 (100%)	3412 (100%)
Birth year	34 859 (98.4%)	9 (100%)	3361 (98.5%)
Within 1 year before or after	316 (0.9%)	0	24 (0.7%)
Sex	35 344 (99.7%)	9 (100%)	3400 (99.6%)
Marital status	73 (0.2%)	9 (100%)	0
State of residence	30 605 (86.4%)	8 (88.9%)	3015 (88.4%)
SSN digit count			
0 or no SSN	0	7 (77.8%)	3403 (99.7%)
7	190 (0.5%)	0	0
8	711 (2.0%)	0	0
9	34 540 (97.5%)	0	0
All data items exactly with the related items on the NDI Records	23 396 (66.0%)	5 (55.6%)	2290 (67.1%)

NDI, National Death Index; NYSIIS, New York State Identification and Intelligence System; SSN, social security number.

regression analysis showed all race and ethnicity categories were significant compared with the white population as a reference group, with Other having the largest OR (OR=14.3 (11.9, 17.1)) followed by the Hispanic population (OR=2.34 (2.30, 2.39)) and by the A/PI population (OR=1.56 (1.53, 1.60)). An analysis of self-reported data showed similar results, except that the Other group had a smaller OR (online supplemental tables 2 and 3).

We obtained summary statistics on 372 865 people that matched records in the NDI (figure 1). We found 663 061 people did not match the NDI database and 522 people were rejected for inclusion of special characters in the fields and one duplicate record was removed. Therefore, we classified 663 583 as alive according to NDI. However, there were 1188 deaths among these patients according to the VDW sources by 31 December 2017.

Above PS cut-off (presumed dead)

Of the 372 865 people that matched records in NDI, we found 38 862 unique people had PSs above the class cut-off threshold (tables 2 and 3, figure 1).

There were no records in Class 1 because state of birth was not submitted. Class 3 had very few matches because father's surname and race were not submitted and very few people submitted had marital status information. Table 2 shows that there was a high percentage of exact matches between the data submitted and the NDI database for each field within each class. Only middle initial

Table 3 NDI matches (n=372 865) stratified by class and probabilistic score, with death dates compared with the VDW

Above cut-off (KPMAS member presumed dead)	Class 2	Class 3	Class 4	Class 5	Sum
NDI and VDW match on exact death date	25 327	7	1972	0	**27 306**
NDI and VDW death dates do not match exactly	1430	0	109	0	**1539**
NDI death date but no death date in VDW	8684	2	1331	0	**10 017**
Sum	35 441	9	3412	0	**38 862**
Below cut-off (KPMAS member presumed alive)	**Class 2**	**Class 3**	**Class 4**	**Class 5**	**Sum**
NDI and VDW match on exact death date	527	0	121	222	**870**
NDI and VDW death dates do not match exactly	23	0	194	771	**988**
NDI (below cut-off) and no death date in the VDW	1022	0	55 942	275 181	**332 145**
Total	1572		56 257	276 174	**334 003**

KPMAS, Kaiser Permanente Mid-Atlantic States; NDI, National Death Index; VDW, Virtual Data Warehouse.

and state of residence had less than 95% matching for all classes. A high percentage of above PS cut-off records (89%) were matched exactly by their 9-digit SSN.

Below PS cut-off (presumed alive)

Of the 372 865 people that matched records in NDI, we found 334 003 had PSs that were below the threshold for class, therefore considered poor matches and the KPMAS member was presumed alive (table 3). Most, (82.7%) were in Class 5 (people who had an SSN that did not match any SSN in NDI).

Comparison between NDI and the KPMAS VDW on death date

Comparing death dates more closely, we analysed the number of NDI records above the PS cut-off, where the death dates matched exactly with the VDW, were missing in the VDW or did not match the VDW (table 3, figure 1).

Of the 38 862 above PS cut-off records, there were 1539 (4.0%) records where NDI and VDW matched in deceased vital status but had non-matching death dates. The death dates differed with a median of 2 days, IQR (1–9 days), maximum of 10 212 days. Of these, there were 1421 that also matched exactly on first and last names, and birth month, and birth year and sex and 1021 that matched on all variables.

For the below PS cut-off records (table 3), we analysed possible matching on 1858 death dates between the VDW and NDI, due to the possibility that some of these NDI records were true matches and therefore could be linked to identifiable deaths and cause of death could be obtained. We found 870 records where death dates matched exactly to our VDW. Of these, 300 also matched exactly on first and last names, and birth month, and birth year and sex; there were only 17 that mismatched on day of birth but 139 had mismatches on state of residence.

Sensitivity and specificity

The submission to NDI provided an additional 10 017 records presumed dead compared with using our VDW alone at the time of submission (tables 3 and 4). Table 4 shows a comparison of deaths found with the NDI best score algorithm to the deaths found in the VDW for all records submitted. We had 74.2% sensitivity and 99.7% specificity on death matches, with NDI as the gold standard.

PSs stratified by sex and race and ethnicity

The above cut-off scores had a higher percentage of men (55.7%) compared with below cut-off scores (52.0%), p<0.0001. Imputed race and ethnicity were significantly different between above and below cut-off scores, p<0.0001: above scores had lower percentages of A/PI and Hispanic people and had higher percentages of white and black people (table 1). Similar results were seen for self-reported race (online supplemental table 1).

Class 4 matches are independent of SSN and dependent on name matching. In multivariate linear regression analyses of Class 4 scores for above and below the PS cut-off, we found sex and imputed race and ethnicity were significantly associated with PS in each group. In the above cut-off-score group, women had slightly higher scores compared with men (online supplemental table 4). Compared with the reference white population, the imputed A/PI populations group had 4.8 times higher PS, while scores for Hispanic people were not significantly different; similar results were seen for self-reported race (online supplemental table 4). In the below cut-off-score group, women had lower PS compared with men. The imputed Hispanic and black populations had higher PS while the A/PI group and the other group had lower PS compared with the reference white population, both statistically significant; similar results were seen for self-reported race, except that the black and other group did not have statistically different PS scores compared with the white population (online supplemental table 4).

DISCUSSION

Data from the NDI can substantially improve the overall capture of deaths over other nationally available sources. We obtained information on an additional 10 017 deaths compared with the KPMAS VDW (sourced from the SSA Death Master File, some state registries and the local KPMAS EHR). However, there were deaths in the KPMAS VDW that were classified as presumed alive by the NDI algorithm. We observed variation in PSs with sex and race and ethnicity. The above cut-off scores had lower percentages of women, A/PI and Hispanic populations compared with the scores below the cut-off. For records not matched on SSN (in Class 4), NDI scores also varied by sex and by race and ethnicity.

The primary source of VDW data is the SSA death master file (SSA-DMF) which, while economic and refreshed quarterly, has been reported to have limitations in the capture of deaths, particularly for younger people and by state.[11 12] The SSA-DMF became more limited when on 1 November 2011, the SSA removed 4.2 million protected state death records and continues to add 1 million fewer annually, as a result of changes in the Social Security Act

Table 4 NDI matches compared with VDW deaths by 31 December 2017 for complete submission

	NDI*		
	Deceased	Alive	Total submitted
VDW			
Deceased	28 845	3027†	31 872
Alive	10 017	994 560	1 003 638
	38 862	997 587	1 036 449

*NDI=patients considered deceased if matching algorithm probabilistic score was above the threshold (best score); matches below the PS and no matches to the database were presumed alive.

†1188 did not match the NDI database.

NDI, National Death Index; PS, probabilistic score; VDW, Virtual Data Warehouse.

(section 205(r)).[12,13] Authors comparing NDI and the SSA Master File emphasise the importance of using multiple sources to ascertain death status for mortality studies and noted the more complete capture of death by NDI even prior to 2011.[6,12,14]

Deaths that appear in our local database but not in NDI could be explained by missing or mismatched fields in the submission which could result in a non-match or lower than expected PS. Geisinger notes that the NDI algorithm did not capture some known deaths in the World Trade Center Health Registry, with higher discrepancies among those missing SSN and non-white populations.[15] Sayer[16] showed that the NDI algorithm is sensitive to mismatch on exact first names when SSN is missing. In multiple matches to the database, the top scoring match may not be the true match. When we also examined 'presumed alive' matches, we found there were over 300 matched people who had exact matching on death date, first name, last name and birth month and year. State of residence appeared to be the most frequently mismatched field for these people. Additional cause of death on these probable matches could be obtained by resubmitting data to NDI.

The A/PI and Hispanic populations were smaller compared with the white and black populations at our institution and had a higher percent missing SSN. Arias *et al* (2016)[17] studied misclassification of race and ethnicity on death certificates used in the National Longitudinal Mortality Study and reported there was accurate race and ethnicity reporting for white and black populations during the 1999–2011 period, but there was 40% misclassification for the AI/AN population and 3% misclassification for A/PI and for Hispanic populations. When we focused on the Class 4 population which were missing SSN, we also found differences in the scores by sex and by race and ethnicity, particularly for the A/PI and Hispanic populations compared with white population. In the above-cut-off Class 4 group, the A/PI had much higher PSs compared with the white population, but the reasons for the higher scores are unclear, possibly due to less variability or missingness for other matching variables. While matches for below the cut-off may be discarded, slight changes in information available on race, sex and state of residence may be influential in accepting a match for borderline cases.

There were some limitations to this study. Several variables that could have improved matching to NDI were not used or were unavailable or were very limited at the time of submission including: self-reported race and ethnicity, father's surname, state of birth and marital status. There may have also been deaths missed due to a reporting lag by the states to NDI. Further, a review of death certificates would have been more definitive for the questionable matches. The primary strengths of this study were the large sample size, a diverse population and the ability to identify key identifiers for high scoring matches to NDI. We also demonstrated that imputed (and self-reported) race and ethnicity is correlated with missing SSN and that both race and ethnicity and sex (directly or indirectly through surname) impact matching to NDI.

CONCLUSION

In conclusion, NDI complements other sources of death data and provides increased information on vital status and cause of death. Other researchers using NDI data may benefit from a comparison group of known deaths either from SSA or a manual validation of internal data. At our institution, for records with a deceased vital status that had scores above the cut-off, over 90% matched on at least 7 digits of the SSN. However, it is important to investigate quality control measures by NDI class and matching variables, particularly for people missing SSN or with specific ethnic backgrounds. This validation step is essential to ensure the accuracy of the NDI best match algorithm and to obtain the maximum return on data submitted to NDI.

Acknowledgements We would like to thank Luther Scott (Kaiser Permanente, Portland, Oregon) for updating imputed race and ethnicity probabilities for all KPMAS members. An abstract was accepted and published for the 2020 Health Care Systems Research Network annual meeting.

Contributors MT-M, guarantor, wrote the manuscript and led analysis plan. SSB, AJD and ESW analysed and interpreted the patient data. All authors were involved in the conceptual design and acquisition of the data. All authors read, edited and approved the final manuscript.

Funding This work was supported by the Kaiser Permanente Mid-Atlantic States Community Benefits Program (RNG210864).

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval This study involves human participants and was approved by Kaiser Permanente Mid-Atlantic States Institutional Review Board (#1284326). Consent to participate was waived by the IRB.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement No data are available. The datasets generated and/or analysed during the current study contain protected health information and are not publicly available. Programming code is available from the corresponding author on reasonable request.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

ORCID iDs

Monica Ter-Minassian http://orcid.org/0000-0002-9497-2585
Sundeep S Basra http://orcid.org/0000-0003-0040-9896

REFERENCES

1 National Center for Health Statistics. In: *National Death Index user's guide*. Hyattsville, MD, 2013.
2 Rogot E, Sorlie P, Johnson NJ. Probabilistic methods in matching census samples to the national death index. *J Chronic Dis* 1986;39:719–34.
3 Skopp NA, Smolenski DJ, Schwesinger DA, *et al*. Evaluation of a methodology to validate national death index retrieval results among a cohort of U.S. service members. *Ann Epidemiol* 2017;27:397–400.

4 Williams BC, Demitrack LB, Fries BE. The accuracy of the national death index when personal Identifiers other than social security number are used. *Am J Public Health* 1992;82:1145–7.

5 Fillenbaum GG, Burchett BM, Blazer DG. Identifying a national death index match. *Am J Epidemiol* 2009;170:515–8.

6 Lash TL, Silliman RA. A comparison of the national death index and social security administration databases to ascertain vital status. *Epidemiology* 2001;12:259–61.

7 Miller EA, McCarty FA, Parker JD. Racial and ethnic differences in a linkage with the national death index. *Ethn Dis* 2017;27:77–84.

8 Elliott MN, Fremont A, Morrison PA, *et al*. A new method for estimating race/Ethnicity and associated disparities where administrative records lack self-reported race/Ethnicity. *Health Serv Res* 2008;43:1722–36.

9 Elliott MN, Morrison PA, Fremont A, *et al*. Using the Census Bureau's surname list to improve estimates of race/Ethnicity and associated disparities. *Health Serv Outcomes Res Method* 2009;9:69–83.

10 Curtis MD, Griffith SD, Tucker M, *et al*. Development and validation of a high-quality composite real-world mortality endpoint. *Health Serv Res* 2018;53:4460–76.

11 Hill ME, Rosenwaike I. The social security administration's death master file: the completeness of death reporting at older ages. *Soc Secur Bull* 2001;64:45–51.

12 Navar AM, Peterson ED, Steen DL, *et al*. Evaluation of mortality data from the social security administration death master file for clinical research. *JAMA Cardiol* 2019;4:375–9.

13 U.S. Department of Commerce National Technical Information Service. Important Notice: Change in Public Death Master File Records. Alexandria, VA, 2011.

14 Pollack AZ, Hinkle SN, Liu D, *et al*. Vital status ascertainment for a historic diverse cohort of U.S. women. *Epidemiology* 2020;31:310–6.

15 Giesinger I, Li J, Takemoto E, *et al*. Confirming mortality in a longitudinal exposure cohort: optimizing national death index search result processing. *Ann Epidemiol* 2021;56:40–6.

16 Sayer B. 2006 Comparing Bigmatch results to current national death index (NDI) selection methods. *Proc Sur Res Methods Section*;2006:3648–55.

17 Arias E, Heron M, Hakes JK. The validity of race and Hispanic-origin reporting on death certificates in the United States: an update. *Vital Health Stat* 2016;2.

BMJ Health & Care Informatics

Social vulnerability and initial COVID-19 community spread in the US South: a machine learning approach

Moosa Tatar ,[1] Mohammad Reza Faraji,[2] Fernando A Wilson[3,4,5]

To cite: Tatar M, Faraji MR, Wilson FA. Social vulnerability and initial COVID-19 community spread in the US South: a machine learning approach. *BMJ Health Care Inform* 2023;**30**:e100703. doi:10.1136/bmjhci-2022-100703

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/bmjhci-2022-100703).

Received 18 November 2022
Accepted 15 June 2023

[1]Center for Value-Based Care Research, Cleveland Clinic, Cleveland, Ohio, USA
[2]Department of Computer Science and Information Technology, Institute for Advanced Studies in Basic Sciences, Zanjan, Iran
[3]Matheson Center for Health Care Studies, University of Utah, Salt Lake, Utah, USA
[4]Department of Population Health Sciences, University of Utah, Salt Lake City, Utah, USA
[5]Department of Economics, University of Utah, Salt Lake City, Utah, USA

Correspondence to
Dr Moosa Tatar; tatarm@ccf.org

ABSTRACT

Background and objectives More than 93 million COVID-19 cases and more than 1 million COVID-19 deaths have been reported in the USA by August 2022. The disproportionate effect of the pandemic and its severe impact on vulnerable communities raised concerns. This research aimed to identify and rank Social Vulnerability Index (SVI) factors highly predictive of the spread of COVID-19 in the US South at the beginning of the pandemic.

Methods We used Extreme Gradient Boosting (XGBoost) machine learning methodology and SVI data, and the number of COVID-19 cases across all counties in the US South to predict the number of positive cases within 30 days of a county's first case.

Results Our results showed that the percentage of mobile homes is the most important feature in predicting the increase in COVID-19. Also, population density per square mile, per capita income, percentage of housing in structures with 10+ units, percentage of people below poverty and percentage of people with no high school diploma are important predictors of COVID-19 community spread, respectively.

Conclusions SVI can help assess the vulnerability or resilience of communities to the spread of COVID-19 and can help identify communities at high risk of COVID-19 spread.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Social and economic factors influence vulnerability to infection and health outcomes and the severe impact of COVID-19 on vulnerable communities.

WHAT THIS STUDY ADDS

⇒ Percentage of mobile homes within a county, population density per square mile and per capita income are important predictors of community spread of COVID-19.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ The Social Vulnerability Index can help assess the resilience of communities to the spread of COVID-19 and can help identify communities at high risk of COVID-19 spread.

INTRODUCTION

More than 93 million COVID-19 cases and more than 1 million COVID-19 deaths have been reported in the USA by August 2022.[1] The pandemic has disproportionally affected minority communities at the local level.[2] Even at the early stages of the pandemic, the severe impact of COVID-19 on vulnerable communities raised concerns.[3] Historically, poverty, inequalities and social determinants of health facilitate the spread of infectious diseases.[4] There is evidence that socioeconomic factors may influence the spatial spread of COVID-19 at the county level.[5] Past pandemics also have shown that social and economic factors influence vulnerability to infection and health outcomes.[6] Further, individuals residing in deprived neighbourhoods (ie, neighbourhoods with higher poverty, lower education, low housing quality and low employment rates) had a higher risk of COVID-19 infection.[7] Also, a recent study analysed the association of social, economic and demographic factors in the initial spread of COVID-19 and reported that social and economic factors are strongly and positively associated with COVID-19.[8]

Many communities in the US South have substantial social vulnerabilities that may worsen the impact of COVID-19. In recent weeks, the US South has become a major region of community spread, ranging from Florida to Texas (figure 1). While studies suggest effective policies, including lockdowns and mandatory mask use, that are effective for controlling the spread of COVID-19 in communities,[9 10] in several of these states, lack of consistent and effective public policies to mitigate infection spread has been a source of debate. In Georgia, for example, the governor filed a lawsuit (later dropped) against the mayor of Atlanta in order to prevent the latter's enforcement of a mask mandate.[11] The city of Atlanta is racially diverse and minority

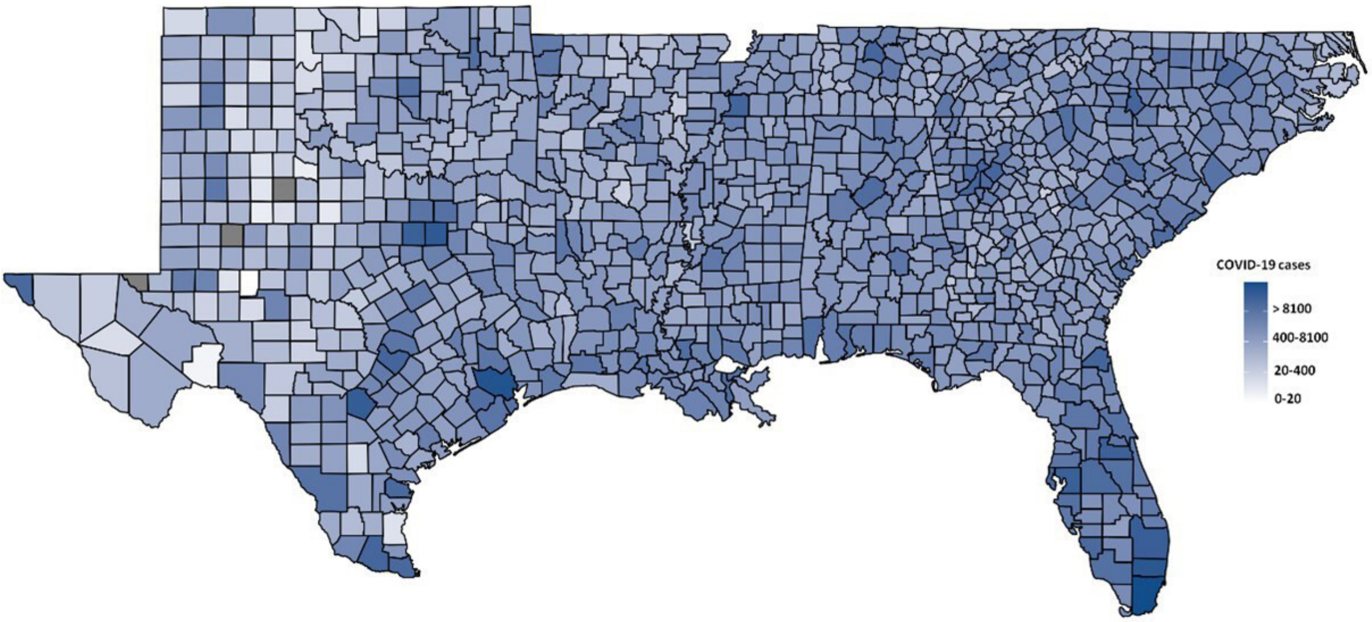


Figure 1 County-level distribution of COVID-19 cases in the US South (August 2020). US South region includes the states of Alabama, Arkansas, Florida, Georgia, Louisiana, Mississippi, North Carolina, Oklahoma, South Carolina, Tennessee and Texas.

communities have experienced both high rates of poverty and other socioeconomic vulnerabilities as well as COVID-19 community spread.

Social vulnerability is the resilience of communities against disease outbreaks and natural or human-caused disasters.[12] It is applicable to identify communities most at risk when faced with adverse events that may impact health (eg, disease outbreaks). Social vulnerability refers to socioeconomic and demographic factors that affect a community's ability and power to prevent human suffering in the event of disaster or outbreaks. The Centers for Disease Control and Prevention (CDC) categorises these socioeconomic and demographic factors into four overall vulnerability domains: socioeconomic status, household composition, and disability, minority status and language, and housing type and transportation.[13] The Social Vulnerability Index (SVI) provides social and spatial information to help public health officials and local emergency response planners to identify communities at high risk of being adversely affected during a crisis.[13] This information helps communities to prepare for a better response to emergency events especially disease outbreaks.[12 13] SVI was associated with increased rates of COVID-19.[14] Also, counties with the highest SVI had a greater risk of COVID-19 infection and death,[3] and most vulnerable counties had higher death rates, especially at the beginning of the pandemic.[15]

Although race/ethnic minority communities have been disproportionately impacted by COVID-19,[3 6 16 17] the role of specific social vulnerabilities such as poverty, housing insecurity and other issues faced in these communities that contribute to the spread of infection at the beginning of the pandemic and spread of the COVID-19 virus is unclear. To address this gap in knowledge, we use machine learning-based analyses of the SVI data to identify and rank SVI factors that are highly predictive of the spread of COVID-19 cases at the county level across 11 states in the US South.

METHODS

Study setting and design

This machine learning-based study included COVID-19 cases and 16 social vulnerability features for all counties across 11 US states located in the South, including: Alabama, Arkansas, Florida, Georgia, Louisiana, Mississippi, North Carolina, Oklahoma, South Carolina, Tennessee and Texas (online supplemental figures A1,A2). To investigate the association of social vulnerability factors and the spread of COVID-19 at the county level, we use an effective prediction algorithm regression method. We regress the number of COVID-19 cases 30 days after the first confirmed COVID-19 case in each county against social vulnerability features (detailed below). We chose to examine the US South because of the number of major COVID-19 'hot spots' located in that region as well as the region's long-standing historical socioeconomic inequities across minority and non-minority communities.[18]

Study sample and data

We used daily COVID-19 cases from January 2020 to August 2020 from the official website of Johns Hopkins University's Coronavirus Resource Center.[1] For each county in the US South (1086 counties), we identified the number of COVID-19 cases 30 days after their first COVID-19 case was confirmed.

We also used the latest SVI data available from the CDC released in 2018.[13] We used 16 social vulnerability features as independent variables: percentage of people below poverty, unemployment rate, per capita income, percentage of people with no high school diploma, percentage of people aged 65 and older, percentage of people aged 17 and younger, percentage of non-institutionalised people with a disability, percentage of single-parent households with children, percentage of minority people (except white, non-Hispanic), percentage of people aged 5+ who speak limited English, percentage of housing in structures with 10+ units, percentage of mobile homes, percentage of overoccupied housing units, percentage of households with no vehicle available, percentage of institutionalised group quarters (eg, correctional institutions, nursing homes) and population density per square mile (see online supplemental table A1 for definitions). All data used in the manuscript are publicly available.

Statistical analysis

We used Extreme Gradient Boosting (XGBoost) to predict the number of positive cases within 30 days of a county's first case. XGBoost is a scalable machine learning system using gradient tree boosting which is available as an open source software package.[19] Chen and Guestrin presented the XGBoost algorithm in 2016.[20] XGBoost is a highly effective and widely used machine learning method that can be used for regression, classification and prediction.[20] Gradient boosted decision trees (GBDT) are an ensemble learning method (ie, a method that aggregates the predictions of a group of predictors) which uses decision trees as their base predictor and sequentially adds decision trees to the ensemble, while each added tree improves the fit of its predecessor to the data.[21] XGBoost benefits from several innovations and optimisation techniques to add scalability to GBDT, making it faster and yielding better performance. In this study, the XGBoost algorithm is used to predict COVID-19 cases as the sum of predictions from thousands of individual decision trees, with each trained on the residual of all previous trees and making marginal improvements to the overall model prediction.[19 21]

While XGBoost learns from the training data and makes predictions with the testing data, it also uses different importance metrics to produce an importance matrix that contains the information gain, cover and frequency of features that have been actually used in the boosted trees. The interpretation of prediction results and how features contribute to the prediction is based on these three importance metrics. Gain is the most relevant attribute to interpret the relative importance of each feature and denotes the relative contribution of a feature in explaining variation in outcomes within the model, that is, a higher feature gain implies that the feature is more important for generating the prediction. Cover denotes the average coverage (the relative number of counties affected) of splits which use a specific feature. It simply corresponds to the percentage of the counties which the feature is used to decide the leaf node for them. Frequency is the percentage representing the relative number of times a specific feature occurs across all the trees estimated within the model.[22] All measures are reported as relative amounts and hence all sum up to 1.

A subset of 869 counties (80% of the total 1086 counties) were used as our training data set, and 217 counties (20% of all counties) were used for our testing data set. We used 10-fold cross-validation, which is a commonly used statistical method in applied machine learning methods, to tune the model's hyperparameters. Cross-validation assesses how the results of a statistical analysis will generalise to an independent data set and tests the model's ability to predict with a new data set. It also points out problems like overfitting or selection bias.[23] Tenfold cross-validation divided the training sample into 10 parts; the model is trained on nine parts (90% of the 869 counties), and performance is measured by the ability to accurately predict COVID-19 cases by the remaining part (the other 10% of 869 counties). When the hyperparameters of the XGBoost model are tuned, the XGBoost is trained using the tuned parameters on all the 869 counties. Finally, the model is used to predict the outcomes (ie, number of positive COVID-19 cases after 30 days of the county's first confirmed case) for the test data (ie, the 217 counties). We also conducted a SHapley Additive exPlanations (SHAP) analysis to explain the predictions of machine learning models. A positive SHAP value means a positive impact of the features on prediction. Finally, for the sensitivity analysis the model was used to predict the outcomes that was number of positive COVID-19 cases after 60 days of the county's first confirmed case. We used the RStudio V.4.0.2 (R Core Team, 2020) statistical package for all analyses.

RESULTS

Table 1 provides sample characteristics of the 16 SVIs and COVID-19 cases and COVID-19 rates per 100 000 population after 30 days of the first COVID-19-positive cases in all counties in the 11 states of the US South (1086 counties). On average, 85.3 COVID-19 cases were reported after 30 days of the first reported case in a county, and a maximum of 6119 COVID-19 cases after 30 days of the first case in a county. Also, on average, 139.5 COVID-19 cases per 100 000 population were reported after 30 days of the first reported case, and a maximum of 4026.8 COVID-19 cases per 100 000 population after 30 days of the first case in a county.

To evaluate the accuracy of our model, we tested the reliability of our predictions on 217 counties in the test data set. Goodness of fit and prediction evaluation (adjusted R-squared=0.59, root mean square error (RMSE)=92.36) indicates that the model was robust (online supplemental table A2). Online supplemental

Table 1 Descriptive statistics of the 16 SVIs and COVID-19 cases and COVID-19 rates per 100 000 population after 30 days of the first COVID-19-positive cases in all counties in the US South (1086 counties)

Feature	Min	Median	Mean	Max	SD
Below poverty, %	2.6	17.9	18.8	49.7	6.4
Unemployment rate, %	0	6.4	6.8	25.8	2.8
Per capita income	12 292	23 540	24 183	50 931	5078
No high school diploma, %	4.4	16.9	17.5	66.3	6.1
Aged 17 and younger, %	7.3	22.9	22.8	36.6	3.2
Non-institutionalised with a disability, %	5.3	17.2	17.3	31	4.1
Single-parent households with children, %	0	9.1	9.3	22.7	2.8
Minority (except white, non-Hispanic), %	1.1	34.1	35.4	99.3	19.7
Aged 5+ who speak limited English, %	0	1.1	2.2	30.4	3.4
Housing in structures with 10+ units, %	0	1.8	3.5	38.5	4.7
Mobile homes, %	0.5	18.0	18.9	59.3	9.9
Overoccupied housing units, %	0	2.4	2.8	21.2	1.8
Households with no vehicle available, %	0	5.9	6.4	20.3	2.9
Institutionalised group quarters, %	0	1.8	3.9	36.5	5.3
Population density per square mile	0.2	49.2	147.1	3499.1	318.9
COVID-19 cases after 30 days	0	23	85.3	6119	317.7
COVID-19 cases per 100 000 population after 30 days	0	63.6	139.5	4026.8	250.9

SVI, Social Vulnerability Index.

figure A5 also shows calibration plot of the predicted versus observed COVID-19 rates. Figure 2 shows the result of XGBoost gain relative importance. The percentage of mobile homes in counties is the most important feature, followed by population density per square mile and per capita income, in predicting the growth of COVID-19 within 30 days of the first case. The relative contributions of percentage of mobile homes, population density per square mile and per capita income to the model for generating predictions are 0.35, 0.12 and 0.12, respectively. Percentage of housing in structures with 10+ units, percentage of people below poverty and percentage of people with no high school diploma have relative contributions of 0.10, 0.08 and 0.04, respectively. The percentage of overoccupied housing units and the percentage of institutionalised group quarters are the

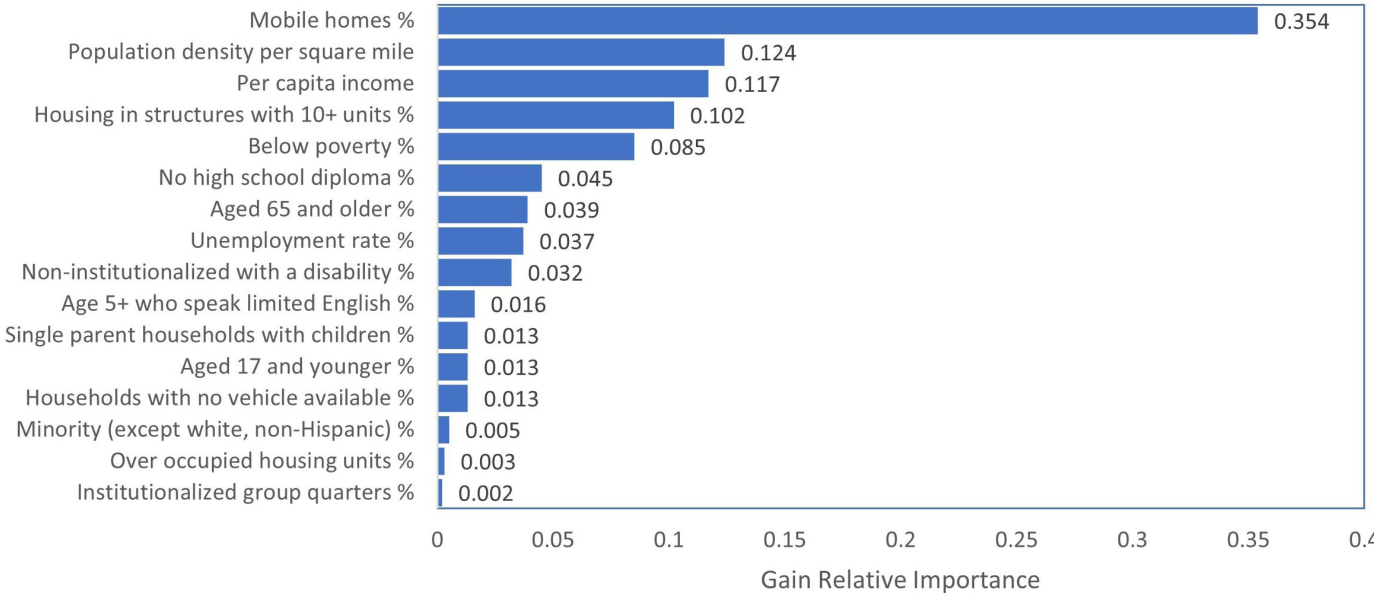


Figure 2 Extreme Gradient Boosting (XGBoost) gain relative importance. The measures are all reported as relative amounts and all sum up to 1.0.

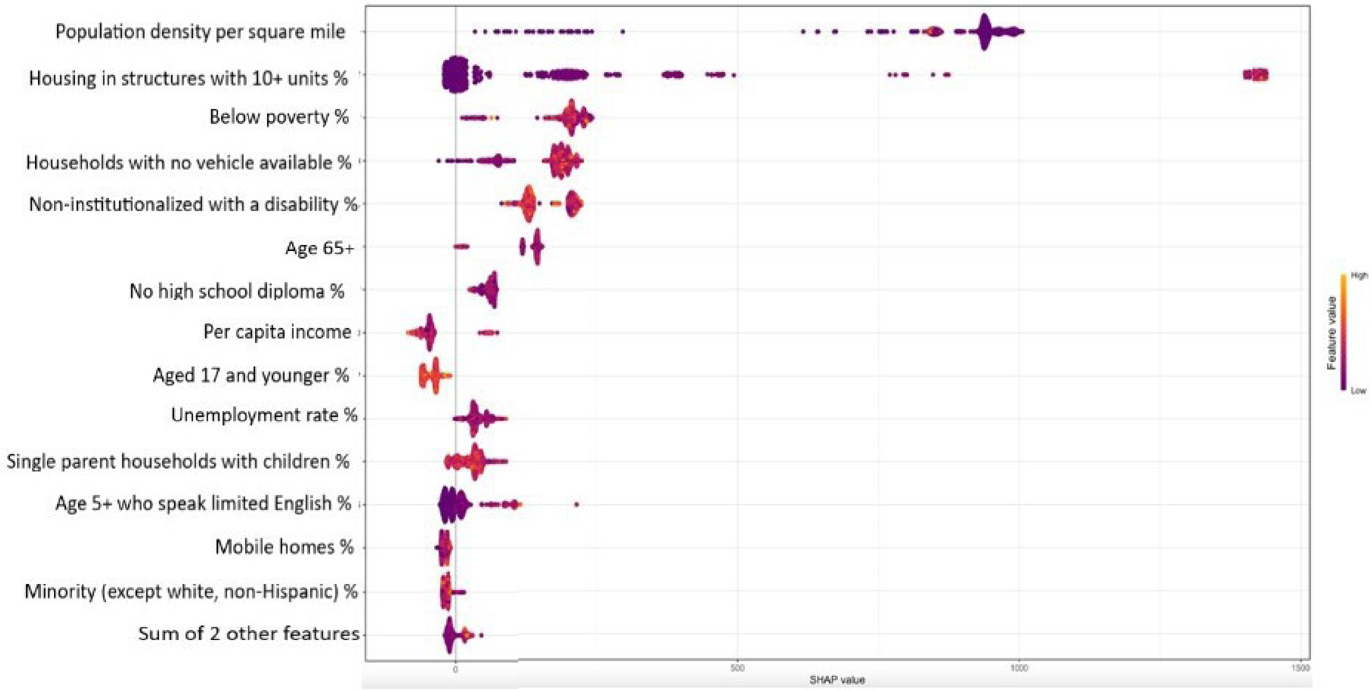


Figure 3 Shapley additive explanations (SHAP) analysis results.

least important features in the model with relative gains of 0.003 and 0.002, respectively.

The relative cover for percentage of mobile homes, population density per square mile and per capita income is 0.09, 0.12 and 0.07, respectively, which shows the relative proportion of counties in our sample that include these features across all the decision trees (online supplemental figure A3). Also, the relative cover for percentage of housing in structures with 10+ units, percentage of people below poverty and percentage of people with no high school diploma is 0.7, 0.06 and 0.06, respectively. Relative frequency is calculated as the proportion of decision tree nodes that include a specific feature. The result of relative frequency shows that percentage of mobile homes, population density per square mile and per capita income occurred in 0.069, 0.093 and 0.079 of nodes within the trees of the model, respectively (online supplemental table A4). In addition, percentage of housing in structures with 10+ units, percentage of people below poverty and percentage of people with no high school diploma accounted for 0.059, 0.085 and 0.061 of nodes in the trees of the model, respectively. Additional XGBoost feature importance matrix details can be found in online supplemental table A3. Figure 3 shows the results of the SHAP analysis. Population density per square mile, percentage of housing in structures with 10+ units and percentage of people below poverty had the most positive impact on the number of COVID-19 cases in a county. Also, per capita income and aged 17 and younger features had the most negative impact on the number of COVID-19 cases in a county.

Online supplemental table A4 shows the result of XGBoost gain relative importance after 60 days of the county's first COVID-19 case. The population density per square mile in counties is the most important feature in predicting the growth of COVID-19 within 60 days of the first case with a relative gain of 31.8%. This is followed by percentage of housing in structures with 10+ units and percentage of mobile homes, with relative gains of 30.4% and 11.2%, respectively. Also, percentage of people aged 65 and older, per capita income and percentage of people aged 5+ who speak limited English have relative contributions of 5.5%, 4.9% and 2.6%, respectively. Additional XGBoost feature importance matrix details can be found in online supplemental table A4.

DISCUSSION

Our machine learning study used SVI data and number of COVID-19 cases across all counties in the US South to analyse the association of social vulnerability features in predicting the community spread of infection. Our analysis suggests that the percentage of mobile homes within a county is the most important feature in predicting the increase in COVID-19. This was followed by population density per square mile and per capita income. Percentage of housing in structures with 10+ units, percentage of people below poverty and percentage of people with no high school diploma were also important predictors of community spread. However, the percentage of large, multifamily housing units and the percentage of institutionalised group quarters were the least important features in predicting COVID-19 spread at the county level.

Our findings are consistent with the results from prior studies that investigated COVID-19 cases and socioeconomic factors and considered the impact of the pandemic on racial and ethnic minorities.[2 3 16 24 25] Studies report a disproportionate rate of infections and deaths among non-Hispanic Blacks and Hispanics.[2 25] For example, a recent study found that minority status and language, household composition and transportation, and housing and disability were associated with the number of COVID-19 cases in the USA.[25] Poverty, crowded housing and lack of vehicle ownership were reported to be associated with increased COVID-19 cases and deaths in urban areas. Also, high population densities catalyse the spread of COVID-19; therefore, avoiding situations with higher population densities will limit the spread of COVID-19.[26] In addition, in rural communities, minority status and language are associated with increases in COVID-19 cases.[3] Another study reported that counties with a higher percentage of minority, high-density housing structures and crowded housing units were at higher risk of becoming a COVID-19 hot spot.[27] A study of urban-rural differences in COVID-19 exposures and outcomes in South Carolina has shown a positive correlation between the case rates, mortality rates and pre-existing social vulnerability. Also, a negative correlation between mortality rates and county resilience patterns suggests that counties with higher levels of inherent resilience had lower death rates.[28]

Although the US South has numerous hot spots of community spread of COVID-19, there are a few prior studies that have systematically investigated the initial spread of COVID-19 in relation to social vulnerabilities across counties in the region. A recent study investigated the spatial association of social vulnerability with COVID-19 prevalence and reported a spatially varying relationship between SVI and COVID-19 cases and deaths.[29] Further, our use of a machine learning approach helped determine the specific community vulnerabilities that are most salient in determining the rapid spread of COVID-19. One study reported that mobility habits (eg, number of citizens who make at least one trip per day; transport accessibility; distance from the main city clusters) have a positive association for the spread of COVID-19.[30] A recent study also forecasted the geographic spread of COVID-19 as a communicable disease by using social structure of networks.[31] Aggregated data from Facebook also showed that COVID-19 cases were more likely to spread between regions that had stronger social network connections.[32] Google COVID-19 Community Mobility Reports also provide a new tool to assess the role of policies to mitigate community spread (eg, to work from home, shelter in place and other recommendations) in flattening the curve of the COVID-19 pandemic.[33]

This study is subject to limitations. The results of this study should not be interpreted in a causality context. There are various state and local policies (eg, lockdown, business closure and facial mask mandate) that may have impacted our findings. Hence, residual confounding should be considered due to omission of important covariates. Also, the number of COVID-19 cases in a county might affect the number of cases in neighbouring counties through the connection between counties. Finally, our results are regional and may not generalise to other regions of the USA. With the availability of various free COVID-19 vaccines, the USA still struggles to fight the pandemic, and new waves of COVID-19 are an ongoing threat to public health in the USA. More studies are needed to investigate the resilience of vulnerable counties against COVID-19.

CONCLUSIONS

Our findings showed that SVI can help assess the vulnerability or resilience of communities to the spread of COVID-19. Thus, our results can help identify communities at high risk of spread and aid in policy efforts tailored to addressing these communities' specific vulnerabilities to COVID-19. An understanding of the role social vulnerabilities have in determining the spread of COVID-19 is critical for forecasting the trajectory of this disease and designing effective mitigation interventions at the community level.

Contributors MT as the corresponding author accepts full responsibility for the work and/or the conduct of the study, had access to the data, and controlled the decision to publish. MT and MRF performed the statistical analyses and had full access to all study data. All authors (MT, MRF and FAW) contributed to concept and design, provided data acquisition and interpretation, drafted the manuscript and critically revised the manuscript for important intellectual content. All authors approved the final version of the manuscript and are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. MT, MRF and FAW jointly supervised this work.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Map disclaimer The inclusion of any map (including the depiction of any boundaries therein), or of any geographic or locational reference, does not imply the expression of any opinion whatsoever on the part of BMJ concerning the legal status of any country, territory, jurisdiction or area or of its authorities. Any such expression remains solely that of the relevant source and is not endorsed by BMJ. Maps are provided without any warranty of any kind, either express or implied.

Competing interests None declared.

Patient consent for publication Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available in a public, open access repository. Data are available in a public, open access data set. Data are available in the GitHub through Novel Coronavirus (COVID-19) Cases provided by JHU CSSE: https://github.com/CSSEGISandData/COVID-19.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

ORCID iD

Moosa Tatar http://orcid.org/0000-0002-0342-4293

REFERENCES

1 Center for Systems Science and Engineering (CSSE). Global cases by the center for systems science and engineering (CSSE) at Johns Hopkins University (JHU)" Johns Hopkins CSSE. Available: https://github.com/CSSEGISandData/COVID-19 [Accessed 24 Aug 2022].

2 Kim SJ, Bostwick W. Social vulnerability and racial inequality in COVID-19 deaths in Chicago. *Health Educ Behav* 2020;47:509–13.

3 Khazanchi R, Beiter ER, Gondi S, *et al*. County-level association of social vulnerability with COVID-19 cases and deaths in the USA. *J Gen Intern Med* 2020;35:2784–7.

4 Singu S, Acharya A, Challagundla K, *et al*. Impact of social determinants of health on the emerging COVID-19 pandemic in the United States. *Front Public Health* 2020;8:406.

5 BaumCFHenry M. Socioeconomic factors influencing the spatial spread of COVID-19 in the United States. *SSRN Journal* 2020.

6 Clark E, Fredricks K, Woc-Colburn L, *et al*. Disproportionate impact of the COVID-19 pandemic on immigrant communities in the United States. *PLoS Negl Trop Dis* 2020;14:e0008484.

7 K C M, Oral E, Straif-Bourgeois S, *et al*. The effect of area deprivation on COVID-19 risk in Louisiana. *PLoS ONE* 2020;15:e0243028.

8 Mogi R, Spijker J. The influence of social and economic ties to the spread of COVID-19 in Europe. *J Popul Res (Canberra)* 2022;39:495–511.

9 Ayouni I, Maatoug J, Dhouib W, *et al*. Effective public health measures to mitigate the spread of COVID-19: a systematic review. *BMC Public Health* 2021;21:1015.

10 Huang X, Shao X, Xing L, *et al*. The impact of lockdown timing on COVID-19 transmission across US counties. *EClinicalMedicine* 2021;38:101035.

11 Reuters Staff. Georgia governor to drop mask lawsuit against Atlanta Mayor and city. Retures; 2020. Available: https://www.reuters.com/article/us-health-coronavirus-usa-georgia/georgia-governor-to-drop-mask-lawsuit-against-atlanta-mayor-and-city-idUSKCN2592VV [Accessed 13 Aug 2020].

12 Flanagan BE, Gregory EW, Hallisey EJ, *et al*. A social vulnerability index for disaster management. *J Homel Secur Emerg Manag* 2011;8.

13 Centers for Disease Control and Prevention. CDC's social vulnerability index (SVI). 2018. Available: https://svi.cdc.gov [Accessed 22 Jun 2020].

14 Karaye IM, Horney JA. The impact of social vulnerability on COVID-19 in the US: an analysis of spatially varying relationships. *Am J Prev Med* 2020;59:317–25.

15 Neelon B, Mutiso F, Mueller NT, *et al*. Spatial and temporal trends in social vulnerability and COVID-19 incidence and death rates in the United States. *PLoS One* 2021;16:e0248702.

16 Alcendor DJ. Racial disparities-associated COVID-19 mortality among minority populations in the US. *J Clin Med* 2020;9:2442.

17 Tai DBG, Shah A, Doubeni CA, *et al*. The disproportionate impact of COVID-19 on racial and ethnic minorities in the United States. *Clin Infect Dis* 2021;72:703–6.

18 Whyte LE. Unpublicized recommendations say states should return to stringent control measures exclusive: White House document. The Center for Public Integrity; 2020.

19 Chen T, He T, Benesty M, *et al*. Xgboost: extreme gradient boosting. R package version 04-2; 2015. 1–4.

20 Chen T, Guestrin C. Xgboost: A Scalable tree boosting system. *Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining* 2016.

21 Géron A. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. O'Reilly Media, 2019.

22 Achiron A, Gur Z, Aviv U, *et al*. Predicting refractive surgery outcome: machine learning approach with big data. *J Refract Surg* 2017;33:592–7.

23 Cawley GC, Talbot NL. On over-fitting in model selection and subsequent selection bias in performance evaluation. *J Mach Learn Res* 2010;11:2079–107.

24 Gaynor TS, Wilson ME. Social vulnerability and equity: the disproportionate impact of COVID-19. *Public Adm Rev* 2020;80:832–8.

25 Karaye IM, Horney JA. The impact of social vulnerability on COVID-19 in the U.S.: an analysis of spatially varying relationships. *Am J Prev Med* 2020;59:317–25.

26 Rocklöv J, Sjödin H. High population densities catalyse the spread of COVID-19. *J Travel Med* 2020;27:taaa038.

27 Dasgupta S, Bowen VB, Leidner A, *et al*. Association between social vulnerability and a county's risk for becoming a COVID-19 hotspot - United States, June 1-July 25, 2020. *MMWR Morb Mortal Wkly Rep* 2020;69:1535–41.

28 Huang Q, Jackson S, Derakhshan S, *et al*. Urban-rural differences in COVID-19 exposures and outcomes in the south: a preliminary analysis of South Carolina. *PLoS ONE* 2021;16:e0246548.

29 Wang C, Li Z, Clay Mathews M, *et al*. The spatial association of social vulnerability with COVID-19 prevalence in the contiguous United States. *Int J Environ Health Res* 2022;32:1147–54.

30 Cartenì A, Di Francesco L, Martino M. How mobility habits influenced the spread of the COVID-19 pandemic: results from the Italian case study. *Sci Total Environ* 2020;741:140489.

31 y Piontti AP, Perra N, Rossi L, *et al*. *Charting the Next Pandemic: Modeling Infectious Disease Spreading in the Data Science Age*. Springer, 2018.

32 Kuchler T, Russel D, Stroebel J. The geographic spread of COVID-19 correlates with structure of social networks as measured by Facebook. National Bureau of economic; 2020. 0898–2937.

33 Aktay A, Bavadekar S, Cossoul G, *et al*. Google COVID-19 community mobility reports: anonymization process description. *arXiv* 2020.