



Varying association of laboratory values with reference ranges and outcomes in critically ill patients: an analysis of data from five databases in four countries across Asia, Europe and North America

Haoran Xu,¹ Louis Agha-Mir-Salim ^{2,3} Zachary O'Brien,⁴ Dora C Huang,⁵ Peiyao Li,^{6,7} Josep Gómez,^{8,9} Xiaoli Liu,^{2,10} Tongbo Liu,¹¹ Wesley Yeung,^{2,12} Patrick Thorat,¹³ Paul Elbers,¹³ Zhengbo Zhang,¹⁴ María Bodí Saera,^{8,9} Leo Anthony Celi ^{2,15}

To cite: Xu H, Agha-Mir-Salim L, O'Brien Z, *et al*. Varying association of laboratory values with reference ranges and outcomes in critically ill patients: an analysis of data from five databases in four countries across Asia, Europe and North America. *BMJ Health Care Inform* 2021;**28**:e100419. doi:10.1136/bmjhci-2021-100419

► Additional supplemental material is published online only. To view, please visit the journal online (<http://dx.doi.org/10.1136/bmjhci-2021-100419>).

HX and LA-M-S are joint first authors.

Received 27 May 2021
Accepted 17 September 2021



© Author(s) (or their employer(s)) 2021. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

For numbered affiliations see end of article.

Correspondence to

Dr Louis Agha-Mir-Salim;
mirsalim@mit.edu

ABSTRACT

Background Despite wide usage across all areas of medicine, it is uncertain how useful standard reference ranges of laboratory values are for critically ill patients.

Objectives The aim of this study is to assess the distributions of standard laboratory measurements in more than 330 selected intensive care units (ICUs) across the USA, Amsterdam, Beijing and Tarragona; compare differences and similarities across different geographical locations and evaluate how they may be associated with differences in length of stay (LOS) and mortality in the ICU.

Methods A multi-centre, retrospective, cross-sectional study of data from five databases for adult patients first admitted to an ICU between 2001 and 2019 was conducted. The included databases contained patient-level data regarding demographics, interventions, clinical outcomes and laboratory results. Kernel density estimation functions were applied to the distributions of laboratory tests, and the overlapping coefficient and Cohen standardised mean difference were used to quantify differences in these distributions.

Results The 259 382 patients studied across five databases in four countries showed a high degree of heterogeneity with regard to demographics, case mix, interventions and outcomes. A high level of divergence in the studied laboratory results (creatinine, haemoglobin, lactate, sodium) from the locally used reference ranges was observed, even when stratified by outcome.

Conclusion Standardised reference ranges have limited relevance to ICU patients across a range of geographies. The development of context-specific reference ranges, especially as it relates to clinical outcomes like LOS and mortality, may be more useful to clinicians.

INTRODUCTION

The care of critically ill patients relies heavily on laboratory data—and, by extension, the laboratory reference ranges associated with them. However, these laboratory reference ranges are typically created by surveying

Summary

What is already known?

- Laboratory results of critically ill patients are interpreted using reference ranges created on the basis of healthy outpatients.
- Correcting abnormal laboratory results to reference range standards can have beneficial or harmful effects.

What does this paper add?

- Laboratory results of critically ill patients often differ significantly from the reference range, even in those with the best clinical outcomes.
- Critically ill patients may require local, context-specific reference ranges for laboratory results to promote appropriate interpretation.

healthy outpatients.¹ It remains unclear if these ranges are applicable to patients admitted to the intensive care unit (ICU).

Previous studies have shown that correcting abnormal values in critically ill patients, such as haemoglobin or glucose, to reference range standards may be harmful.^{2–5} For example, clearly defined thresholds have been established for the initiation of packed red blood cell (PRBC) transfusions.^{6–8} However, observational studies show that PRBCs are routinely administered at higher haemoglobin levels,^{7 9 10} suggesting that clinicians may strive to correct laboratory values towards normality rather than adhere to evidence-based targets. As previously hypothesised, specific reference ranges tailored to scenarios and populations may be more meaningful, if these reference ranges are shown to relate to clinical outcome.^{11–13}

Previous research considered whether the distributions of laboratory values for critically ill patients differed from reference ranges and if these differences were associated with outcomes. The single-centre, cross-sectional study found that laboratory values of ICU patients differed significantly from the reference range, even in those with the best clinical outcomes, suggesting that normal reference ranges may not apply to critically ill patients in reference or outcome.¹⁴ This adds to ongoing discussions regarding the need to consider context in interpreting laboratory values, particularly in critical care settings, and advocates for further research into contextualising laboratory values.^{11 15}

This study aims to expand on previous work by evaluating data from five ICU databases located across different continents to consider if similar patterns hold worldwide. In particular, this work aims to characterise how ICU laboratory values differ from typical reference ranges in ICUs across the USA, Netherlands, China and Spain, and determine if the relationship between laboratory values and patient outcomes varies across contexts.

METHODS

Design

We conducted an international, multi-centre, retrospective, cross-sectional study examining the most severely deranged (minimum or maximum as appropriate) laboratory results within the first 24 hours of a patient's first admission to the ICU.

Setting

We included all patients from five ICU databases: the Medical Information Mart for Intensive Care (MIMIC), the eICU Collaborative Research Database (eICU-CRD), the Amsterdam University Medical Center database (AUMCdb), the Chinese PLA General Hospital ICU database (PLAGH-ICUdb) and the Unitat de Cures Intensives de l'Hospital Joan XXIII database (UCIHJ23db). An overview of all databases is displayed in online supplemental table 1.

MIMIC contains data from the Beth Israel Deaconess Medical Center ICU, a tertiary hospital located in Boston, Massachusetts, USA, which comprises more than 70 beds with a broad case mix. We used data from the latest version available at the time of analysis, MIMIC-III, which contained granular data on more than 38 000 admissions between 2001 and 2012.¹⁶ eICU-CRD contains similarly detailed, patient-level data on more than 200 000 admissions across 335 ICUs in the USA between 2014 and 2015.¹⁷ AUMCdb is the first freely accessible European ICU database and contains data from the Amsterdam University Medical Center ICU, a mixed medical-surgical ICU with data on more than 20 000 patients including admissions between 2003 and 2016 (V.1.0.2).¹⁸ The PLAGH-ICUdb integrates data from nine ICUs in the Chinese People's Liberation Army General Hospital in Beijing, China.¹⁹ PLAGH-ICUdb includes data on more than 74 000 adult

patients admitted between 2008 and 2019. Finally, the UCIHJ23db includes 4840 admissions between 2015 and 2019 from the Joan XXIII University Hospital in Tarragona, Spain.

Primary analysis

Our analysis was performed between February and May 2020. We included the first ICU admission of all adult patients from the five databases. Patients were stratified into those with the 'best' and 'worst' clinical outcomes. The best outcome group was defined as patients who survived the ICU admission and had an ICU length of stay (LOS) in the lowest quartile (shortest LOS). The worst outcome group was defined as those who died during the ICU admission.

For each patient, we extracted the most severely deranged laboratory values of commonly ordered investigations collected within the first 24 hours of ICU admission. The included investigations were maximum creatinine, minimum haemoglobin, maximum lactate and maximum sodium. No imputation was performed to replace missing data. We calculated the 95% CI for each investigation, stratified by the best and worst outcome patients within each database, and presented these as distribution plots. The locally used normal reference range for each investigation was added to these plots, to allow a visual assessment of the variance between these reference ranges and patient outcomes. We compared the difference in laboratory result distributions between the best and worst outcome groups by calculating the degree of overlap and divergence.

Statistical analysis

Data extraction was performed using SQL. Statistical analysis was then conducted using R and Python. The queries and code used for analyses were uploaded to a public GitHub repository.²⁰ Kernel density estimation plots were used to present the distribution of laboratory results. To then quantify the difference in distribution between best and worst outcome groups, we calculated the overlapping coefficient (OVL) and the Cohen standardised mean difference (SMD), as have been used for this purpose previously.^{14 21 22} OVL quantifies the overlap of two distributions, with an OVL of 1 representing complete overlap and an OVL of 0 representing no overlap. SMD describes the difference in group means, relative to the variability observed within each group. The SMD value represents the divergence between groups in SD. An SMD of 0 indicates no difference in the means of the two groups; less than 0.2 is considered a small effect size, 0.2 to 0.8 a moderate effect size and greater than 0.8 a large effect size.²¹

Given the large sample size included in our analysis, tests of statistical significance were not performed, as it was anticipated that even very small and clinically irrelevant differences between groups would demonstrate statistical significance and may consequently have undue importance assigned to them.

RESULTS

Patients

Our study population included a total of 259 382 patients from five databases (MIMIC $n=38\,508$, eICU-CRD $n=132\,994$, PLAGH-ICUdb $n=63\,515$, AUMCdb $n=20\,127$ and UCIHJ23db $n=4238$). Substantial heterogeneity existed across the databases in patient demographics, the interventions they received and their clinical outcomes, as displayed in [table 1](#). Notably, the proportion of patients who were admitted electively varied widely from 4.2% in UCIHJ23db to 71.7% in AUMCdb. By extension, the case mix also varied greatly, as reflected by the proportion of patients admitted following cardiac surgery (0.0% in UCIHJ23db vs 35.0% in AUMCdb). The interventions that patients received differed across databases, most appreciably in the delivery of mechanical ventilation and the administration of intravenous crystalloids, colloids and PRBCs.

The proportion of patients who died in the ICU ranged from 5.73% in eICU-CRD to 14.61% in UCIHJ23db. Similarly, the median ICU LOS ranged from 25 (20–73) hours in AUMCdb to 95 (46–173) hours in PLAGH-ICUdb.

Laboratory results

The IQR and median values of the most severely deranged measured laboratory investigations, for creatinine (maximum), haemoglobin (minimum), lactate (maximum) and sodium (maximum), stratified by database and patient outcomes, are displayed in [online supplemental table 2](#). The locally used normal reference ranges for each database are also reported.

Regarding the distribution of investigation results and their corresponding reference ranges, the sodium measurements of best outcome patients consistently fell within the corresponding normal range ([online supplemental figure 1](#)), though other laboratory results did so variably. While the upper margin of the creatinine reference range includes the vast majority of best outcome patients from the PLAGH-ICUdb and UCIHJ23db, increasing proportions of best outcome patients had creatinine values beyond the upper margin in AUMCdb, MIMIC and eICU-CRD ([online supplemental figure 2](#)). The distribution of haemoglobin results shows that the majority of patients tended to record values below the lower margin of their local reference range across all databases, irrespective of whether they had best or worst outcomes ([figure 1](#)). Similarly, the distribution of lactate measurements indicates that a substantial proportion of patients with best outcomes had a measured lactate above the upper margin of local reference ranges, particularly in MIMIC and eICU-CRD ([figure 2](#)).

The 95% CIs for each laboratory value, stratified by best and worst clinical outcome group and by database, are reported in [table 2](#). Overlapping and divergence coefficients are reported in [table 3](#) and summarise the degree to which the distribution of laboratory results differed between best and worst outcome patients.

The best and worst outcome patients in the UCIHJ23db demonstrated the greatest overlap in the distribution of both creatinine (OVL=0.67, SMD=-0.46) and haemoglobin (OVL=0.86, SMD=0.32), while those from the PLAGH-ICUdb demonstrated the least overlap in the distribution of these laboratory results (creatinine OVL=0.48, SMD=-0.92 and haemoglobin OVL=0.67, SMD=0.8) ([online supplemental figure 2](#) and [figure 1](#)).

Best and worst outcome patients from MIMIC demonstrated the greatest overlap in the distribution of highest measured lactate (OVL=0.65, SMD=-0.65), while those from AUMCdb demonstrated the least overlap (OVL=0.47, SMD=-1.01) ([figure 2](#)). AUMCdb also demonstrated the least overlap in highest measured sodium between best and worst outcome patients (OVL=0.67, SMD=-0.74), while the remaining databases consistently demonstrated OVL of approximately 0.75 for sodium measurements ([online supplemental figure 1](#)).

Overall, the mean overlap between best and worst patients across databases was greatest for measurements of haemoglobin (OVL=0.79) and lowest for measurements of lactate (OVL=0.45).

DISCUSSION

Differences between the most severely deranged laboratory results of patients admitted to the ICU and locally used normal reference ranges were observed in every database studied. These differences persisted even when comparing those patients with the best outcomes against the normal reference range.

In addition, among the databases, differences in the degree of overlap between best and worst group laboratory distributions were observed, which may represent variability in case mix and therapies applied, and/or imply variable discriminatory function among laboratory values based on region. Our findings build on the single-centre work by Tyler *et al*¹⁴ by replicating similar observations across different contexts and geographies. They further support the need to consider context in reacting to abnormal laboratory values, as correcting abnormal values may not always be beneficial or benign.^{5 11}

The differences observed between the reference range and selected ICU values across all five databases suggest that normal reference ranges are not useful in managing critically unwell patients. For instance, the haemoglobin results of most ICU patients fell outside normal reference ranges, irrespective of whether they had the best or worst outcome ([figure 1](#)). Patients with the best outcomes (here, ICU survival and shortest LOS) would be expected to have laboratory results that more closely align with the reference range, while those who die in the ICU should have results which are significantly worse. This is based on the assumption that the further patients' results deviate from the reference range, the more severely deranged their physiology and the more likely they are to have a poor clinical outcome. However, given the difference observed between the reference range and best outcome

Table 1 Baseline characteristics, interventions and outcomes of study population

	MIMIC		eICU-CRD	PLAGH-ICUdb	AUMCdb	UCIHJ23db
	USA	USA	USA	China	Netherlands	Spain
Country						
ICU admissions, No	38 508		132 994	63 515	20 127	4238
Male sex, No (%)	21 793 (56.59)		71 983 (54.15)	38 155 (60.07)	12 807 (65.16)	2730 (64.42)
Age, years, median (IQR)	65.70 (52.40–77.90)		65.00 (53.00–76.00)	56.89 (45.89–66.77)	Not available*	63.00 (50.00–73.00)
Admission type, No (%):						
Elective	6039 (15.83)		24 631 (18.52)	29 398 (46.29)	14 429 (71.69)	179 (4.22)
Emergency:	32 122 (84.17)		108 363 (81.48)	34 117 (53.71)	5698 (28.31)	4059 (95.78)
Via emergency department	19 152 (49.77)		69 077 (51.94)	29 790 (46.90)	2515 (12.50)	2220 (52.38)
Via other (eg, ward)	13 248 (34.40)		39 312 (29.56)	4327 (6.81)	3183 (15.81)	1839 (43.39)
Cardiac surgery, No (%)	7606 (19.75)		12 103 (9.10)	7265 (11.44)	7036 (34.96)	0 (0)
Mechanical ventilation,						
No (%)	19 156 (49.78)		49 159 (36.96)	Not available†	14 055 (69.83)	1844 (43.51)
Duration—hours, median (IQR)	19.00 (6.53–82.25)		27.82 (11.28–74.35)	Not available†	Not available	94.65 (28.40–256.82)
RRT,						
No (%)	1790 (4.65)		2154 (4.38)‡	Not available†	918 (4.56)	343 (8.09)
Crystalloid administration,						
No (%)	27 218 (70.74)		Not available§	52 165 (82.13)	19 079 (94.79)	1894 (44.69)
mL, median (IQR)	2000 (900–3650)		Not available§	1500 (750–2500)	1500 (1000–2469)	1000 (500–1500)
Colloid administration, No (%)	2587 (6.72)		Not available§	19 527 (30.74)	8642 (42.94)	217 (5.12)
Packed red blood cell transfusion, No (%)	12 361 (32.12)		Not available§	8584 (13.51)	5435 (27.00)	1007 (23.76)
ICU length of stay, hours, median (IQR)	50.30 (28.50–98.20)		39.00 (21.00–71.00)	95.01 (45.67–173.16)	25.00 (20.00–73.00)	93.70 (48.20–191.52)
ICU mortality, No (%)	2947 (7.66)		7618 (5.73)	3743 (5.89)	1929 (9.58)	619 (14.61)

*AUMCdb only contains age categories.

†Data not recorded in PLAGH-ICUdb.

‡Data from subset of eICU-CRD sites with reliable data regarding RRT therapy.

§Excluded due to poor data collection regarding IV fluid and PRBC administration.

ICU, intensive care unit; PRBC, packed red blood cell; RRT, renal replacement therapy.

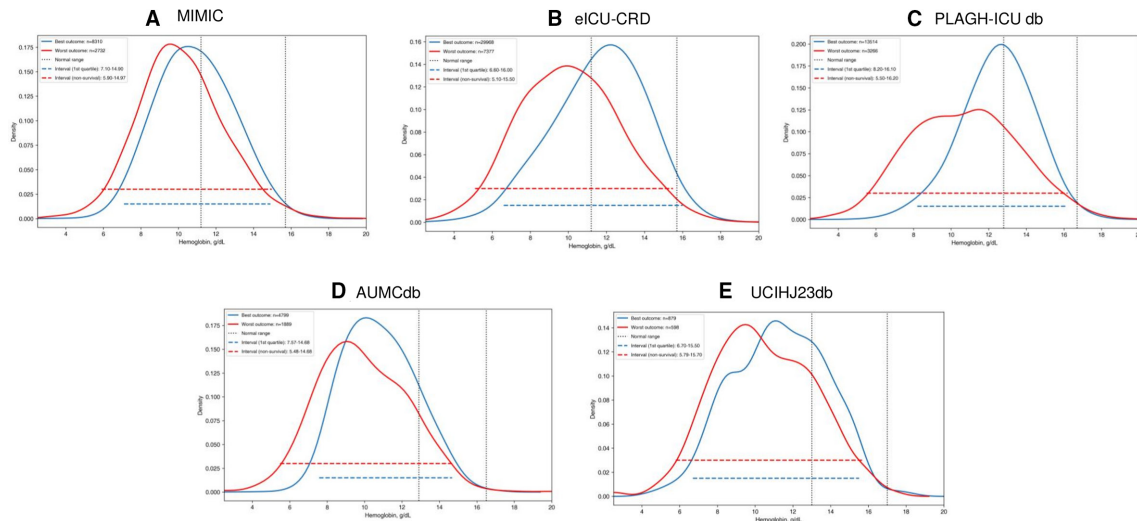


Figure 1 Minimum haemoglobin measurement on first intensive care unit admission—best versus worst outcome per database (A-E).

group, it is clear that normal reference ranges are not meaningful in ICU contexts. This discrepancy likely represents how reference ranges are formulated: reference ranges, though used to define normal and abnormal for both healthy individuals and critically unwell patients, are typically derived from samples of healthy outpatients.¹

As expected, patients with the best and worst clinical outcomes had differing laboratory results across databases. However, between databases, we found that the extent to which these groups differed was variable, as the degree of overlap in distributions changed across investigations and the context in which they were utilised. For

example, in the UCIHJ23db from Spain, the creatinine of patients with the best and worst clinical outcomes had substantial overlap (OVL=0.67), suggesting a decreased ability for creatinine to differentiate between patients with good and bad outcomes in this context. By comparison, creatinine results in the PLAGH-ICUdb from China demonstrated a lower overlap between groups (OVL=0.48). Consequently, creatinine may serve as a better prognosticator in this database, as it better discriminates between those with good and bad outcomes.

Variation in the overlap of laboratory results between patients with the best and worst outcomes was also seen

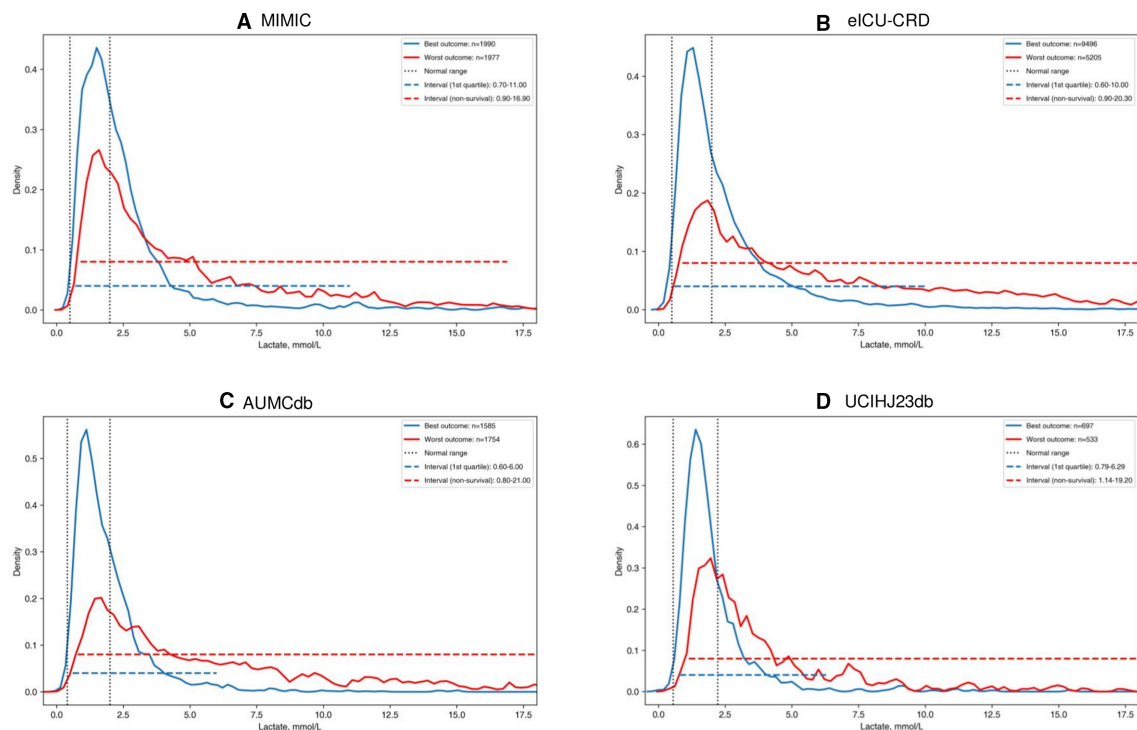


Figure 2 Maximum lactate measurement on first intensive care unit admission—best versus worst outcome per database (A-D). Data not recorded in PLAGH-ICUdb.

Table 2 Ninety-five per cent CIs of laboratory results stratified by best and worst outcome patient groups

	Creatinine, mg/dL		Haemoglobin, g/dL		Lactate, mmol/L		Sodium, mEq/L	
	Best	Worst	Best	Worst	Best	Worst	Best	Worst
MIMIC	0.50–4.50	0.50–6.40	7.10–14.90	5.90–14.97	0.70–11.00	0.90–16.90	131.00–146.00	129.00–155.78
eICU-CRD	0.50–6.45	0.54–7.50	6.60–16.00	5.10–15.50	0.60–10.00	0.90–20.30	130.00–147.00	128.00–157.00
PLAGH-ICUdb	0.44–1.51	0.42–7.49	8.20–16.10	5.50–16.20	Not available*	Not available*	133.10–147.90	128.80–156.60
AUMCdb	0.45–5.12	0.54–5.72	7.57–14.68	5.48–14.68	0.60–6.00	0.80–21.15	132.00–147.00	132.00–158.00
UCIHJ23db	0.40–5.04	0.44–6.33	6.70–15.50	5.79–15.70	0.79–6.29	1.14–19.20	132.00–148.93	129.00–155.28

*Data not recorded in PLAGH-ICUdb.

to different extents in the measurement of haemoglobin, lactate and sodium. These results imply that different investigations may represent good prognosticators in one context but not another and question the value of attempting to return these results to a healthy patient's reference range. As with currently utilised reference ranges, context-specific reference ranges developed from a heterogeneous cohort of patients would need to be interpreted with an understanding of individual patient factors and how their acute pathology may alter the significance of specific results.

While outside the scope of our project, we included data regarding patient demographics and common critical care interventions (renal replacement therapy (RRT), PRBC transfusion and crystalloid vs colloid resuscitation), which may have indicated mechanisms contributing to these variations in overlap. For instance, PLAGH-ICUdb was seen to have a narrower distribution of creatinine results. While RRT data was not available from this database, it can be seen that intravenous fluid administration was similar to that in other databases, so is unlikely to explain variations in renal function. However, the overall younger age of patients in PLAGH-ICUdb may have contributed to their lower creatinine. Furthermore, PLAGH-ICUdb also displayed the lowest overlap in haemoglobin results between best and worst outcome patients. While they also had the lowest PRBC administration rate, whether this represents a causative relationship is unknown. More broadly, we have demonstrated that substantial variability in the case mix and therapies provided existed across the databases, which may have contributed to differences in results across countries. Considering the differences between centres worldwide, the concept of context-specific reference ranges may prove even more useful by guiding practice with the goal of improving patient outcomes rather than unnecessarily normalising pathology results. Future research must consider the impact of case mix and clinical practices when developing new reference ranges, which would then require prospective validation to confirm them as appropriate treatment targets.

As such, our current study forms the foundation for several avenues of future enquiry. First, we intend to analyse and compare more homogeneous subgroups of patients (eg, cardiac surgery) and define context-specific laboratory result ranges, which are associated with the best clinical outcomes, and may therefore represent 'normality' for these groups of critically ill patients. Such reference ranges may then be prospectively validated to determine if they represent appropriate treatment targets and whether deviation from these ranges are associated with poorer outcomes. Furthermore, prospective studies will allow for the collection of data regarding the therapies provided to patients and thereby an investigation of the mechanism through which context-specific variations may arise. In addition, among databases that have collected data over a greater length of time (eg, PLAGH-ICUdb from 2008 until 2019), we intend to investigate

Table 3 Overlap between laboratory distributions of best and worst outcome patients with the local reference range

	Creatinine		Haemoglobin		Lactate		Sodium	
	OVL	SMD	OVL	SMD	OVL	SMD	OVL	SMD
MIMIC	0.65	-0.53	0.85	0.35	0.65	-0.65	0.76	-0.31
eICU-CRD	0.62	-0.48	0.74	0.61	0.57	-1.03	0.77	-0.35
PLAGH-ICUdb	0.48	-0.92	0.67	0.8	Not available*	Not available*	0.75	-0.09
AUMCdb	0.61	-0.48	0.81	0.42	0.47	-1.01	0.67	-0.74
UCIHJ23db	0.67	-0.46	0.86	0.32	0.57	-0.73	0.75	-0.31
Mean	0.61	-0.57	0.79	0.50	0.45	-0.68	0.74	-0.36

*Data not recorded in PLAGH-ICUdb.

OVL, overlapping coefficient; SMD, standardised mean difference.

whether the association between laboratory results and outcomes varies over time and therefore suggests that the prognostic value of results and their corresponding reference ranges require periodic review.

Strengths and limitations

Our study has several strengths. It is an analysis of an extremely large dataset, including more than 250 000 patients from three continents. Moreover, the ICUs included are varied in their case mixes and the corresponding severity of illness of their patients.

However, several limitations exist within this study. As with all retrospective research involving multiple large databases, variation in the design, collection and coding of variables may vary across datasets, creating inaccuracy in results. In our study, this is mitigated through the use of objective variables including laboratory results, ICU LOS and ICU mortality. Retrospective research of this nature is also inherently limited by missing data. Notably in our study this included missing data regarding lactate and interventions from PLAGH-ICUdb and intravenous fluid therapy and transfusions in eICU-CRD, respectively. However, other than lactate, these variables were used purely for hypothesis-generating purposes and do not alter our primary findings. The included databases collected information from ICU admissions across varying years, so differences in results may reflect changes in global practices over time rather than differences between centres or countries. Dichotomising patients into those with best and worse outcomes using ICU LOS and mortality does not reflect patient outcomes beyond ICU discharge. This includes the possibility that patients classified as having the 'best' outcome may have been discharged quickly from the ICU to receive end-of-life care. However, these definitions improved interpretability of our results and are consistent with those used previously.¹⁴ Further, our study includes descriptive analyses without adjustment for potential confounders. Therefore, the associations between individual laboratory results and patient outcomes do not indicate independent causative relationships and should not be interpreted as such. Finally, comparing heterogeneous patient populations comprising varied case mixes is problematic. The

possibility that context-specific reference ranges would also need to vary based on patient factors or specific conditions exists, though could not be concisely investigated in our present work.

CONCLUSION

In a cohort of more than 250 000 patients admitted to ICUs across four countries and three continents, there was substantial deviation in laboratory results when compared with normal reference ranges, even for those with the best clinical outcomes. Furthermore, when stratified by patients with the best and worst clinical outcomes, the degree of overlap between these patient groups varied widely across investigations and databases. These results suggest not only that specific reference ranges may be required for critically ill patients in different contexts but also that investigations may have a varying ability to discriminate between patients' outcomes depending on the setting.

Author affiliations

¹School of Medicine, Chinese PLA General Hospital, Beijing, China

²Laboratory for Computational Physiology, Harvard-MIT Division of Health Sciences and Technology, Cambridge, Massachusetts, USA

³Institute of Medical Informatics, Charité - Universitätsmedizin Berlin (corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health), Berlin, Germany

⁴School of Public Health and Preventive Medicine, Monash University, Melbourne, Victoria, Australia

⁵Department of Internal Medicine, Beth Israel Deaconess Medical Center, Boston, Massachusetts, USA

⁶Global Health Drug Discovery Institute, Beijing, China

⁷Department of Computer Science and Technology, Tsinghua University, Beijing, China

⁸Department of Intensive Care Medicine, Joan XXIII University Hospital in Tarragona, Tarragona, Catalunya, Spain

⁹Pere Virgili Health Research Institute, Reus, Catalunya, Spain

¹⁰School of Biological Science and Medical Engineering, Beihang University, Beijing, China

¹¹Information Department, Chinese PLA General Hospital, Beijing, China

¹²Department of Cardiology, National University Health System, Singapore

¹³Department of Intensive Care Medicine, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

¹⁴Medical Innovation Research Department, Chinese PLA General Hospital, Beijing, China

¹⁵Division of Pulmonary Critical Care and Sleep Medicine, Beth Israel Deaconess Medical Center, Boston, Massachusetts, USA

Twitter Leo Anthony Celi @MITCriticalData

Acknowledgements Article funding was supplied by MIT Libraries, the Beijing Municipal Science and Technology Project (Z181100001918023) and the Big Data R&D Project of Chinese PLA general hospital (2018MBD-009).

Funding LAC is funded by the National Institute of Health through NIBIB R01 EB017205.

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval The study data and protocol were approved, where applicable, by the respective institutional review board (IRB) as follows: The data in MIMIC-III has been previously de-identified, and the IRBs of the Massachusetts Institute of Technology (0403000206) and Beth Israel Deaconess Medical Center (2001-P-001699/14) approved the use of the database for research. The use of eICU-CRD is exempt from IRB approval due to the retrospective design, lack of direct patient intervention, and the security schema, for which the re-identification risk was certified as meeting safe harbor standards by an independent privacy expert (Privacert, Cambridge, MA Massachusetts, USA). The use of AmsterdamUMCdb is likewise exempt from IRB approval due to a combination of de-identification, contractual and governance strategies where re-identification is not reasonably likely and can therefore be considered anonymous information in the context of the General Data Protection Regulation. Data from UCIHJ23db has been previously anonymized, removing any link with real identifiers. The IRB of the Hospital Universitari de Tarragona Joan XXIII approved the anonymization mechanism used for the present study. Finally, the data from People's Liberation Army (PLA) General Hospital was de-identified and approved for research use by the hospital ethics committee (S2021-050-01). Due to the de-identified nature of the data, the analysis is not considered human subject research. Hence, informed consent was waived for this study as approved by Beth Israel Deaconess Medical Center, Boston, MA Massachusetts, USA, and all other partaking hospitals and universities.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available in a public, open access repository. Data are available upon 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.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>.

ORCID iDs

Louis Agha-Mir-Salim <http://orcid.org/0000-0002-2733-5084>

Leo Anthony Celi <http://orcid.org/0000-0001-6712-6626>

REFERENCES

- Horowitz G, Jones G. Establishment and use of reference intervals. In: Burtis C, Ashwood E, Bruns E, eds. *Tietz textbook of clinical chemistry and molecular diagnostics*, 2017: 170–94.
- Takala J, Ruokonen E, Webster NR, *et al*. Increased mortality associated with growth hormone treatment in critically ill adults. *N Engl J Med* 1999;341:785–92.
- NICE-SUGAR Study Investigators for the Australian and New Zealand Intensive Care Society Clinical Trials Group and the Canadian Critical Care Trials Group, Finfer S, Chittock D, *et al*. Intensive versus conventional glucose control in critically ill patients with traumatic brain injury: long-term follow-up of a subgroup of patients from the NICE-SUGAR study. *Intensive Care Med* 2015;41:1037–47.
- Holst LB, Haase N, Wetterslev J, *et al*. Lower versus higher hemoglobin threshold for transfusion in septic shock. *N Engl J Med* 2014;371:1381–91.
- Aberegg SK, O'Brien JM. The normalization heuristic: an untested hypothesis that may misguide medical decisions. *Med Hypotheses* 2009;72:745–8.
- Carson JL, Trulzi DJ, Ness PM. Indications for and adverse effects of red-cell transfusion. *N Engl J Med* 2017;377:1261–72.
- Carson JL, Stanworth SJ, Roubinian N, *et al*. Transfusion thresholds and other strategies for guiding allogeneic red blood cell transfusion. *Cochrane Database Syst Rev* 2016;10:CD002042.
- Carson JL, Guyatt G, Heddle NM, *et al*. Clinical practice guidelines from the AABB: red blood cell transfusion thresholds and storage. *JAMA* 2016;316:2025–35.
- Soril LJJ, Noseworthy TW, Stelfox HT, *et al*. A retrospective observational analysis of red blood cell transfusion practices in stable, non-bleeding adult patients admitted to nine medical-surgical intensive care units. *J Intensive Care* 2019;7:19.
- Sadana D, Kummangal B, Moghekar A, *et al*. Adherence to blood product transfusion guidelines—An observational study of the current transfusion practice in a medical intensive care unit. *Transfus Med* 2021;31:227–35.
- Aberegg S. The normalization fallacy: why much of “critical care” may be neither. *Pulm CCM*, 2017. Available: <http://pulmccm.org/main/2017/critical-care-review/normalization-fallacy-much-critical-care-may-neither/> [Accessed 7 Jan 2021].
- Manrai AK, Patel CJ, Ioannidis JPA. In the era of precision medicine and big data, who is normal? *JAMA* 2018;319:1981–2.
- Gräsbeck R. The evolution of the reference value concept. *Clin Chem Lab Med* 2004;42:692–7.
- Tyler PD, Du H, Feng M, *et al*. Assessment of intensive care unit laboratory values that differ from reference ranges and association with patient mortality and length of stay. *JAMA Netw Open* 2018;1:e184521.
- Ezzie ME, Aberegg SK, O'Brien JM. Laboratory testing in the intensive care unit. *Crit Care Clin* 2007;23:435–65.
- Johnson AEW, Pollard TJ, Shen L, *et al*. MIMIC-III, a freely accessible critical care database. *Sci Data* 2016;3:1–9.
- Pollard TJ, Johnson AEW, Raffa JD, *et al*. The eICU Collaborative research database, a freely available multi-center database for critical care research. *Sci Data* 2018;5:1–13.
- Thoral PJ, Peppink JM, Driessen RH. Sharing ICU patient data Responsibly under the SCCM/ESICM joint data science collaboration: the AmsterdamUMCdb example. *Crit Care Med* 2021;49.
- Liu T, Liu X, Fan Y. Constructing a Comprehensive Clinical Database Integrating Patients' Data from Intensive Care Units and General Wards. *Proc - 2019 12th Int Congr Image Signal Process Biomed Eng Informatics, CISP-BMEI 2019*, 2019.
- Xu H. Online appendix—ICU reference ranges code. GitHub, 2021. Available: <https://github.com/xhr0506/ICU-labtest>
- Inman HF, Bradley EL. The overlapping coefficient as a measure of agreement between probability distributions and point estimation of the overlap of two normal densities. *Commun Stat Theory Methods* 1989;18:3851–74.
- Cohen J. *Statistical power analysis for the behavioral sciences*. L. Erlbaum associates, 1988.

© 2021 Author(s) (or their employer(s)) 2021. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ. This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>. Notwithstanding the ProQuest Terms and Conditions, you may use this content in accordance with the terms of the License.

Reliability of COVID-19 symptom checkers as national triage tools: an international case comparison study

Fatma Mansab,^{1,2} Sohail Bhatti,^{1,2} Daniel Goyal ^{1,3}

To cite: Mansab F, Bhatti S, Goyal D. Reliability of COVID-19 symptom checkers as national triage tools: an international case comparison study. *BMJ Health Care Inform* 2021;**28**:e100448. doi:10.1136/bmjhci-2021-100448

► Additional supplemental material is published online only. To view, please visit the journal online (<http://dx.doi.org/10.1136/bmjhci-2021-100448>).

Received 19 July 2021
Accepted 30 September 2021

ABSTRACT

Objectives Triage is a critical component of the pandemic response. It affects morbidity, mortality and how effectively the available healthcare resources are used. In a number of nations the pandemic has sponsored the adoption of novel, online, patient-led triage systems—often referred to as COVID-19 symptom checkers. The current safety and reliability of these new automated triage systems remain unknown.

Methods We tested six symptom checkers currently in use as triage tools at a national level against 52 cases simulating COVID-19 of various severities to determine if the symptom checkers appropriately triage time-critical cases onward to healthcare contact. We further analysed and compared each symptom checker to determine the discretionary aspects of triage decision-making that govern the automated advice generated.

Results Of the 52 clinical presentations, the absolute rate of onward referral to any form of healthcare contact was: Singapore 100%, the USA 67%, Wales 65%, England 62%, Scotland 54% and Northern Ireland 46%. Triage decisions were broadly based on either estimates of 'risk' or 'disease severity'. Risk-based symptom checkers were more reliable, with severity-based symptom checkers often triaging time-critical cases to stay home without clinical contact or follow-up.

Conclusion The COVID-19 symptom checkers analysed here were unable to reliably discriminate between mild and severe COVID-19. Risk-based symptom checkers may hold some promise of contributing to pandemic case management, while severity-based symptom checkers—the CDC and NHS 111 versions—confer too much risk to both public and healthcare services to be deemed a viable option for COVID-19 triage.

INTRODUCTION

Symptom checkers are online platforms where the public can enter details of their illness, answer set questions about their symptoms and then receive advice on what to do next. During the pandemic, many nations have deployed symptom checkers to help identify potential COVID-19 cases and provide advice to the public. Some nations have gone further, using symptom checkers in place of more typical clinical triage systems.¹

Despite a number of studies highlighting the diagnostic sensitivity of various online

Summary

What is already known?

- Symptom checkers have been deployed at a national level in a number of countries to support the pandemic response.
- There are no quality, safety or efficacy studies supporting the use of COVID-19 symptom checkers as triage tools.

What does this paper add?

- The COVID-19 symptom checkers analysed here are currently in use at a national level as stand-alone triage services. They are all freely accessible to the public.
- Out of the symptom checkers analysed, only the UK version (NHS 111 COVID-19 Symptom Checker) has been formally integrated into the national clinical pathway.
- None of the symptom checkers analysed here could reliably distinguish between mild and severe COVID-19.

COVID-19 symptom checkers, we could find no studies (apart from our previous analysis)² examining the safety and reliability of online COVID-19 symptom checkers as a standalone triage tool.

The difference is stark. On the one hand, these accessible web-based questionnaires can direct potential cases toward SARS-CoV-2 testing services—answering the question: should you be tested? In such circumstances, symptom checkers act more as a prompt, conveying the national advice. On the other hand, there is another category of symptom checkers attempting to answer a much more complicated question: do you need medical help?

It is quite an ask, of an automated system. And the stakes are high. There is the unavoidable direct morbidity and mortality impact when triaging acute medical problems.^{3,4} There is also an operational consideration, whereby delaying treatment in



© Author(s) (or their employer(s)) 2021. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

¹School of Postgraduate Medicine, University of Gibraltar, Gibraltar, Gibraltar

²Department of Public Health, Gibraltar Health Authority, Gibraltar, Gibraltar

³Department of Medicine, Lorn and Islands Hospital, Oban, UK

Correspondence to

Dr Daniel Goyal;
daniel.goyal@gha.gi

time-critical conditions leads to a higher overall healthcare burden.^{3–6}

Online COVID-19 symptom checkers may carry the potential to offset the inevitable high healthcare demands of pandemic management, allowing valuable resources to be focused on those with real clinical need. This benefit can only be realised if the symptom checker successfully triages COVID-19 pneumonia (and other serious conditions mimicking COVID-19) on to further care early enough for maximal treatment benefits to be achieved. Missing the opportunity to prevent disease progression will invariably lead to higher mortality,^{4,6} delayed recovery (eg, higher rates of long COVID-19)⁷ and a more lengthy inpatient stay.^{4,6} There is a very real possibility that symptom checkers, if inappropriately used, can significantly increase the healthcare burden associated with COVID-19—compromising healthcare capacity sooner than is necessary.^{2,5} Both at a patient level and operational level, quality and efficacy studies are essential if clinical activities are to be replaced or even augmented with such patient-led, automated clinical services.

Back in April 2020, we undertook an analysis on national symptom checkers from the UK, USA, Singapore and Japan. At that time, the results of our case-simulation study revealed a low rate of onward referral for both the UK's '111 COVID-19 Symptom Checker' and the US's 'CDC Coronavirus Symptom Checker'—44% and 38%, respectively. It was noted that both symptom checkers triaged simulated cases of severe COVID-19 pneumonia, bacterial pneumonia and sepsis, to stay home with no further healthcare contact.² In short, both the US and UK symptom checkers maintained a high threshold to refer patients onward for healthcare contact.

During this previous analysis, we compared and contrasted the four symptom checkers, looking specifically for points of divergence. The most notable difference was whether known COVID-19 disease 'risk factors' or an estimated 'disease severity' were used to calculate triage disposition. Both Singapore's and Japan's symptom checkers focused specifically on risk factors—age, duration of symptoms, the presence of breathlessness and comorbidities—and made no attempt to quantify disease severity. We have termed these, 'risk-based symptom checkers'. Both the US and UK symptom checkers, in April 2020, relied more on qualitative questions designed to estimate how severe a case was, and made no account of the most consistent risk factors—age, duration of symptoms or the presence of breathlessness. In addition, only moderately severe comorbidities affected the CDC symptom checker triage decision and only 'shielding category' comorbidities affected triage dispositions in the NHS 111 symptom checker.² We have termed these, 'severity-based symptom checkers'.

At the time of our initial study, the impact of the CDC symptom checker on actual case presentations to US hospitals was unknown. However, the 'NHS 111 COVID-19 Symptom Checker' was a known gatekeeper for UK patients with suspected or confirmed COVID-19.^{8–11} As

such, the symptom checker's 'decision' to triage cases simulating time critical urgent medical problems—severe COVID-19, bacterial pneumonia and sepsis—to remain at home was considered by the authors as unsafe. No data existed on internal or external quality assurance studies, further compounding the concerns regarding the use of the NHS 111 COVID-19 symptom checker and patient safety. These concerns were raised with NHS Digital—the body responsible for the NHS 111 symptom checker—both prior to publication and following. NHS Digital considered the concerns to be historic and not representative of the improved version of the symptom checker.¹² We, therefore, undertook a repeat analysis in June 2021.

METHODS

During the first week in June 2021, we undertook a follow-up analysis on national symptom checkers. In summary, we generated four distinct patient scenarios relating to COVID-19. The four scenarios included were fever with cough, comorbidity with fever and cough, immunosuppression with fever and cough, and shortness of breath with fever. We varied patient age, duration of symptoms and the severity of symptoms. In total, this generated 52 separate case simulations, including mild, moderate, severe and critical COVID-19, and COVID mimickers such as bacterial pneumonia and sepsis. Each case was applied by a single investigator (DG) to each symptom checker and the triage decision was recorded. We then calculated the total referral ratio (ie, proportion of all 52 cases that were referred for clinical contact, regardless of the level of designation—call centre, primary care provider, or emergency department). A percentage ratio was generated of total referrals made by each symptom checker. We also noted specific features of each symptom checker for comparison.

In addition to the methods as described previously, we also completed analysis of symptom checkers from Scotland,¹³ Wales¹⁴ and Northern Ireland,¹⁵ and thus differentiated the previously analysed NHS England symptom checker from the other three nations.

A more detailed methodology is explained by Mansab *et al.*²

RESULTS

Of the four nations included in the initial analysis of April 2020, Singapore, the USA and all four nations within the UK continue to use symptom checkers as part of the national response to COVID-19. Japan was no longer using an accessible symptom checker, it being replaced by a flow chart. As such, it was excluded from further analysis.

Triage dispositions were slightly different for each symptom checker, but generally followed: stay home, or contact service provider/General Practitioner (GP)/111, or go straight to emergency department/999.

Table 1 National inpatient healthcare burden and key population statistics. Source: World Bank and WHO

	Singapore	USA	UK
Population data			
Patients currently admitted to hospital per 10 000 inhabitants (rate as per April 2020)	0.38 (2.23)	0.21 (0.43)	0.3 (2.29)
Mean national age (years)	44.2	38.4	40.5
Gross Domestic Product per capita (thousands of US dollars)	59.8	63.5	40.3
Physicians per 10 000 head of capita	24	25	28
Total case fatality rate (%)	0.05	1.8	2.8

The current analysis was undertaken during a relatively low prevalence time, when COVID-19 inpatient burden was relatively low (table 1).^{16–18}

Referral ratio

The rates of onward referral to any further healthcare contact for each national symptom checker were: Singapore 100%, the USA 67%, Wales 65%, England 62%, Scotland 54% and Northern Ireland 46% (figure 1). Previous referral rates in April 2020 were: Singapore 88%, the USA 38% and England 44%. For the triage disposition of individual case simulations, see online supplemental data.

Specific features

The Singapore symptom checker currently refers all cases for a same day assessment at one of the nation's public health clinics. It continues to refer all patients with any degree of breathing problems directly to the emergency department (scenario 4, online supplemental material).

The US (CDC) symptom checker referred twice as many cases on to clinical care than it did the year before. Notably, the advice for those referred had changed from 'contact medical provider within 24 hours' to 'contact medical provider as soon as possible'. Age was also now a considered risk factor for disease severity with all patients over the age of 65 years with suspected COVID-19 being advised to contact their medical provider regardless of disease severity or other comorbidity. The CDC symptom

checker continued to triage those under 65 years of age with mild to moderate shortness of breath to stay home with no further clinical contact (scenario 4, online supplemental material).

The UK symptom checkers do not account for age in the triage decision in the case simulations undertaken, except for NHS Wales. NHS Wales '111' symptom checker triaged all cases over the age of 70 years onward to call '111'. Scotland, Northern Ireland and England continued to, for example, triage a 72 years old with cough and fever for 7 days to stay home with no healthcare contact or follow-up (scenario 1, online supplemental material).

In comparison to the year previously, the NHS England symptom checker now triaged any case with the subjective sense of shortness of breath onward to further healthcare contact ('call 111' for cases with self-rated mild to moderate shortness of breath, and the emergency department for severe shortness of breath). If, though, shortness of breath is a secondary symptom (ie, feeling flu-like with shortness of breath), then patients are still advised to stay home, unless self-identified as severe (table 2).

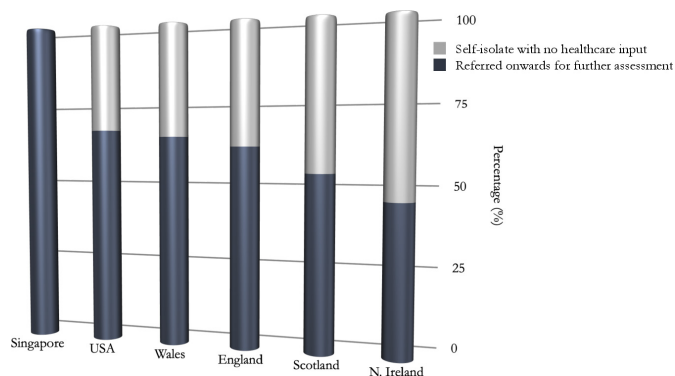


Figure 1 Percentage ratio of absolute onward referrals of each national symptom checker. Black represents the total percentage of cases triaged onward to further healthcare contact. Grey represents the percentage of cases triaged to remain at home with no further planned healthcare follow-up.

DISCUSSION

The WHO guidelines on triage recommend that all patients with suspected or confirmed COVID-19 are clinically triaged.¹⁹ In many countries, this proves challenging due to national policy, healthcare accessibility, healthcare resources and the level of community transmission of SARS-CoV-2. Most patients with COVID-19 do not develop complicated illness and will resolve the infection without clinical intervention. Successful clinical pathways are able to pick out the COVID-19 pneumonia reliably and, crucially, early enough for basic medical care to prevent progression of disease.

There have been noticeable improvements in the symptom checkers that were first analysed in April 2020. Notably, the overall referral rate onward to healthcare contact has increased in the USA, UK and Singapore. Other features of the symptom checkers have also been improved. Singapore has included a question on immunosuppression, although it still does not confer such patients to more urgent care. The CDC coronavirus symptom checker has now included age as a defining risk factor. Most symptom checkers have expanded

Table 2 Summary of the national COVID-19 symptom checkers' triage criteria

Triage criteria	Risk-based symptom checker	Severity-based symptom checker	
	Singapore	CDC	NHS 111
Duration of symptoms	Duration of symptoms affects triage outcomes. Patients with symptoms over 4 days are always triaged in to further care.	Duration of symptoms does not alter triage outcomes. No length of illness leads to triage in to further healthcare contact.	Duration of symptoms does not alter triage outcomes. No length of illness leads to triage in to further healthcare contact.
Age	Age affects triage outcomes. Cases over the age of 65 years are always triaged on for further healthcare contact. During times of low SARS-CoV-2 healthcare burden, age is removed as a restriction to further clinical assessment.	Age affects triage outcomes. Cases over the age of 65 years are now triaged on to further healthcare contact.	Age has no bearing on triage advice for the NHS symptom checkers in England, Scotland and Northern Ireland. NHS Wales triages all cases over the age of 70 years to contact '111'.
Comorbidity	Comorbidity affects triage outcomes. Cases with any comorbidity are triaged on to further care.	Comorbidity affects triage outcomes. Cases with moderately severe comorbidities are triaged on for further healthcare contact.	Comorbidity affects triage outcomes. The type of comorbidity triggering triage on to further healthcare contact differs across the four nations.*
Shortness of breath	Any degree of shortness of breath is triaged straight to the emergency department.	Patients with severe breathlessness are triaged to the emergency department. Patients with mild to moderate shortness of breath are advised to stay home with no clinical follow-up.	Patients with severe breathlessness are triaged to the emergency department. NHS England triage cases in to further care if self-rated breathlessness is mild to moderate and is the primary symptom, but not if a secondary symptom. Scotland, Northern Ireland and Wales triage all patients self-reporting breathlessness to urgent '111'/General Practitioner. If severe, the patient is advised to call '999'.

*NHS Scotland only triages cases with comorbidities that are on the shielding category list for further care. NHS Northern Ireland only triages cases with immunosuppression or conditions that have become more difficult to control since symptoms began. NHS Wales relies on shielding categories and also includes diabetes and pregnancy with heart conditions. NHS England use shielding categories and also consider immunosuppression, diabetes, heart disease, respiratory disease, kidney failure, liver disease or neurological disease.

the list of comorbidities affecting triage decisions. And while the seriousness of breathlessness is given variable attention by each national symptom checker, all seem to now give it more weight in triage decisions. While these changes move towards re-establishing the usual standards of care for patients with COVID-19, our results indicate that COVID-19 symptom checkers are too unreliable in discriminating mild from severe COVID-19, and, as such, are likely to confer too great a risk to public, and, if used in place of clinical triage, will only weaken the healthcare response to the pandemic.²⁰

Specifics

Whether the CDC coronavirus symptom checker has in fact worsened outcomes for the USA or not remains unknown. It has not been formally integrated into the US healthcare system or COVID-19 clinical pathways. The fact it functions as a stand-alone patient-led triage service and is freely accessible to the public is concerning. While not formally or preferentially directed to the symptom checker, the US public can seek clinical advice from the automated service, and thereby, potentially—in the patient's mind—negate the need to undertake actual clinical triage. Given it cannot reliably differentiate between mild and severe COVID-19, any reassurance provided is neither dependable nor evidence based. For example, the CDC coronavirus symptom checker still advises the 63years old with a 7-day history of persistent fever and

worsening cough to remain at home without contacting their healthcare service provider. In the absence of any discoverable quality and safety studies supporting its use, it would seem prudent to discontinue the use of the CDC coronavirus symptom checker until the CDC can prove its efficacy and safety, as would be the case for any diagnostic test.

Singapore remains consistent in its approach to the clinical management of COVID-19.²¹ All suspected cases of COVID-19 are clinically assessed by a physician and followed-up by primary care.²² This is consistent with the WHO technical guidelines and clinical recommendations for triage and management of COVID-19.¹⁹ During our previous analysis in April 2020, Singapore was suffering a surge of COVID-19 infections. The symptom checker was then set to advise the young, non-breathless patient with no comorbidities and a short duration of illness to self-isolate and contact the public health clinic if symptoms worsened or had not improved by day 4.² This remains a compromise to normal clinical care but may have achieved a low-risk reduction in healthcare burden. Now, with SARS-CoV-2 prevalence less than in April 2020, all suspected cases are clinically assessed.

Unlike the CDC and Singapore COVID-19 symptom checkers, the NHS '111' symptom checkers have a clear and critical role in the national clinical response to COVID-19.^{8–11} Current advice in the UK to the public

(including from contact tracers) is to self-isolate if COVID-19 is suspected or confirmed and if concerned about symptoms to use NHS 111 online services or call '111'.²³ Altogether this has generated—even during the low SARS-CoV-2 prevalence period of this analysis—around 30 000 online triages per month in England, including around 900 online triages in those 70 years old and over.²⁴ The NHS 111 COVID-19 symptom checkers act as a gatekeeper for further clinical contact. As such, the reliability of the UK's symptom checkers and the triage criteria set by NHS 111 is critical to COVID-19 patient outcomes and associated healthcare usage.

The NHS England symptom checker has increased the rate of onward referral since April 2020 and now refers patients with self-rated mild or moderate breathing difficulties to call '111'. From both a patient safety and healthcare burden perspective, this is a welcomed improvement. However, the NHS England symptom checker (and to some degree the NHS Wales symptom checker) continues to attempt to quantify disease severity with subjective questions and qualifiers likely to cause under-reporting of true disease severity, as it did previously.² It fails to reliably identify severe COVID-19 or other time-critical COVID mimickers.

Only the NHS Wales symptom checker accounted for age in the triage decision of the cases simulated. The reason for NHS Scotland, England and Northern Ireland not accounting for age—the most reliable predictor of disease severity—remains unclear.

None of the NHS symptom checkers account for 'silent hypoxia' (case scenario 1, online supplemental data). Silent hypoxia is the presence of hypoxia (low blood oxygen levels) without any sensation of breathlessness. It indicates severe or critical COVID-19 pneumonia and requires immediate inpatient care. Silent hypoxia affects up to one-third of patients presenting to hospital and carries a poorer prognosis.^{25 26} The inability of the NHS 111 symptom checkers to identify cases suffering silent hypoxia is likely a terminal limitation to the success of such severity-based symptom checkers as viable triage tools for COVID-19.

The attempt of the NHS 111 COVID-19 symptom checker to determine if COVID-19 is present, then to assign a severity level (ie, non-severe), constitutes a diagnostic process.²⁷ Given the NHS England symptom checker (and all UK national symptom checkers) then provide the clinical advice to 'self-isolate' and detailed advice on how to manage symptoms such as cough and breathlessness at home,²⁸ it breaches the boundary between simple signposting (simply deciding who is the most appropriate next healthcare contact) and ventures into the area of diagnosis and clinical management (deciding what treatment is appropriate based on an assessment). The NHS 111 COVID-19 symptom checker should then, at the very minimum, be subject to the same quality standards as any other diagnostic test, including national regulation.

As the UK clinical COVID-19 pathway is heavily reliant on such symptom checkers, together with the subsequent

diversion of patients away from actual clinical triage, the NHS 111 symptom checkers are likely to be contributing to the UK's poor pandemic response, including the high morbidity and mortality. Also of growing concern is the impact the NHS 111 symptom checkers are likely to have on the resilience of society to tolerate background levels of SARS-CoV-2 and post-pneumonia complications (eg, long COVID-19) by delaying presentation of COVID-19 pneumonia to timely, appropriate medical care.

Given the NHS '111' symptom checkers have ventured into diagnosis and, arguably, clinical management, are currently gatekeepers to further healthcare access, fail to reliably triage severe COVID-19 on to further care, fail to account for age as a risk factor (except NHS Wales) and are likely to miss COVID-19 mimickers such as bacterial pneumonia, considerable improvements are needed to render the current NHS 111 COVID-19 symptom checkers fit for purpose.

Future Direction

The use of symptom checkers as part of the national clinical care pathway for COVID-19 (and future pandemics) requires considerably more research and validation.²⁹ Currently, none of the symptom checkers pose a viable option in replacing clinical triage, and as such, effort should focus on resourcing clinical triage services.

Data relating to the use of symptom checkers and the effects these have on future healthcare burden have not yet been analysed. At an operational level, the possibility of severity-based symptom checkers leading to an increase in healthcare burden, including an increase in high-dependency admissions, should sponsor caution and an urgent review of any care pathways depending on such forms of patient-led triage. Our analysis suggests, severity-based COVID-19 symptom checkers (such as the NHS 111 or CDC versions) are likely to increase the healthcare burden associated with the pandemic (in comparison to clinically led, remote triage).

There is the equally challenging obstacle of national versus local triage to overcome. The current NHS 111 symptom checker triages nationally using the same referral thresholds. This may have contributed to the disproportionate healthcare activity across the UK.²⁷ Where a national symptom checker is 'set' to respond to critical demand in, for example, London, those using the symptom checker in an area of low demand, for example, the Lake District, will also be held to the same, compromised and rationed access to healthcare. This goes against the principles of triage, in that triage decisions must be responsive to resource availability. It is not justifiable—or logistically savvy—to ration access to healthcare preemptively or without a definitive need to.

In the short term—pending further safety studies—a 'risk-based symptom checker' may provide a possible low-risk solution to signposting potential COVID-19 cases, under pandemic conditions. The usual standard remains an actual clinical assessment, but where healthcare resources are insufficient for such a standard of

care, an untested symptom checker that can be adjusted in response to risk and demand would be preferable to an untested symptom checker attempting to determine clinical severity from an automated algorithm. A national symptom checker may still have a role in risk stratifying, and with the benefit of postcode localisation, there may be an ability to adjust the 'risk necessary to take' more accurately and based on local demands. Such a national service providing local risk stratification must be dynamic and responsive to demands, be under constant data collection and review, and be viewed as a considerable compromise to usual standards of care.

Whatever future version of COVID-19 symptom checkers manifest, they must be designed with the intention of detecting progressive COVID-19 or those at risk of severe disease, not designed with the intention of preventing healthcare contact. Triage itself is not resource saving. But the effort invested in the triage process yields high returns when cases of progressive COVID-19 pneumonia are detected early enough to avoid costly, protracted and complicated admissions. Triage systems must be viewed for what they are: an opportunity to maximise the use of available resources to prevent death, avoid disability and improve healthcare resilience.

CONCLUSION

The use of symptom checkers to triage patients during a pandemic or major incident is novel and untested. Our case simulation study provides little reassurance for their ongoing use. Even during a period of low healthcare burden, the symptom checkers deployed by both the USA and UK maintained a high threshold for onward referral. Neither symptom checker reliably triaged treatable, time-critical cases in to healthcare contact or follow-up and were unable to consistently differentiate mild from severe COVID-19. Of further concern, age is not factored in the triage decisions of the NHS 111 symptom checkers (except NHS Wales)—an unusual practice in clinical triage and well-below national and international standards of care.

Beyond the patient safety concerns, there is no evidence that COVID-19 symptom checkers reduce the healthcare burden associated with the pandemic. Our results suggest, by delaying the presentation of time-critical cases to medical care, it is quite likely the NHS 111 symptom checkers increase the healthcare burden associated with the SARS-CoV-2 pandemic in the UK.

In the absence of any safety, efficacy or quality assurance studies to support the use of symptom checkers as triage tools, our results necessitate a recommendation for the NHS 111 symptom checker and CDC coronavirus symptom checker to be subject to further analysis prior to their ongoing use in COVID-19 clinical care pathways. The stakes of patient triage are simply too high, and the reliability of symptom checkers is simply too poor, to justify their ongoing use.

Contributors All authors contributed to the conception, methodology and analysis and did final review and edit. FM undertook the majority of the write-up. DG undertook edits and contributed to the write-up and revision of the manuscript. DG is the guarantor of 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 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.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>.

ORCID iD

Daniel Goyal <http://orcid.org/0000-0003-0418-8859>

REFERENCES

- Morse KE, Ostberg NP, Jones VG, *et al*. Use characteristics and triage acuity of a digital symptom Checker in a large integrated health system: population-based descriptive study. *J Med Internet Res* 2020;22:e20549.
- Mansab F, Bhatti S, Goyal D. Performance of national COVID-19 'symptom checkers': a comparative case simulation study. *BMJ Health Care Inform* 2021;28:e100187.
- Lim WS, Baudouin SV, George RC, *et al*. Bts guidelines for the management of community acquired pneumonia in adults: update 2009. *Thorax* 2009;64 Suppl 3:iii1–55.
- Phua J, Dean NC, Guo Q, *et al*. Severe community-acquired pneumonia: timely management measures in the first 24 hours. *Crit Care* 2016;20:237.
- Haase CB, Bearman M, Brodersen J, *et al*. 'You should see a doctor', said the robot: Reflections on a digital diagnostic device in a pandemic age. *Scand J Public Health* 2021;49:33–6.
- , Horby P, Lim WS, *et al*, RECOVERY Collaborative Group. Dexamethasone in hospitalized patients with Covid-19. *N Engl J Med* 2021;384:693–704.
- Nalbandian A, Sehgal K, Gupta A, *et al*. Post-Acute COVID-19 syndrome. *Nat Med* 2021;27:601–15.
- NHS England. Novel coronavirus (COVID-19) standard operating procedure - COVID Oximetry @home. Available: <https://www.england.nhs.uk/coronavirus/wp-content/uploads/sites/52/2020/11/C0817-standard-operating-procedure-covid-oximetry-@home-v1.1-march-21.pdf> [Accessed 2 May 2021].
- NHS24, coronavirus (COVID-19): general advice. Available: <https://www.nhsinform.scot/illnesses-and-conditions/infections-and-poisoning/coronavirus-covid-19/coronavirus-covid-19-general-advice> [Accessed 27 May 2021].
- Department of Health. Northern Ireland. Covid-19 urgent and emergency care action: no more silos. Available: <https://www.health-ni.gov.uk/sites/default/files/publications/health/doh-no-more-silos.pdf> [Accessed 4 May 2021].
- Welsh Government cabinet statement. Timely presentation of COVID-19 disease. 4 Aug 2020. Available: <https://gov.wales/written-statement-timely-presentation-covid-19-disease> [Accessed 3 May 2021].
- BBC News. Covid online symptom checker 'may delay treatment'. Available: <https://www.bbc.co.uk/news/health-56323915>

- 13 NHS. NHS Inform Scotland, Self Help Guide: COVID-19, 2021. Available: <https://www.nhsinform.scot/self-help-guides/self-help-guide-coronavirus-covid-19>
- 14 NHS Wales. Coronavirus COVID-19 symptom Checker. Available: <https://111.wales.nhs.uk/SelfAssessments/symptomcheckers/COVID19.aspx>
- 15 Health and social care Northern Ireland, COVID-19 symptom Checker. Available: <https://covid-19.hscni.net/symptoms/>
- 16 NHS Digital. Healthcare in the UK | coronavirus in the UK (data.gov.uk). Available: <https://coronavirus.data.gov.uk/details/healthcare> [Accessed 14 Aug 2021].
- 17 WHO. Situation report: Singapore. Available: https://www.who.int/docs/default-source/wpro-documents/countries/singapore/singapore-situation-report/covid19_sitrep_sgp_20210606.pdf?sfvrsn=d536e556_7 [Accessed 6 Jun 2021].
- 18 John Hopkins University. COVID-19 situation Tracker. Available: <https://coronavirus.jhu.edu/us-map> [Accessed 14 Aug 2021].
- 19 World Health Organization. Operational considerations for case management of COVID-19 in health facility and community: interim guidance, 19 Mar 2020. World Health Organization, 2020. Available: <https://apps.who.int/iris/handle/10665/331492>
- 20 Christian MD. Triage. *Crit Care Clin* 2019;35:575–89.
- 21 Chotirmall SH, Wang L-F, Abisheganaden JA. Letter from Singapore: the clinical and research response to COVID-19. *Respirology* 2020;25:1101–2.
- 22 Lim WH, Wong WM. COVID-19: notes from the front line, Singapore's primary health care perspective. *Ann Fam Med* 2020;18:259–61.
- 23 England NHS. Novel coronavirus (COVID-19) standard operating procedure - COVID Oximetry @home. Available: <https://www.england.nhs.uk/coronavirus/wp-content/uploads/sites/52/2020/11/C0817-standard-operating-procedure-covid-oximetry-@home-v1.1-march-21.pdf> [Accessed 2 May 2021].
- 24 NHS24. Coronavirus (COVID-19): general advice. Available: <https://www.nhsinform.scot/illnesses-and-conditions/infections-and-poisoning/coronavirus-covid-19/coronavirus-covid-19-general-advice> [Accessed 27 May 2021].
- 25 Brouqui P, Amrane S, Million M, *et al.* Asymptomatic hypoxia in COVID-19 is associated with poor outcome. *Int J Infect Dis* 2021;102:233–8.
- 26 Committee on Diagnostic Error in Health Care, Board on Health Care Services, Institute of Medicine. The National Academies of Sciences, Engineering, and Medicine. In: Balogh EP, Miller BT, Ball JR, eds. *Improving diagnosis in health care*. Washington (DC): National Academies Press (US), 2015.
- 27 NHS England. How to look after yourself at home if you have coronavirus (COVID-19). Available: <https://www.nhs.uk/conditions/coronavirus-covid-19/self-isolation-and-treatment/how-to-treat-symptoms-at-home/> [Accessed 4 Jun 2021].
- 28 Burn S, Propper C, Stoye G. What happened to English NHS Hospital activity during the COVID-19 pandemic? Institute for Fiscal Studies. Available: <https://ifs.org.uk/publications/15432> [Accessed 13 May 2021].
- 29 Akbar S, Coiera E, Magrabi F. Safety concerns with consumer-facing mobile health applications and their consequences: a scoping review. *J Am Med Inform Assoc* 2020;27:330–40.

© 2021 Author(s) (or their employer(s)) 2021. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ. This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>. Notwithstanding the ProQuest Terms and Conditions, you may use this content in accordance with the terms of the License.

Dashboards for visual display of patient safety data: a systematic review

Daniel R Murphy,^{1,2} April Savoy,^{3,4,5} Tyler Satterly,^{1,2} Dean F Sittig ,^{6,7} Hardeep Singh^{1,2}

To cite: Murphy DR, Savoy A, Satterly T, *et al.* Dashboards for visual display of patient safety data: a systematic review. *BMJ Health Care Inform* 2021;**28**:e100437. doi:10.1136/bmjhci-2021-100437

Received 06 July 2021

Accepted 22 September 2021



© Author(s) (or their employer(s)) 2021. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

¹Center for Innovations in Quality, Effectiveness and Safety, Michael E DeBakey VA Medical Center, Houston, Texas, USA

²Department of Medicine, Baylor College of Medicine, Houston, Texas, USA

³Purdue School of Engineering and Technology, Indiana University Purdue University at Indianapolis, Indianapolis, Indiana, USA

⁴Center for Health Information and Communication, Richard L Roudebush VA Medical Center, Indianapolis, Indiana, USA

⁵Center for Health Services Research, Regenstrief Institute, Inc, Indianapolis, Indiana, USA

⁶School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, Texas, USA

⁷The UT-Memorial Hermann Center for Healthcare Quality & Safety, Houston, Texas, USA

Correspondence to

Dr Dean F Sittig;
dean.f.sittig@uth.tmc.edu

ABSTRACT

Background Methods to visualise patient safety data can support effective monitoring of safety events and discovery of trends. While quality dashboards are common, use and impact of dashboards to visualise patient safety event data remains poorly understood.

Objectives To understand development, use and direct or indirect impacts of patient safety dashboards.

Methods We conducted a systematic review in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses guidelines. We searched PubMed, EMBASE and CINAHL for publications between 1 January 1950 and 30 August 2018 involving use of dashboards to display data related to safety targets defined by the Agency for Healthcare Research and Quality's Patient Safety Net. Two reviewers independently reviewed search results for inclusion in analysis and resolved disagreements by consensus. We collected data on development, use and impact via standardised data collection forms and analysed data using descriptive statistics.

Results Literature search identified 4624 results which were narrowed to 33 publications after applying inclusion and exclusion criteria and consensus across reviewers. Publications included only time series and case study designs and were inpatient focused and emergency department focused. Information on direct impact of dashboards was limited, and only four studies included informatics or human factors principles in development or postimplementation evaluation.

Discussion Use of patient-safety dashboards has grown over the past 15 years, but impact remains poorly understood. Dashboard design processes rarely use informatics or human factors principles to ensure that the available content and navigation assists task completion, communication or decision making.

Conclusion Design and usability evaluation of patient safety dashboards should incorporate informatics and human factors principles. Future assessments should also rigorously explore their potential to support patient safety monitoring including direct or indirect impact on patient safety.

INTRODUCTION

Since the 2000 release of the Institute of Medicine's landmark report, *To Err is Human: Building a Safer Healthcare System*,¹ healthcare organisations have increasingly gathered, analysed and used data to improve the safety

of healthcare delivery. Despite increased research and quality improvement efforts, how data on patient safety events is communicated to people who will act on these data is not well understood. For instance, due to national quality reporting programmes, such as the Centers for Medicare & Medicaid Services' Quality Payment Program,² which adjusts healthcare organisation's reimbursement rates based on meeting certain quality measures, dashboards have been used extensively to visualise and disseminate process-based quality measures such as understanding how well haemoglobin A1c is controlled across all of a clinic's patients. However, an understanding of how commonly dashboards are used for patient safety-specific measures and how effective they are at advancing patient safety efforts and safety culture remains unknown.

Dashboards have been used extensively within and outside healthcare and serve as a form of visual information display that allows for efficient data dissemination.^{3 4} Dashboards aggregate data to provide overviews of key performance indicators to facilitate decision making, and when used correctly, enable efforts to improve an organisation's structure, process and outcomes.^{4 5} For dashboards to play a strategic role in communicating patient safety data, it is essential they are designed to relay key information about performance effectively.⁶ Thus, the dashboard design must consider informatics and human factors principles to ensure information is efficiently communicated. Informatics and human factors approaches have been successful in the design and evaluation of user interfaces in healthcare, and have variably been applied to dashboard development.⁷ One common approach is user-centred design, which is an iterative design process that aims to optimise usability of a display by focusing on users and their needs through requirement analysis, translation of requirements into design elements, application of design principles

Table 1 Agency for Healthcare Research and Quality Safety Targets

No	Safety topic	Examples
1	Alert Fatigue	Failure to recognise ventilator alarm
2	Device-related complication	Device malfunction
3	Diagnostic errors	Delayed stroke diagnosis, test misinterpretation
4	Discontinuities, gaps and hand-off problems	Missed critical lab result
5	Drug shortages	Antibiotics shortage
6	Failure to rescue	Death from postpartum haemorrhage
7	Fatigue and sleep deprivation	Resident errors due to sleep deprivation
8	Identification errors	Wrong-patient procedures
9	Inpatient suicide	Death of hospitalised patient
10	Interruptions and distractions	Incorrect surgical counts due to distractions
11	Medical complications	Falls, pressure ulcers, nosocomial infections, thromboembolism
12	Medication safety	Dispensing errors, medication-related hypoglycaemic or renal failure
13	MRI safety	Harm related to unsafe MRI practice
14	Nonsurgical procedural complications	Bedside procedure complications
15	Overtreatment	Complications after inappropriate antibiotic use
16	Psychological and social complications	Privacy violations
17	Second victims	Clinician emotional harm after adverse event
18	Surgical complications	Unexpected return to surgery, surgical site infection
19	Transfusion complications	Transfusion of incompatible blood types

From: <https://psnet.ahrq.gov/Topics>.
MRI, Magnetic Resonance Imaging.

and evaluation.⁸ Considering dashboards, usability would be defined as the extent to which a dashboard can be used by clinicians to understand and achieve specified goals with effectiveness, efficiency and satisfaction in clinical settings.⁹

Three main goals that guided this study were: (1) To understand the frequency and settings of use of patient safety dashboards in healthcare, (2) To determine the effectiveness of dashboards on directly or indirectly impacting patient safety at healthcare organisations and (3) To determine whether informatics and human factors principles are commonly used during dashboard development and evaluation. Our study focused on dashboards that displayed the frequency or rate of events, that is, those that facilitated retrospective review of past safety events to reduce these types of events in the future or dashboards that identified safety events of individual patients in real-time in order to mitigate further harm. We excluded dashboards that only displayed risk of an event.

METHODS

Design

We conducted a systematic literature review in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses guidelines.

Search strategy and inclusion criteria

We searched all available published and unpublished works in English using three literature databases

(MEDLINE via PubMed, EMBASE and CINAHL). Publications were eligible for inclusion if they included discussion about a dashboard for displaying patient safety event data in the healthcare setting. Patient safety event data were based on the list of ‘Safety Targets’ (table 1) on the Agency for Healthcare Research and Quality’s (AHRQ) Patient Safety Network (PSNet),¹⁰ and excluded process measures. Because of the variety of topics within patient safety, we ultimately used only the word ‘dashboard’ in our keyword and title search of all three databases, since this maximised the number of known publications identified without excluding relevant publications. Thus, our inclusion parameters, in PICOS format, were:

Population: Organisations providing medical care.

Interventions: Dashboards used to disseminate patient safety data (defined as measures related to any topic defined as a ‘Safety Target’ (table 1) by the AHRQ).¹⁰

Comparators: Settings with and without the use of patient safety dashboards.

Outcomes: (1) Settings where patient safety dashboards were used and (2) Impact of use of patient safety dashboards on reducing patient safety events.

Time frame: Studies published in English from 1 January 1950 to 30 August 2018.

Setting: Ambulatory care, inpatient and emergency department settings.

Screening process

After manually removing duplicates and non-journal publications (eg, magazine articles and book chapters),

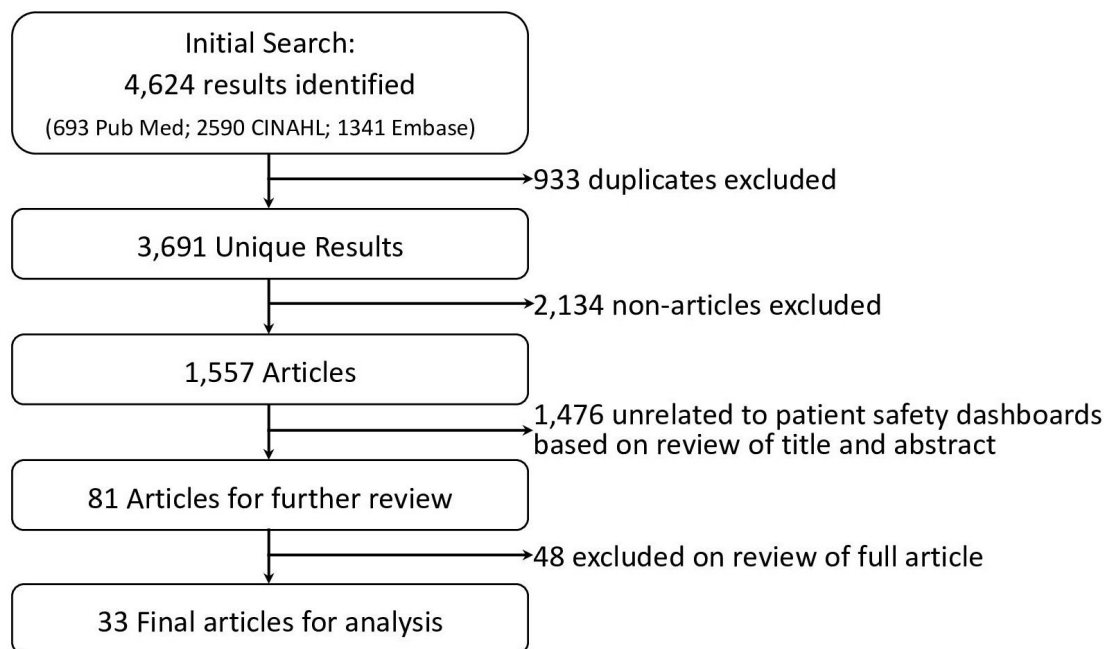


Figure 1 Flow chart of literature search results and the selection process of accepted/excluded publications.

two authors (DRM and TS) with expertise in clinical care, informatics and human factors reviewed titles and abstracts of each remaining article or abstract. Works were only included if they described display of patient safety event data (based on AHRQ's PSNet list of Patient Safety Targets) on a dashboard. Publications that discussed only non-safety event-related aspects of quality (eg, haemoglobin A1c control or rates of mammography screening) were excluded. Similarly, literature on dashboards displaying risk factors to prevent patient safety events rather than events themselves (eg, intensive care screens that display a particular patient's heart rate and oxygenation saturation or calculate a real-time risk level) were beyond the scope of this study and were excluded. We reviewed all publications potentially meeting study criteria in full. Reviewers discussed each inclusion, and disagreements regarding whether an article or abstract met criteria were resolved by consensus.

Publication evaluation

Three authors (DRM, TS and AS) independently extracted data from each identified publication using a structured review form. Reviewers specifically identified (1) the setting the dashboard was used in, (2) the patient safety topic displayed on the dashboard, (3) the type of informatics or human factors principles used in dashboard design or usability evaluation performed on the final dashboard and (4) the impact of the dashboard, both related to reducing patient safety events in the setting where it was used and other impacts identified by each publication's authors. To assess the level of evidence in improving patient safety, reviewers also assessed the study type and whether a control or other comparison group was used. Findings are aggregated and reported using descriptive statistics.

RESULTS

Our literature search identified a total of 4624 results (PubMed: 693, CINAHL: 2590, Embase: 1341). After 933 duplicates were removed, 3691 result entries remained. One reviewer (TS) subsequently removed 2134 magazine articles, newspaper articles, thesis papers, conference papers, reports that were unrelated to the topic of patient safety, as well as publications not in English. Titles and abstracts of the remaining 1557 articles and conference abstracts were independently reviewed by two reviewers (TS and DRM). Reviewers manually reviewed titles and abstracts and excluded (A) publications that did not include discussion of a dashboard as a primary or secondary focus, and (B) publications where dashboards were mentioned, but the dashboard did not include measures related to any of the AHRQ 'Safety Targets' (table 1). After exclusions, reviewers identified a combined total of 81 publications that warranted further review of the entire publication. Reviewers discussed each publication, and after consensus, identified 33 final publications that warranted inclusion in the analysis. Reference sections of each publication were reviewed for additional sources but did not identify additional publications. Figure 1 displays a flow chart of the search strategy.

Search results

The final set included 33 publications, including 5 conference abstracts and 28 full articles (table 2). The earliest publications describe use of patient safety measures on a dashboard in 2004, 2005 and 2006,^{11–13} followed by a paucity of additional publications until 2010.

Clinical settings

All patient safety dashboards were used in the hospital setting, often at the level of the entire hospital or hospital

Table 2 Final studies using patient safety dashboards identified during literature search

Citation	Type	Setting	Safety topic	Study type
Anand (2015) ¹⁴	Article	Paediatric cardiac ICU	Pressure ulcers, unplanned extubation, hospital infections (CAUTI, CLABSI, VAP)	Case report
Bakos (2012) ²⁴	Article	Trauma centre	Hospital infections (CLABSI)	Case report
Chandharan (2010) ¹⁷	Article	Maternity ward	Postpartum haemorrhage	Case report
Coleman (2013) ¹⁸	Article	Hospital wards	Medication-related events	Time series
Collier (2015) ²²	Article	Inpatient maternity and paediatrics wards	Pressure ulcers	Case report
Conway (2012) ²⁵	Article	Trauma centre	Surgical site infections	Case report
Dharamshi (2011) ²⁷	Article	Surgery	Return to surgery	Case report
Donaldson (2005) ¹²	Article	Surgery, critical care floors	Pressure ulcers, falls	Case report
Fong (2017) ²³	Article	Pharmacy	Medication-related events	Case report
Frazier (2012) ²⁹	Article	Whole hospital	Falls, hospital infections (MRSA, C. Diff, VAP, CLABSI, CAUTI), Pressure ulcers	Case report
Gardner (2015) ³⁶	Article	Whole hospital	Falls	Case report
Hebert (2018) ¹⁵	Article	Cardiac surgery unit and ICU	Hospital infections (VAP)	Time series
Hendrickson (2013) ³⁰	Abstract	Whole hospital	Hospital infections (VAP, CLABSI, CAUTI, MRSA, VRE, C. Diff)	Case report
Hyman (2017) ³⁷	Article	Whole hospital	Hospital infections (CLABSI, CAUTI, CAP), falls, VTE	Case report
Johnson (2006) ¹³	Article	Whole hospital	Medication-related events	Case report
Lau (2012) ¹⁹	Abstract	Hospital oncology and GI departments	Delays in biopsy follow-up	Case report
Lo (2014) ³³	Article	Whole hospital	Hospital infections (CAUTI)	Case report
Mackie (2014) ³⁵	Article	Whole hospital	Pressure ulcers	Time series
Madison (2013) ³¹	Abstract	Whole hospital	Hospital infections (CLABSI)	Case report
Mane (2018) ²⁶	Article	Emergency department	Delays in CVA diagnosis	Case report
Mayfield (2013) ¹⁶	Abstract	ICU, oncology ward	Hospital infections (CLABSI, VAP)	Case report
Mazzella-Ebstein (2004) ¹¹	Article	Hospital wards	Pressure ulcers, falls, DVTs	Case report
Milligan (2015) ⁴¹	Article	Whole hospital	Hypoglycaemic	Time series
Mlaver (2017) ²⁰	Article	Hospital floor	Pressure ulcers, hypoglycaemic	Case report
Nagelkerk (2014) ⁵⁰	Article	Paediatrics ward	Hospital deaths	Case report
Pemberton (2014) ⁵¹	Article	Dental hospital	Wrong-site surgery, falls, medication errors	Case report
Rao (2011) ³²	Abstract	Whole hospital	Hospital infections (VAP)	Case report
Ratwani (2015) ³⁸	Article	Whole hospital	Falls	Case report
Riley (2010) ³⁴	Article	Whole hospital	Hospital infections (MRSA, C. Diff), falls, pressure ulcers, medication errors	Case report
Rioux (2007) ²⁸	Article	Surgery	Surgical site infections	Time series
Skledar (2013) ³⁹	Article	Whole hospital	Medication-related events	Case report
Stone (2018) ⁴⁰	Article	Whole hospital	Medication-related events	Case report
Waitman (2011) ²¹	Article	Hospital wards, pharmacy	Renal failure	Case report

CAUTI, catheter-associated urinary tract infection; C. Diff, *Clostridium difficile*; CLABSI, central line-associated blood stream infection; CVA, cerebrovascular accident; DVT, deep vein thrombosis; GI, gastrointestinal; ICU, intensive care unit; MRSA, methicillin-resistant *Staphylococcus aureus*; VAP, ventilator-associated pneumonia; VRE, vancomycin-resistant enterococcus infection; VTE, Venous Thromboembolism;

system. Several patient safety dashboards were used in ICUs,^{12 14–16} hospital wards,^{11 12 17–22} pharmacies,^{21 23} emergency departments and trauma centres,^{24–26} and surgical settings.^{12 27 28} No use of patient safety dashboards was identified in the ambulatory care setting.

Patient safety topics

The most common use of patient safety dashboards (11 of 33) was tracking hospital infections (figure 2). Types

of infection tracked included central line-related blood stream infections,^{14 16 29–31} ventilator-associated pneumonia,^{14 16 29 30 32} catheter-associated urinary tract infections,^{14 29 30 33} methicillin-resistant *Staphylococcus aureus* infections,^{29 30 34} vancomycin-resistant *Enterococcus* infections³⁰ and *Clostridium difficile* infections.^{29 30} Dashboards additionally displayed rates of pressure ulcers,^{11 12 14 20 22 34 35} patient falls^{11 29 36–38} and medication-related errors,^{13 18 23 39 40}

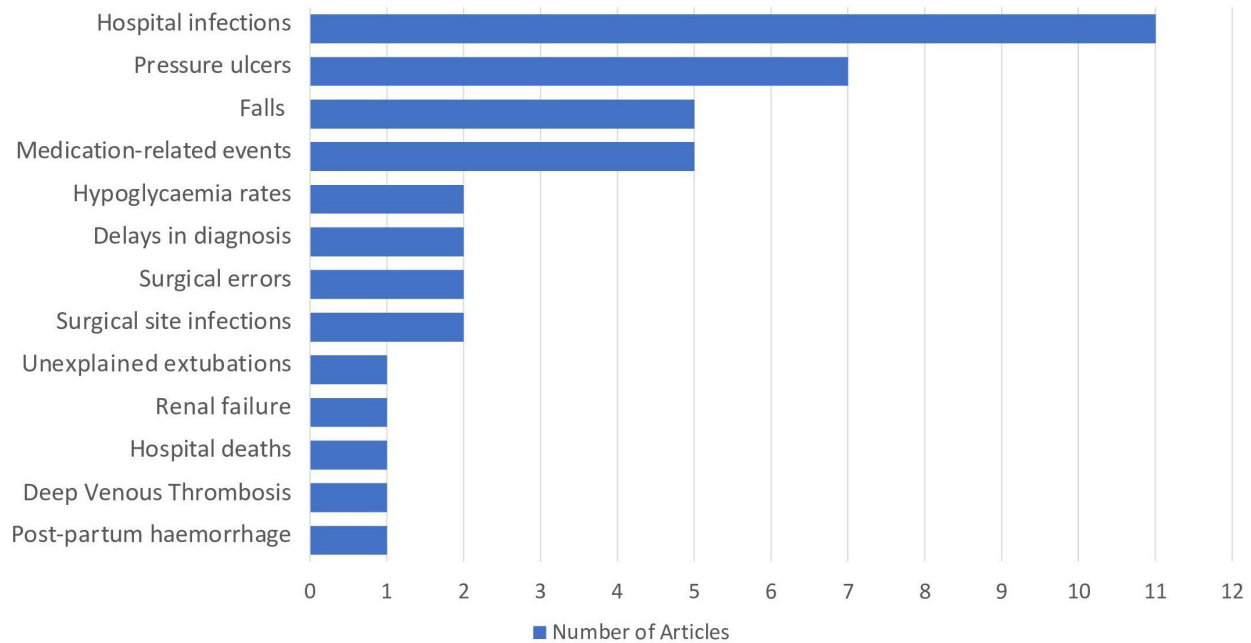


Figure 2 Number of publications identified by dashboard patient safety topic.

followed less commonly by other patient safety topics (See table 2 for all safety topics and figure 2 for chart of topic frequencies).

Impact of Dashboard use and level of evidence

Of all studies identified, 5 used a time series design^{15 18 28 35 41} while the remaining 28 used case report designs describing specific implementations of patient safety dashboards without statistical analyses performed. Of the five time series studies, Coleman *et al*¹⁸ identified a 0.41% decrease in missed doses of medications other than antibiotics ($p=0.007$); however, it was part of four concurrent interventions to reduce missed and delayed medication doses, and thus, the specific impact of the dashboard was unclear. Similarly, Milligan *et al*⁴¹ reported a reduction in hypoglycaemic rates, Rioux *et al*²⁸ reported a decrease in surgical site infections over a 6-year period after dashboard implementation, and Mackie *et al*³⁵ reported a reduction in hospital-acquired pressure ulcers; however, in each case, the dashboard was one aspect of a broader campaign to reduce the respective patient safety events. Other studies, including Bakos, Chandraharan, Collier, Conway, Hebert, Hendrickson and Hyman,^{15 17 22 24 25 30 37} reported a subjective reduction in patient safety events, but did not describe a statistical analysis. The remaining publications did not include discussion of the direct or indirect impact of the dashboard on patient safety events.

Most publications that evaluated the dashboard focused instead on sensitivity and specificity of dashboard measures, employee satisfaction with the dashboards and reduction in time required to gather data for the dashboard compared with previous manual data collection. Another impact of dashboards described included dissemination of patient event data in real time or closer to real time than previously possible due to algorithms that monitor electronic

patient safety data and automatically update dashboards. Direct impact on culture and staffing levels of patient safety personnel were not described in any of the studies. However, as described above, several studies implemented dashboards as a package with other patient safety-focused efforts, suggesting changes in culture, infrastructure, and staffing likely occurred, but concomitantly with the dashboard implementation rather in response to it.

Usability

Only two studies used a human factors approach for design and evaluation of dashboards. Ratwani and Fong³⁸ described a development process employing commonly accepted human factors design principles,⁴² followed by focus groups with users and a 2-week pilot phase to collect usability data and make improvements to the dashboard. Mlaver *et al*²⁰ used a participatory design approach that employed collaboration with users during iterative refinements. Two additional studies discussed more limited efforts to obtain feedback. Dharamshi *et al*²⁷ performed a limited usability analysis with an anonymous survey of dashboard users at 6-months after implementation to understand factors that limited the usability of the dashboard. Stone *et al*⁴⁰ iteratively obtained feedback from physician users between dashboard revisions. However, the majority of studies did not describe the use of an informatics or human factors approach that considered usability design principles, user-centred design processes or usability evaluation methods. Thus, there was little evidence of design elements that were most useful or usable across scenarios or settings.

DISCUSSION

Our systematic review identified 33 publications discussing the use of dashboards to communicate and visualise

patient safety data. All publications were published since 2004, suggesting increased measurement of patient safety after the 1999 publication of *To Err is Human*. All publications involved display of patient safety events in the inpatient setting, the most common of which were hospital acquired infections. There may, thus, exist opportunities for similar efforts in the ambulatory setting (eg, falls, lost referrals, abnormal test results lost to follow-up or medication prescribing errors).

Overall, the level of evidence that dashboards directly or indirectly impact patient safety was limited. Only five of the publications used time series designs with the remaining designs comprised of case reports of dashboard implementations either alone or as part of broader patient safety interventions. No interventional studies were identified. Most studies reported on accuracy of the measures displayed or survey-based user satisfaction with the dashboard, rather than the dashboards' impact on patient safety events. Studies that provided data on reductions in patient safety events either did not report statistical analyses to support the reduction, or more commonly, were part of a broad process improvement effort containing multiple interventions, making it difficult to tease out which intervention truly impacted safety. While it can be argued that the intent of a patient safety dashboard is to communicate data about the extent of safety issues at an organisation and support other improvement efforts, the act of showing data via a dashboard may alone have an impact of motivating quality and safety efforts. Dashboards likely have impacts on safety culture and indirectly lead to allocation of resources to reducing patient safety events. The studies identified did not describe these impacts in response to dashboard implementation, and thus, this topic warrants future exploration.

Most publications described dashboard development as a quality improvement approach to addressing a specific organisational problem or to meet institutional or national standards. Several studies reported high user satisfaction with the dashboard, though these were often limited assessments and did not capture whether users fully understood the content of the dashboard. With four exceptions, studies lacked informatics or human factors design approaches during development, application of standardised design principles and use of usability evaluations. Without informatics, human factors or user-centred design approaches, information requirements from users may not be well understood. Thus, there is limited evidence about the dashboard acceptance, frequency of use or whether dashboards satisfactorily met the needs of intended users. For example, a common mention was use of colour coding following a traffic light scheme (red=poor status, yellow=warning, green=good status), without a formal evaluation of the usability for the 8% of men and 0.5% of women in the population with red-green colour blindness.⁴³

Some dashboards were implemented within a bundle of other interventions. The lack of dashboard usability testing before and after implementation made it difficult

to identify the impact or effect of the dashboard. As with many clinical informatics interventions, there could be numerous social and/or technical factors that may have influenced the reported outcomes beyond the dashboard. Rigorous informatics and human factors design approaches^{44–47} are needed to improve the use and impact of patient safety dashboards. Because intervention development is often time constrained, rapid qualitative assessment approaches or human factors methods involving rapid prototyping,^{48–49} for example, can be adapted to meet the shorter timelines needed for rapid cycle quality improvement. This will ensure dashboards are useful and usable and generate much needed evidence about efficiency, effectiveness and satisfaction in various care settings.

Our study has several limitations. First, it is subject to a potential reporting bias. While we analysed publications based on the content reported, it is possible that additional statistical analyses and usability assessments were performed that were not reported. Furthermore, there is likely to be greater use of patient safety dashboards developed as part of routine quality improvement efforts within healthcare organisations, but these may not be published. Nevertheless, this is an area that is ripe for additional research. Second, there was a significant variability in how dashboards were described, ranging from basic text descriptions to full-colour screenshots. This variability made performing standardised usability assessments impossible. Finally, our search was limited to the publications present in the databases we searched. While we used three different databases to mitigate this impact, if publications did not appear in any of our search databases, they would have been missed.

In conclusion, we identified a growing use of patient safety dashboards, largely focused on displaying inpatient safety events. Due to limited use of informatics and human factors-based approaches during development or postimplementation evaluation, the usability of such dashboards was difficult to assess. Furthermore, because of limited evaluation of the impact of dashboards and because dashboards were often implemented as part of a variety of process improvement efforts, the literature is not clear on direct impact of dashboard implementation on patient safety events. Because well-designed dashboards have potential to support patient safety monitoring, our study should encourage integration of informatics and human factors principles into design and usability evaluation of dashboards as well as assessment of their direct or indirect impact on patient safety.

Twitter Dean F Sittig @DeanSittig and Hardeep Singh @HardeepSinghMD

Contributors DRM, DS and HS developed idea for this systematic review. DRM and TS performed the literature search. DRM, TS and AS critically reviewed and extracted data from the publications identified. All authors contributed to the writing of the initial manuscript and of revising subsequent versions. All authors had control over the decision to publish. DRM had access to the full data set and accepts full responsibility for the finished article.

Funding This project was funded by an Agency for Healthcare Research and Quality Mentored Career Development Award (K08-HS022901) and partially

funded by the Houston VA HSR&D Center for Innovations in Quality, Effectiveness and Safety (CIN 13-413). HS is additionally supported by the VA Health Services Research and Development Service (IIR17-127; Presidential Early Career Award for Scientists and Engineers USA 14-274), the VA National Center for Patient Safety, the Agency for Health Care Research and Quality (R01HS27363), and the Gordon and Betty Moore Foundation (GBMF 5498 and GBMF 8838). AS is additionally supported by the VA HSR&D Center for Health Information and Communication (CIN 13-416), National Institutes of Health, National Center for Advancing Translational Sciences, and Clinical and Translational Sciences Award (KL2TR002530 and UL1TR002529). There are no conflicts of interest for any authors.

Disclaimer These funding sources had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; and preparation, review, or approval of the manuscript.

Competing interests None declared.

Patient consent for publication Not applicable.

Provenance and peer review Commissioned; externally peer reviewed.

Data availability statement Data sharing not applicable as no datasets generated and/or analysed for this study. Not applicable.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>.

ORCID iD

Dean F Sittig <http://orcid.org/0000-0001-5811-8915>

REFERENCES

- Kohn LT, Corrigan JM, Donaldson MS, Committee on Quality of Health Care in America, Institute of Medicine. *To err is human: building a safer health system*. The National Academies Press, 2000.
- Quality payment program. United States Centers for Medicare & Medicaid Services. Available: <https://qpp.cms.gov/> [Accessed 6/6/2021].
- Hugo JV, S SG. *Human factors principles in information Dashboard design*. Idaho Falls, ID (United States: Idaho National Lab. (INL), 2016.
- Dowding D, Randell R, Gardner P, et al. Dashboards for improving patient care: review of the literature. *Int J Med Inform* 2015;84:87–100.
- Sarikaya A, Correll M, Bartram L, et al. What do we talk about when we talk about Dashboards? *IEEE Trans Vis Comput Graph* 2019;25:682–92.
- Dowding D, Merrill JA. The development of Heuristics for evaluation of Dashboard Visualizations. *Appl Clin Inform* 2018;9:511–8.
- Carayon P, Hoonakker P. Human factors and usability for health information technology: old and new challenges. *Yearb Med Inform* 2019;28:071–7.
- Norman DA, Draper SW. *User centered system design: new perspectives on Human-Computer interaction*. 1st ed. CRC Press, 1986.
- 20282-2 usability of consumer products and products for public use, 2013 International Organization of Standards. Available: <https://www.iso.org/obp/ui/#iso:std:iso:ts:20282:-2:ed-2:v1:en>
- PSNet topics, 2021. Available: <https://psnet.ahrq.gov/topics-0> [Accessed 6/6/2021].
- Mazzella-Ebstein AM, Saddul R. Web-Based nurse executive dashboard. *J Nurs Care Qual* 2004;19:307–15.
- Donaldson N, Brown DS, Aydin CE, et al. Leveraging nurse-related dashboard benchmarks to expedite performance improvement and document excellence. *J Nurs Adm* 2005;35:163–72.
- Johnson K, Hallsey D, Meredith RL, et al. A nurse-driven system for improving patient quality outcomes. *J Nurs Care Qual* 2006;21:168–75.
- Anand V, Cave D, McCrady H, et al. The development of a congenital heart programme quality dashboard to promote transparent reporting of outcomes. *Cardiol Young* 2015;25:1579–83.
- Hebert C, Flaherty J, Smyer J, et al. Development and validation of an automated ventilator-associated event electronic surveillance system: a report of a successful implementation. *Am J Infect Control* 2018;46:316–21.
- Mayfield J, Wood H, Russo AJ, et al. Facility level dashboard utilized to decrease infection preventionists time disseminating data. *Am J Infect Control* 2013;41:S54.
- Chandrarahan E. Clinical dashboards: do they actually work in practice? Three-year experience with the maternity Dashboard. *Clin Risk* 2010;16:176–82.
- Coleman JJ, Hodson J, Brooks HL, et al. Missed medication doses in hospitalised patients: a descriptive account of quality improvement measures and time series analysis. *Int J Qual Health Care* 2013;25:564–72.
- Lau S, Wehbi M, Varma V. Electronic tool for tracking patient care after positive colon biopsy. *Am J Clin Pathol* 2012;138:A122.
- Mlaver E, Schnipper JL, Boxer RB, et al. User-Centered collaborative design and development of an inpatient safety dashboard. *Jt Comm J Qual Patient Saf* 2017;43:676–85.
- Waitman LR, Phillips IE, McCoy AB, et al. Adopting real-time surveillance dashboards as a component of an enterprisewide medication safety strategy. *Jt Comm J Qual Patient Saf* 2011;37:326–AP4.
- Collier M. Pressure Ulcer Incidence: The Development and Benefits of 10 Year's-experience with an Electronic Monitoring Tool (PUNT) in a UK Hospital Trust. *EWMA J* 2015;15:15–20.
- Fong A, Harriott N, Walters DM, et al. Integrating natural language processing expertise with patient safety event review committees to improve the analysis of medication events. *Int J Med Inform* 2017;104:120–5.
- Bakos KK, Zimmermann D, Moriconi D. Implementing the clinical Dashboard at VCUHS. *NI* 2012;2012:11.
- Conway WA, Hawkins S, Jordan J, et al. The Henry Ford health system no harm campaign: a comprehensive model to reduce harm and save lives. *Jt Comm J Qual Patient Saf* 2012;38:318–AP1.
- Mane KK, Rubenstein KB, Nassery N, et al. Diagnostic performance dashboards: tracking diagnostic errors using big data. *BMJ Qual Saf* 2018;27:567–70.
- Dharamshi R, Hillman T, Shaw R. Increasing engagement with clinical outcome data. *Br J Healthc Manag* 2011;17:585–9.
- Rioux C, Grandbastien B, Astagneau P. Impact of a six-year control programme on surgical site infections in France: results of the INCISO surveillance. *J Hosp Infect* 2007;66:217–23.
- Frazier JA, Williams B. Successful implementation and evolution of Unit-Based nursing Dashboards. *Nurse Leader* 2012;10:44–6.
- Hendrickson C, Guelcher A, Guspiel AM, et al. Results of increasing the frequency of healthcare associated infections (HAI) data reported to managers with an embedded quality improvement process. *Am J Infect Control* 2013;41:S114.
- Madison AS, Johnson D, Shepard J. Leveraging internal knowledge to create a central line associated bloodstream infection surveillance and reporting tool. *Am J Infect Control* 2013;41:S56.
- Rao R. Application of six sigma process to implement the infection control process and its impact on infection rates in a tertiary health care centre. *BMC Proc* 2011;5.
- Lo Y-S, Lee W-S, Chen G-B, et al. Improving the work efficiency of healthcare-associated infection surveillance using electronic medical records. *Comput Methods Programs Biomed* 2014;117:351–9.
- Riley S, Cheema K. Quality Observatories: using information to create a culture of measurement for improvement. *Clin Risk* 2010;16:93–7.
- Mackie S, Baldie D, McKenna E, et al. Using quality improvement science to reduce the risk of pressure ulcer occurrence – a case study in NHS Tayside. *Clin Risk* 2014;20:134–43.
- Gardner LA, Bray PJ, Finley E, et al. Standardizing falls reporting: using data from adverse event reporting to drive quality improvement. *J Patient Saf* 2019;15:135–42.
- Hyman D, Neiman J, Rannie M, et al. Innovative use of the electronic health record to support harm reduction efforts. *Pediatrics* 2017;139:e20153410–e10.
- Ratwani RM, Fong A. 'Connecting the dots': Leveraging visual analytics to make sense of patient safety event reports. *J Am Med Inform Assoc* 2015;22:312–7.
- Skledar SJ, Niccolai CS, Schilling D, et al. Quality-Improvement analytics for intravenous infusion pumps. *Am J Health Syst Pharm* 2013;70:680–6.
- Stone AB, Jones MR, Rao N, et al. A Dashboard for monitoring Opioid-Related adverse drug events following surgery using a national administrative database. *Am J Med Qual* 2019;34:45–52.
- Milligan PE, Bocox MC, Pratt E, et al. Multifaceted approach to reducing occurrence of severe hypoglycemia in a large healthcare system. *Am J Health Syst Pharm* 2015;72:1631–41.
- Shneiderman B. The eyes have it: a task by data type taxonomy for information Visualizations. *Proceedings 1996 IEEE Symposium on Visual Languages* 1996:336–43.



- 43 Roskoski R. Guidelines for preparing color figures for everyone including the colorblind. *Pharmacol Res* 2017;119:240–1.
- 44 Holden RJ, Carayon P. SEIPS 101 and seven simple SEIPS tools. *BMJ Qual Saf* 2021. doi:10.1136/bmjqs-2020-012538. [Epub ahead of print: 26 May 2021].
- 45 Carayon P, Hoonakker P, Hundt AS, *et al.* Application of human factors to improve usability of clinical decision support for diagnostic decision-making: a scenario-based simulation study. *BMJ Qual Saf* 2020;29:329–40.
- 46 Savoy A, Militello LG, Patel H, *et al.* A cognitive systems engineering design approach to improve the usability of electronic order forms for medical consultation. *J Biomed Inform* 2018;85:138–48.
- 47 Harte R, Glynn L, Rodríguez-Molinero A, *et al.* A Human-Centered design methodology to enhance the usability, human factors, and user experience of connected health systems: a three-phase methodology. *JMIR Hum Factors* 2017;4:e8.
- 48 Wilson J, Rosenberg D. Chapter 39 - Rapid Prototyping for User Interface Design. In: Helander M, ed. *Handbook of Human-Computer Interaction*. North. Holland, 1988: 859–75.
- 49 McMullen CK, Ash JS, Sittig DF, *et al.* Rapid assessment of clinical information systems in the healthcare setting: an efficient method for time-pressed evaluation. *Methods Inf Med* 2011;50:299–307.
- 50 Nagelkerk J, Peterson T, Pawl BL, *et al.* Patient safety culture transformation in a children's Hospital: an interprofessional approach. *J Interprof Care* 2014;28:358–64.
- 51 Pemberton MN, Ashley MP, Shaw A, *et al.* Measuring patient safety in a UK dental Hospital: development of a dental clinical effectiveness dashboard. *Br Dent J* 2014;217:375–8.

© 2021 Author(s) (or their employer(s)) 2021. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ. This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>. Notwithstanding the ProQuest Terms and Conditions, you may use this content in accordance with the terms of the License.

Clinical risk prediction models: the canary in the coalmine for artificial intelligence in healthcare?

Videha Sharma , Angela Davies , John Ainsworth

To cite: Sharma V, Davies A, Ainsworth J. Clinical risk prediction models: the canary in the coalmine for artificial intelligence in healthcare? *BMJ Health Care Inform* 2021;**28**:e100421. doi:10.1136/bmjhci-2021-100421

Received 28 May 2021
Accepted 03 September 2021

Clinical risk prediction models (CRPMs) are statistical models that aim to improve medical decision making by providing an objective measure of potential health outcomes based on data.¹ In recent years, there has been an explosion in the development of such models across all areas of medical care. Additionally, the adjunct of artificial intelligence (AI), which builds on traditional statistical prediction methods using machine learning, models are increasing in complexity and potential accuracy.² However, the adoption pathway of CRPMs and AI models is similar, with both requiring access to data for model development, regulatory approval and effective implementation into the workflow of clinicians.

Despite significant potential, only few CRPMs have been implemented into practice and achieved patient benefit.³ In this editorial we explore the incentives of producers, intermediaries and users of CRPMs and discuss how a lack of alignment has failed to realise their potential to achieve intended benefits. AI in healthcare faces a similar threat and we propose a novel solution to mitigate this for the future.

Currently, models are mainly produced in the academic context, where there is access to data and methodological expertise. Researchers are incentivised by traditional academic objectives, such as published papers, conference presentations or other scientific accolades. This results in a failure to pursue a successfully validated model beyond these goals. If further motivation (and often funding) allows, efforts are focused on improving the statistical accuracy or undertaking external validation studies, rather than exploring implementation or clinical usability.

Furthermore, any software, such as a CRPM or a clinical AI model, is by definition a medical device.⁴ This means that they are

subject to conformity assessments and regulatory approval prior to being placed on the market.⁵ Most researchers will not have the expertise to undertake this, or have access to the relevant support to navigate this process. Additionally, software requires regular updates and maintenance, which may come with considerable running costs. Currently, there is no clear understanding of whom to attribute these to, yet they are critical to the longer-term safety, efficacy and viability of a CRPM.

Clinically validated and approved models are typically implemented as stand-alone web or mobile applications. This creates usability barriers, as users access an external interface and manually transcribe data to receive results.⁶ In reality, electronic health record (EHR) vendors could act as intermediaries and integrate models directly into their systems, complementing the clinical workflow. However, the vendors would have to take responsibility for the medical device regulation, maintenance and associated costs. The value proposition for vendors to foot these costs and risks is currently not there, especially as clinical stakeholders do not yet expect such functionalities in EHRs. An alternative could be for vendors to provide third party companies an application programming interface (API) to their EHR. However, as there is no single API standard, third party model suppliers would have to integrate with each EHR individually. This would be compounded with ongoing licensing fees, making for a precarious business model for what are usually small enterprises.

ALIGNING INCENTIVES WITH BLOCKCHAIN TECHNOLOGY

The individual incentives of the producers (researchers), intermediaries (EHR vendors) and users (healthcare providers) are



© Author(s) (or their employer(s)) 2021. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

Centre for Health Informatics,
School of Health Sciences,
The University of Manchester,
Manchester, UK

Correspondence to

Dr Videha Sharma;
videha.sharma@postgrad.
manchester.ac.uk

currently unaligned. The academic environment or EHR market has not incentivised the conversion of technical discovery to integrated product development, limiting the ‘bench-to-bedside’ pathway of CRPMs. To prevent a similar experience for AI models, we must develop strategies that align incentives and create a value proposition for all involved parties.

A potential solution could be a national infrastructure; a marketplace for models, all clinically validated and compliant with medical device regulations. Blockchain, a form of distributed ledger technology (DLT), may facilitate such an infrastructure by securely hosting the marketplace and allowing the producers to be remunerated when their model is used through smart contracts.

Blockchain is an open network of distributed data stored in secure blocks, which are available to all participants (known as ‘nodes’) on a network.⁷ By distributing blocks across all nodes, the data in the network is difficult to hack, change or corrupt, creating a traceable, immutable and secure record of transactions between nodes.⁸ Blockchain has therefore been widely discussed in the context of sharing electronic patient records.⁹ Smart contracts are a digital technology that execute a financial transaction recorded in a blockchain when a predefined condition is met.¹⁰

Blockchain and DLT could support the implementation and financial reward for CRPMs: models could be published to the national marketplace, hosted on the blockchain and clinical data could be entered securely to receive results with a micro-payment triggered at every use, via smart contracts. Defining a national vendor-neutral API standard for models would make the marketplace accessible from all EHRs that implement it. A recognised body could regulate this process alongside an established framework, such as the UK government’s guide to good practice for digital and data-driven health technologies.¹¹ The traceability provided through a DLT-based solution would build trust among all stakeholders and allow a shared interest to develop.

An example of a CRPM that this could apply to is the CHA(2)DV(2)-VASc score, which is used to predict the risk of stroke in patients with atrial fibrillation, and thus guide the need for blood-thinning medication.¹² The producers would publish their model to the marketplace, who would take the responsibility of assessment conformity and regulatory approval. Once the model is live, EHR vendors could integrate it into their interface using the standardised API.

This would increase the use of CRPMs by clinicians as they are incorporated into their workflow, provide a monetary incentive for researchers to pursue models to implementation and integration and finally, make EHRs that integrate CRPMs more attractive for healthcare providers to procure. [Figure 1](#) illustrates this concept, highlighting how blockchain technology can align incentives and operationalise current and future CRPMs and AI models.

The traditional medical research path is linear with rigid objectives and little concern for commercialisation. However, it is evidence-focused and rightly, prioritises safety and regulation. In contrast, technology

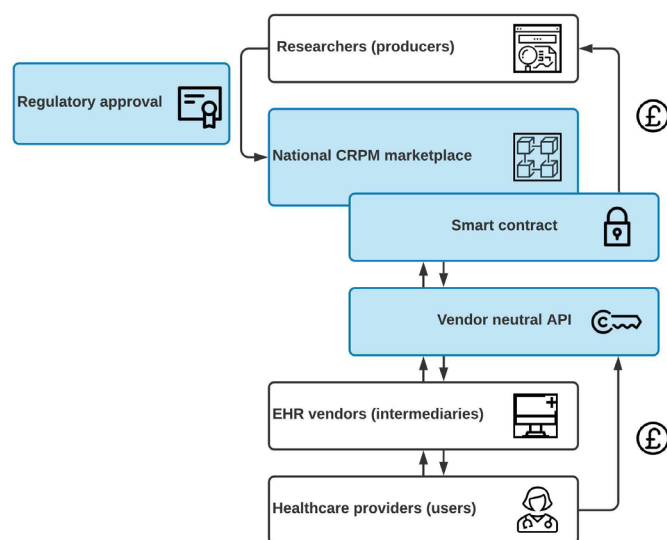


Figure 1 Conceptual representation of a vendor-neutral distributed ledger-based CRPM marketplace to maintain data security with the use of smart contracts to facilitate micro-payments. API, application programming interface; CRPMs, clinical risk prediction models; EHR, electronic health record.

development is agile, iterative and focused on real-world application.¹³ There remains a need to create a joint culture across academic and industry stakeholders to harmonise expertise and develop meaningful digital health solutions. Recent efforts, such as the proposed Decision-support systems driven by artificial intelligence guidelines, support this by calling for early clinical evaluation with a view to bridging the current implementation gap of AI models.¹⁴

The introduction of Chief Clinical Informatics Officers and digital strategies by healthcare providers will help regulate and adopt these technologies going forward,¹⁵ however, a collaboration across the vendor industry remains essential. A drive towards business success may incentivise researchers, vendors and healthcare providers appropriately to pursue solutions and achieve intended benefits. Interdisciplinary and cross-industry health research, with a long-term focus on clinical impact can thus unlock the potential of CRPMs and AI, leading to radical change in patient care and outcomes.

Twitter Videha Sharma @VidehaSharma

Contributors JA conceptualised the manuscript. VS reviewed the literature and wrote the manuscript. JA and AD reviewed and edited the manuscript. JA and VS cocreated the figure.

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 AD is a member of the editorial board of *BMJ Health and Care Informatics*.

Patient consent for publication Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is

properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>.

ORCID iDs

Videha Sharma <http://orcid.org/0000-0001-7640-1239>

Angela Davies <http://orcid.org/0000-0002-3365-7231>

REFERENCES

- 1 Steyerberg EW. *Clinical prediction models*. Springer, 2019.
- 2 Bzdok D, Altman N, Krzywinski M. Statistics versus machine learning. *Nat Methods* 2018;15:233–4.
- 3 Kwan JL, Lo L, Ferguson J, et al. Computerised clinical decision support systems and absolute improvements in care: meta-analysis of controlled clinical trials. *BMJ* 2020;370:m3216.
- 4 Spanou D. *Software as a medical device (SaMD): key definitions*. IMDRF SaMD Working Group, 2013.
- 5 MHRA G. *Medical device stand-alone software including apps (including IVDMDs)*. UK Government Policy, 2019.
- 6 Sharma V, Ali I, van der Veer S, et al. Adoption of clinical risk prediction tools is limited by a lack of integration with electronic health records. *BMJ Health Care Inform* 2021;28:e100253.
- 7 Leeming G, Ainsworth J, Clifton DA. Blockchain in health care: hype, trust, and digital health. *The Lancet* 2019;393:2476–7.
- 8 Cunningham J, Ainsworth J. Enabling patient control of personal electronic health records through distributed ledger technology. *Stud Health Technol Inform* 2018;245:45–8.
- 9 Mayer AH, da Costa CA, Righi RdaR. Electronic health records in a blockchain: a systematic review. *Health Informatics J* 2020;26:1273–88.
- 10 Cannarsa M. Interpretation of contracts and smart contracts: smart interpretation or interpretation of smart contracts? In: *European review of private law*. 26, 2018.
- 11 Public Health England Guidance. *Guidance on social distancing for everyone in the UK*, 2020.
- 12 Lip GYH, Nieuwlaat R, Pisters R, et al. Refining clinical risk stratification for predicting stroke and thromboembolism in atrial fibrillation using a novel risk factor-based approach: the euro heart survey on atrial fibrillation. *Chest* 2010;137:263–72.
- 13 Steinberg D, Horwitz G, Zohar D. Building a business model in digital medicine. *Nat Biotechnol* 2015;33:910–20.
- 14 Watkinson P, Clifton D, Collins G, et al. DECIDE-AI: new reporting guidelines to bridge the development-to-implementation gap in clinical artificial intelligence. *Nat Med* 2021;27:186–7.
- 15 Wachter R. *Making it work: harnessing the power of health information technology to improve care in England*. London, UK: Department of Health, 2016.

© 2021 Author(s) (or their employer(s)) 2021. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ. This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>. Notwithstanding the ProQuest Terms and Conditions, you may use this content in accordance with the terms of the License.

COVID-19 pandemic and artificial intelligence: challenges of ethical bias and trustworthy reliable reproducibility?

Casimir Kulikowski ,¹ Victor Manuel Maojo²

To cite: Kulikowski C, Maojo VM. COVID-19 pandemic and artificial intelligence: challenges of ethical bias and trustworthy reliable reproducibility? *BMJ Health Care Inform* 2021;**28**:e100438. doi:10.1136/bmjhci-2021-100438

Received 01 July 2021
Accepted 17 September 2021

Rapid vaccine breakthroughs for the SARS-CoV-2 viral pandemic have been enabled by genomics-based designs and biomedical informatics-driven experimentation relying on many algorithmic and artificial intelligence (AI) methods. Great hopes expressed about informatics for the humanitarian amelioration of pandemics internationally depend on data analytics and AI for predicting COVID-19 spread and public health prevention measures, diagnoses and treatments. Yet, bioinformatics-enabled vaccine development has turned out to be the only truly indispensable technological work-around compensating for the tragic worldwide shortcomings in pandemic responses and insufficiencies in epidemiological genomics data infrastructures.¹

Any AI in a healthcare informatics system must target recommendations and actions to individual patients and this requires high-quality relevant data to be extracted and prioritised from heterogeneous mixes of statistics, for which much more sophisticated and reproducible methods of semantic annotation, knowledge-based design and cross-validation are needed than commonly used today. These need to build on experience with multiple methods of expert-knowledge representation and inference beyond purely data-driven machine learning. Especially important is to identify high-risk or vulnerable subpopulations to avoid biased misapplication of machine learning and other AI techniques that could exacerbate healthcare inequalities during the COVID-19 pandemic and beyond.² Natural language analysis has become a major enabling breakthrough for extracting information from the literature and from big data sources, such as electronic health records, laboratory tests, public databases and others. Combined with

image analysis, there are initial prototypes and great expectations reported for tracking the COVID-19 pandemic.³ Yet, unfortunately, machine learning methodologies for producing personalised diagnostics and therapeutics are still largely fragile, unexplainable and often insufficiently reproducible.⁴ Serious medical actions cannot be algorithmically and automatically taken without review and integration with final decision-making judgments of human experts, who draw not only on their experiences in interpreting statistical data subjectively but are also required to take clinical and legal responsibility for the ethical treatment of patients.⁵ Expert professionals cannot be totally replaced by algorithmic or AI 'Chatbots', so admired for efficiency in business or entertainment IT. And even in these less ethically challenged fields, automated software rarely truly satisfies the needs of customers. An extensive review of AI machine learning methods for predictive modelling of COVID-19 infections from lung CT images concluded that a majority of models were at risk of being biased, leading to unreliable results, noting that: 'In their current reported form, none of the machine learning models included in this review are likely candidates for clinical translation for the diagnosis/prognosis of Covid-19'.⁶

The above conclusions coincide with the authors' experience in biomedical AI over many decades.⁷ Better and thoroughly tested and evaluated models are needed to explain human-machine reasoning under risk and uncertainty. Because the rapid onset of the COVID-19 pandemic required correspondingly urgent responses, most COVID-related AI tools did not undergo comprehensive evaluations, including for those for ethical use, although history has shown this to be essential for clinical systems. An urgent



© Author(s) (or their employer(s)) 2021. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

¹Department of Computer Science, Rutgers University, New Brunswick, New Jersey, USA

²Artificial Intelligence, Universidad Politecnica de Madrid, Madrid, Spain

Correspondence to
Dr Casimir Kulikowski;
kulikows@cs.rutgers.edu

undertaking to make the current predominantly data-driven AI methods (eg, deep learning) clinically usable is to develop innovative advanced cognitive models that are humanely explainable, and ethically driven knowledge-and-experience-based. The COVID-19 pandemic reinforces lessons that for AI to be effective, unbiased and reliably trustworthy for patient care in clinical epidemiological settings, novel AI approaches are urgently needed. These will have to be highly problem focused,⁸ so the best expert judgments can exploit specific clinical phenotypes from precision medicine developments to interactively, securely, and in clearly explained ways take advantage of the latest computational techniques of structured, indexed data and knowledge base design.

In summary, AI has been key in producing computational genomic analyses and techniques essential for the exceptionally rapid development of COVID-19 vaccines, but expectations that it will play a substantial role in clinically helping handle the current pandemic remain premature, largely based on inadequately tested early prototypes. Lessons learnt during the present COVID-19 pandemic will all have to be critically reviewed and completely new, human-interactive and humanely tested AI developed beyond current data-analytical insights, so the world can respond more effectively with unbiased ethical responsibility to pandemics in the future.

Contributors Each author contributed equally to the writing of this article.

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 Commissioned; externally peer reviewed.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>.

ORCID iD

Casimir Kulikowski <http://orcid.org/0000-0002-0625-1666>

REFERENCES

- 1 Kulikowski CA. Pandemics: Historically “slow learning curve” leading to biomedical informatics and vaccine breakthroughs. *Yearb Med Inform* 2021;30:290–301.
- 2 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.
- 3 Cury RC, Megyeri I, Lindsey T, *et al*. Natural language processing and machine learning for detection of respiratory illness by chest CT imaging and tracking of COVID-19 pandemic in the US. *Radiol Cardiothorac Imaging* 2021;3:e200596.
- 4 Roberts M, Driggs D, Thorpe M, *et al*. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nat Mach Intell* 2021;3:199–217.
- 5 Grote T, Berens P. On the ethics of algorithmic decision-making in healthcare. *J Med Ethics* 2020;46:205–11.
- 6 Wynants L, Van Calster B, Collins GS, *et al*. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. *BMJ* 2020;369:m1328.
- 7 Maojo V, Crespo J, García-Remesal M, *et al*. Biomedical ontologies: toward scientific debate. *Methods Inf Med* 2011;50:203–16.
- 8 Larson EJ. *The Myth of Artificial Intelligence: Why Computers Can't Think the Way We do*. Cambridge: Bellknap-Harvard University Press, 2021.

© 2021 Author(s) (or their employer(s)) 2021. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ. This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>. Notwithstanding the ProQuest Terms and Conditions, you may use this content in accordance with the terms of the License.