BMJ Health & Care Informatics

Maturity assessment of Kenya's health information system interoperability readiness

Job Nyangena,[1,2] Rohini Rajgopal ,[3] Elizabeth Adhiambo Ombech,[1] Enock Oloo,[1] Humphrey Luchetu,[1] Sam Wambugu,[4] Onesmus Kamau,[5] Charles Nzioka,[5] Samson Gwer,[1,6] Moses Ndiritu Ndirangu[1]

To cite: Nyangena J, Rajgopal R, Ombech EA, *et al*. Maturity assessment of Kenya's health information system interoperability readiness. *BMJ Health Care Inform* 2021;**28**:e100241. doi:10.1136/bmjhci-2020-100241

► Additional material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/bmjhci-2020-100241).

Received 18 September 2020
Revised 02 February 2021
Accepted 06 February 2021

[1]Research and Evidence Department, Afya Research Africa, Nairobi, Kenya
[2]Institute of Biomedical Informatics, Moi University, Eldoret, Kenya
[3]Gillings School of Global Public Health, Department of Health Policy and Management, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA
[4]ICF International, Fairfax, Virginia, USA
[5]Kenya Ministry of Health, Nairobi, Kenya
[6]School of Medicine, Kenyatta University, Nairobi, Kenya

Correspondence to
Moses Ndiritu Ndirangu;
mndiritu@afyaresearch.org

ABSTRACT

Background The use of digital technology in healthcare promises to improve quality of care and reduce costs over time. This promise will be difficult to attain without interoperability: facilitating seamless health information exchange between the deployed digital health information systems (HIS).

Objective To determine the maturity readiness of the interoperability capacity of Kenya's HIS.

Methods We used the HIS Interoperability Maturity Toolkit, developed by MEASURE Evaluation and the Health Data Collaborative's Digital Health and Interoperability Working Group. The assessment was undertaken by eHealth stakeholder representatives primarily from the Ministry of Health's Digital Health Technical Working Group. The toolkit focused on three major domains: leadership and governance, human resources and technology.

Results Most domains are at the lowest two levels of maturity: nascent or emerging. At the nascent level, HIS activities happen by chance or represent isolated, ad hoc efforts. An emerging maturity level characterises a system with defined HIS processes and structures. However, such processes are not systematically documented and lack ongoing monitoring mechanisms.

Conclusion None of the domains had a maturity level greater than level 2 (emerging). The subdomains of governance structures for HIS, defined national enterprise architecture for HIS, defined technical standards for data exchange, nationwide communication network infrastructure, and capacity for operations and maintenance of hardware attained higher maturity levels. These findings are similar to those from interoperability maturity assessments done in Ghana and Uganda.

Summary

What is already known?

- ► In Kenya and other sub-Saharan African countries, there has been a proliferation of digital health solutions implemented over the past decade aimed at improving health service delivery. However, these implementations have been found to be uncoordinated, fragmented and not integrated into a cohesive national health information network. This fragmentation has led to the duplication of effort by different implementors and the inability to scale pilots, diminishing the potential benefits of digital health interventions.

What does this paper add?

- ► This paper provides a comprehensive review of Kenya's health information system interoperability readiness and identifies priorities for intervention.

INTRODUCTION

Digital technology has transformed the global way of life over the past three decades. The healthcare space has been part of this revolution with the ubiquitous implementation of digital solutions to tackle healthcare delivery challenges.[1–3] The WHO defines digital health as an umbrella term that includes previous terms such as eHealth and mHealth as well as emerging concepts like the use of advanced computing techniques to manage big data in health, genomics and artificial intelligence.[4] Digital health has the potential to improve the safety and quality of care, reduce the skyrocketing costs of healthcare and increase the patient's participation in their own care.[5–7]

The WHO recognises that digital health presents a unique opportunity for the development and strengthening of public health systems.[8] The recent rise in the number of cell phone users and internet technologies in developing countries, coupled with a reduction in the price of devices and services, has made digital health an attractive potential solution to the challenges of a resource-constrained health system.[9] In Kenya, there has been a proliferation of digital health solutions implemented over the past decade aimed at improving health service delivery. However, these implementations have been found to be uncoordinated, fragmented and not integrated into a cohesive national health information network.[9 10] This fragmentation has led to the duplication of effort by different implementors and the lack of scaling of piloted implementations, among other issues that limit the potential benefits of digital health interventions.[11]

To realise the potential of digital health interventions, they need to be implemented in an interoperable environment. Interoperability refers to the capacity for different information systems to meaningfully exchange data. In the context of health information systems (HIS), this enables them to be implemented across organisational boundaries to effectively deliver healthcare services and advance the health status of individuals and communities.[12] Globally, there have been a few successful implementations of HIS interoperability such as in Estonia and in the state of Indiana, USA.[13 14] These examples demonstrate that the goal of HIS interoperability is achievable, and the lessons learnt from their experiences may be useful in our situation.

In Kenya, the National Government, through the Ministry of Health (MoH), has taken steps to facilitate a more conducive environment for health information exchange across different information systems. These include the development of guidance documents on digital health standards for electronic HIS, a national enterprise architecture, a master health facility list and a health worker registry, among others.[15–17] While these are significant milestones in health system interoperability, much is yet to be done. We conducted an assessment of the current state of interoperability in Kenya to determine the progress made so far and to identify gaps that need intervention.

For our assessment, we used the HIS Interoperability Maturity Toolkit by the MEASURE Evaluation project in collaboration with the Health Data Collaborative. This toolkit provides a comprehensive framework for evaluating HIS interoperability at a national level. The toolkit was extensively validated within low-income countries, including Kenya, and has been used to evaluate the HIS maturity for Ghana and Uganda.[18 19] By using it, we were sure to have a comprehensive and comparable measure for HIS maturity for Kenya. It was developed with the following objectives in mind: to identify the domains and subdomains for HIS interoperability and stages of their development toward maturity; to assess and understand where they are on the path to HIS interoperability and identify actions that can accelerate interoperability maturation; to use the results of the assessment to plan, prioritise, and coordinate resources to support a strong, responsive and sustainable national HIS; and to monitor, evaluate, and report on individual or all components of HIS interoperability.

We assessed the state of national HIS interoperability in Kenya, where studies and surveys have reported little or no interoperability among the increasing number of digital health systems and products.

METHODS

Assessment tool

We applied the MEASURE Evaluation project's HIS Interoperability Maturity Toolkit as a framework for the assessment of the HIS interoperability landscape in Kenya. We chose this toolkit as it had already been developed and validated by the MEASURE team and had been used for similar assessments in Uganda and Ghana (see online supplemental appendix 1 for the Uganda and Ghana assessments). The toolkit addresses three maturity domains: leadership and governance, human resources and technology. Each domain is divided into subdomains, making a total of 18 subdomains as summarised in table 1.

Table 1 Domains and subdomains of the interoperability maturity framework

Domain	Subdomains
Leadership and governance	1. Governance structure for HIS 2. Interoperability guidance documents 3. Compliance with data exchange standards 4. Data ethics 5. HIS interoperability monitoring and evaluation 6. Business continuity 7. Financial management 8. Finance resource mobilisation
Human resources	1. Human resources policy 2. Human resources capacity (skills and numbers) 3. Human resources capacity development
Technology	1. National HIS enterprise architecture 2. Technical standards 3. Data management 4. HIS subsystems 5. Operations and maintenance 6. Communication network: LAN and WAN 7. Hardware

HIS, health information systems; LAN, local area network; WAN, wide area network.

During an assessment, each domain and subdomain is assigned a maturity level in accordance with user guidelines for the maturity toolkit. The maturity levels are described below.

Level 1 (nascent)

The country lacks HIS capacity or does not follow processes systematically. HIS activities happen by chance or represent isolated, ad hoc efforts.

Level 2 (emerging)

The country has defined HIS structures, but they are not systematically documented. No formal or ongoing monitoring or measurement protocol exists.

Level 3 (established)

The country has documented HIS structures. The structures are functional. Metrics for performance monitoring, quality improvement and evaluation are used systematically.

Level 4 (institutionalised)

Government and stakeholders use the national HIS and follow standard practices.

Level 5 (optimised)

The government and stakeholders routinely review interoperability activities and modify them to adapt to changing conditions.

For a domain to be at a given defined maturity level, all its subdomains need to be at or above that level. The score of a domain determines its level maturity, taking the floor of the level if the score is between one level and the next. For example, a domain/subdomain that scores 3+ is judged at level 3 (established) and not level 4 (institutionalised).

For the assessment, we involved a number of Kenya's digital health stakeholders through a workshop, mostly constituting the Digital Health Technical Working Group (TWG) led by the digital health unit of the MoH and represented by different sectors: academia, research, professional bodies, non-governmental organisations and other entities (see online supplemental appendix 2 for the list and classification of participating entities). The participants were individuals and organisational representatives who had experience working within the digital health ecosystem in Kenya at local, county and national levels. These participants, by virtue of being members of the TWG, were best placed to understand the parameters within the MEASURE toolkit and respond to them appropriately. Routine users were not the target of this assessment as this assessment was for national level HIS interoperability and as such, the participants needed to have a national level outlook to be able to respond appropriately to the parameters in the assessment tool.

Participants were presented with the assessment goals, scope and process. They were divided into three groups corresponding to the three domains of HIS interoperability. The groups discussed the maturity domains and subdomains and completed the assessment questionnaire as defined by the toolkit. A consensus-building session on the results was conducted to present the findings from each group and develop a final harmonised set of answers for both the domains and subdomains.

RESULTS

A total of 25 different entities with 39 representatives were involved in the interoperability maturity assessment and discussions. There were 11 representatives from the MoH and other government agencies, 4 representatives from academia, 5 representatives from the private sector and 19 from non-governmental organisations.

Kenya's HIS interoperability maturity matrix

In this assessment, the majority of interoperability subdomains were still in the nascent stage of maturity. In the leadership and governance domain, the 'governance structure for HIS' and 'interoperability guidance documents' subdomains had the highest maturity score at established and institutionalised, respectively, while 'financial management' and 'financial resource mobilisation' subdomains were judged as emerging. The other subdomains were in the nascent stage of maturity. Overall, the human resources domain, comprised of three subdomains, was emerging in maturity. Of the seven subdomains of the technology domain, one (communication network: LAN and WAN) had institutionalised maturity; three (national HIS enterprise architecture, technical standards and HIS subsystems) were established in maturity; two (operations and maintenance, and hardware) were emerging, while data management was the least mature at nascent maturity and thus pulled the entire technology domain to its level. The assessment is summarised in table 2.

DISCUSSION

The HIS interoperability maturity model addresses the components that are critical to interoperability: technology, the broad area of leadership and governance of the HIS, and human resources. The maturity model concept is used to measure the ability of an organisation or government entity, such as a MoH, to continuously improve in a specific discipline until it reaches the desired level of development or maturity.[20] Overall, our findings reveal that the Kenya HIS (KHIS) interoperability subdomains were at the nascent or emerging stage.

While there was no subdomain that had achieved the highest maturity level, there is some progress that should be acknowledged. There is a relatively robust technological environment to support HIS activities with a defined national enterprise architecture for HIS, defined technical standards for data exchange, a nationwide communication network infrastructure and capacity for operations and maintenance of hardware. This shows a clear bias towards the technology that facilitates interoperability and neglect of the other two domains that are important for interoperability.

The leadership and governance domain has two subdomains that are well established. These are governance structure for HIS and availability of interoperability guidance documents. The governance structure for HIS subdomain includes TWGs that support the MoH in its HIS agenda. Interoperability is handled under the Digital Health TWG. The TWGs, as presently constituted, lack defined terms of reference that outline the scope of their mandate. This can potentially result in the lack of focus and difficulty in the monitoring and evaluation of the TWG activities and mandates. Such terms of reference should be reviewed regularly and align with the emerging digital health trends and the ever-increasing number of digital health stakeholders. Its deliberations should be firmly anchored in an evolving interoperability roadmap for the KHIS.

The MoH has published several documents to provide guidance on the implementation of different aspects of

Table 2 Interoperability domain maturity scores

Leadership and governance		
Subdomain	**Level**	**Comment**
Governance structure for HIS	(3+) established	Kenya's Ministry of Health has an established governance structure for the management of HIS activities. There are technical working groups (TWGs) that meet regularly, namely the HIS TWG, eHealth TWG, Monitoring and Evaluation TWG and the Central Registration of Vital Statistics TWG. Their activities are coordinated through a ministry-led, interagency coordinating committee. These working groups comprise of stakeholders from both the public and private sectors. However, a routine HIS curriculum focused on building an environment that enables policy, building a resource pipeline and creating champions does not exist.
Interoperability guidance documents	(4) institutionalised	The National Government has developed and launched guidance documents to support different aspects of digital health implementation. The Kenya eHIS interoperability standards document is specific to interoperability in the health sector and is based on and supported by other guidance documents in place: the Kenya National eHealth Policy, the Kenya National eHealth Strategy (2011–2017), the Kenya HIS Policy, the Kenya Standards and Guidelines for mHealth systems and the Kenya Health Enterprise Architecture.[15–17 21 22] In general, these documents are intended to guide implementation of HIS interoperability. Plans are underway to review the interoperability document.
Compliance with data exchange standards	(1) nascent	The Kenya eHIS interoperability standards document outlines the data exchange standards that are recommended for system interoperability.[21] Despite its existence, there are no structures, processes or procedures in place to guide or enforce compliance with the data exchange, messaging and data security standards as envisaged in the guidelines.
Data ethics	(2) emerging	This subdomain addresses the moral dimensions of data management, including the policing of adherence to ethical principles throughout data generation, recording, curation, processing, dissemination, sharing and use. No enacted general or healthcare-specific data protection laws, regulatory frameworks or ethics provisions exist to guide data ethics around security, privacy and confidentiality. While the 2018 Data Protection Bill is a good start (currently under review before parliament), it may not adequately address the unique and specific nuances of healthcare data.
HIS interoperability monitoring and evaluation	(1) nascent	This subdomain refers to the use of indicators/attributes from the maturity model to facilitate the tracking of inputs, processes and outputs against desired results of HIS interoperability implementation, and the use of these data to make decisions. The Ministry of Health has a monitoring and evaluation framework that focuses on the improvement of information systems at all levels and a stewardship goal of establishing common data architecture to ease the sharing of data.
Business continuity	(1) nascent	The interoperability maturity tool defines business continuity as the capability of an organisation to continue the delivery of products or services at acceptable predefined levels following a disruptive incident. It entails devising plans and strategies that enable an organisation to continue operations and to recover quickly from any type of disruption. There is currently no government-approved business continuity plan in place for both the national and county levels of HIS.
Financial management Financial resource mobilisation	(2+) emerging (2) emerging	Financial management includes the legal and administrative systems, and procedures that permit a government ministry, its agencies and organisations to conduct activities that adhere to procedural and appropriate use of public funds. Resource mobilisation includes the activities involved in securing new and additional financial resources for HIS management. The government has budgeted for digital health including interoperability activities. Furthermore, it was found that a significant proportion of financial resources for HIS strengthening including HIS interoperability were donor driven.
Domain total	(1) nascent	
Human resources		

Continued

Table 2 Continued

Leadership and governance		
Subdomain	**Level**	**Comment**
Human resources policy	(2) emerging	The maturity assessment did not identify the presence of a human resources policy that recognises HIS-related cadres. A national needs assessment has been completed showing the number of staff and types of skills needed to support HIS including digital HIS and interoperability. However, there is an absence of a long-term plan to grow and sustain staff with the skills needed to sustain HIS and digital HIS and interoperability. Further, HIS-related cadre roles such as health records and information officers (HRIOs) at county level are mapped to the government's workforce and schemes of work.
Human resources capacity (skills and numbers)	(2) emerging	The country does not have enough staff dedicated to maintaining digital HIS and interoperability. The HRIOs are involved in all aspects of health records and information, but not necessarily digital HIS. Furthermore, it was found that the country depends on technical assistance from external stakeholders to support the national and county digital HIS.
Human resources capacity development	(2+) emerging	Tertiary education institutions such as Moi University and Kenyatta University have started programmes to build capacity for digital health roles. However, there is no plan for or ongoing in-service training for HIS staff to build their skills around digital HIS and interoperability. Furthermore, the country does not have the capacity to train enough staff to support digital HIS and interoperability through in-country, preservice and in-service training institutions or partnerships with other training institutions.
Domain total	(2) emerging	
Technology		
National HIS enterprise architecture	(3+) established	A national enterprise architecture for an HIS defines how HIS subsystems interact and exchange data and shows necessary services for data exchange. Kenya has a validated national HIS enterprise architecture that defines technology requirements and exchange formats for interoperability.[16] There are also foundational tools and rules for HIS interoperability including health information management systems for routine and surveillance data and core authoritative registries (facility registry and health worker registry). These tools are owned and implemented by the National Government.
Technical standards	(3+) established	The technical standards provide a common language and set of expectations that enable interoperability among systems and/or devices. They include standards for data exchange, transmission, messaging, security, privacy and hardware. The National Government, through the Ministry of Health, has published and disseminated standards for data exchange. There are plans to develop a certification mechanism for new HIS subsystems to be integrated into a national HIS using the specified standards. Additionally, an interoperability laboratory, Digital Health Applied Research Centre, has been set up by a collaboration between the Jomo Kenyatta University of Agriculture and Technology and a development partner to test technical standards and new digital HIS. [23]
Data management	(1) nascent	There was no national document for data management procedures for the Kenya HIS.
HIS subsystems	(3) established	Although the standards and guidelines for digital health system interoperability are published, most digital HIS in the country consist of standalone program-specific subsystems working in silos addressing only the basic needs such as routine HIS, surveillance systems and human resource management systems. The government requires that all HIS subsystems comply with the country's interoperability plan, but this has not been effectively enforced.
Operations and maintenance	(2+) emerging	This refers to a set of procedures to ensure a high uptime for computer hardware, software and network resources. Kenya has strong in-country capacity for computer technology maintenance, but the maintenance for network and hardware is a mix of reactive and evolving preventive procedures.

Continued

Table 2 Continued

Leadership and governance		
Subdomain	**Level**	**Comment**
Communication network: LAN and WAN	(4) institutionalised	Through the National Fiber Optic Backbone network, the government has begun implementing a technical solution to ensure permanent connectivity to HIS services. [24] All national offices of the Ministry of Health have a strong and reliable network connection to access the various HIS network services.
Hardware	(2) emerging	These are the physical parts of a system of computers including desktop computers, laptop computers and servers that provide services to a user in the HIS. The country has inadequate hardware (eg, servers, computers, printers and supportive accessories) to support a national HIS.
Domain total	(1) nascent	

HIS, health information systems; LAN, local area network; WAN, wide area network.

digital health in the country. However, the policies and strategies outlined in these documents have received little to no attention. There is potential for future research to further investigate the reasons behind our findings, as this assessment was a snapshot of the state of interoperability at a particular time.

So while other domains and subdomains have received some appreciable progress in maturation, the implementation of subdomains on compliance with data exchange standards, data ethics, monitoring and evaluation, business continuity and financial resource mobilisation has been left out. This gap in policy implementation shows that a holistic approach is indispensable to the attainment of HIS interoperability.

A skilled workforce is central to any enterprise and the HIS domain is no exception. From our findings, human resource capacity has not been adequately addressed. At present, HIS are managed by health records and information officers who have little or no training in digital health. Furthermore, there are currently no plans to provide in-service training on digital health to these staff or long-term plans to grow and sustain staff with required digital health skills needed to maintain modern HIS. This means that even if the other domains are adequately addressed, there will be inadequately skilled manpower in the country to support the maturation of health interoperability. Investment in preservice and in-service national training programmes to build human resource capacity on digital HIS, including interoperability, based on a training curriculum that outlines the required competencies, can catalyse the emergence of skilled digital health practitioners.

The technology domain had four of its seven subdomains being at or above established, with the 'operations and maintenance' and 'hardware' subdomains at the emerging level. The overall domain, however, was nascent due to the nascent score of the 'data management' subdomain. The KHIS lacks a national document for data management procedures yet holds tens to hundreds of millions of data entries and generates more every month. Developing and implementing a data management document will help in the utilisation of the available data for studying patterns of ill-health to inform health policies for better health outcomes.

The findings from this assessment mirror those of similar assessments done in Ghana and Uganda where the results revealed that most subdomains are at the lowest two levels: nascent or emerging. The maturation of country level interoperability is key to regional and continental HIS interoperability.

Moving forward, the MoH and other digital health stakeholders need to continue the collaborative efforts to achieve digital health system interoperability at local, national and regional levels.

CONCLUSION

The maturity model we used provides a holistic framework that the MoH can use to implement its national HIS interoperability vision. It identifies the three domains of leadership and governance, human resources and technology that need to be developed concurrently to achieve interoperability. Our findings show that some domains are more developed than others and this may be one of the reasons that HIS interoperability has so far proven elusive.

Overall, the National Government has made significant steps towards achieving HIS interoperability. We emphasise focusing on the domain of KHIS leadership and governance that is still in the nascent stage for its importance in the coordination and the growth of the human resources and technology domains.

Twitter Rohini Rajgopal @rohinirajgopal

Acknowledgements Members of Kenya's Ministry of Health eHealth's Technical Working Group were instrumental to this assessment and were involved in workshop discussions that supported our findings.

Contributors JN, MNN, EAO and SG contributed to the design and implementation of the research. OK and CN were involved in planning and supervised the work. JN and MNN processed the collected data, performed the analysis and drafted the manuscript. JN, MNN, RR, EAO, EO, HL and SW reviewed and edited the final draft. All authors discussed the results and commented on the manuscript.

Funding The work evaluated by this study was funded through the UKAID-County Innovation Challenge Fund that aimed to transform maternal and newborn healthcare by supporting innovations that revolutionise access and utilisation. This funding was complemented with funding from Pfizer Foundation's Global Health Innovation Grant Program and Grand Challenges Canada".

Disclaimer The fund was not involved in the evaluation, writing or decision to submit this publication. No author was paid directly or indirectly by a

pharmaceutical company or other agency to write this article. The corresponding author had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Competing interests None declared.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

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

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

ORCID iD

Rohini Rajgopal http://orcid.org/0000-0002-7515-3568

REFERENCES

1 Hillestad R, Bigelow J, Bower A, *et al*. Can electronic medical record systems transform health care? potential health benefits, savings, and costs. *Health Aff* 2005;24:1103–17.
2 Devaraj S, Kohli R. Information technology payoff in the health-care industry: a longitudinal study. *J Manag Inf Syst* 2000;16:41–67.
3 Adeola O, Evans O. Digital health: ICT and health in Africa. *Actual Problems of Economics* 2019;10:66–83.
4 World Health Organization. Who guideline: recommendations on digital interventions for health system strengthening, 2019. Available: https://apps.who.int/iris/bitstream/handle/10665/311941/9789241550505-eng.pdf?ua=1
5 Blumenthal D. Launching HITECH. *N Engl J Med* 2010;362:382–5.
6 Buntin MB, Burke MF, Hoaglin MC, *et al*. The benefits of health information technology: a review of the recent literature shows predominantly positive results. *Health Aff* 2011;30:464–71.
7 Goldzweig CL, Towfigh A, Maglione M, *et al*. Costs and benefits of health information technology: new trends from the literature. *Health Aff* 2009;28:w282–93.
8 ITU-WHO. National eHealth Strategy Toolkit [Internet]. International Telecommunications Union & World Health Organization, 2012. Available: https://www.itu.int/pub/D-STR-E_HEALTH.05-2012 [Accessed 19 Nov 2018].
9 Lewis T, Synowiec C, Lagomarsino G, *et al*. E-Health in low- and middle-income countries: findings from the center for health market innovations. *Bull World Health Organ* 2012;90:332–40.
10 Njoroge M, Zurovac D, Ogara EAA, *et al*. Assessing the feasibility of eHealth and mHealth: a systematic review and analysis of initiatives implemented in Kenya. *BMC Res Notes* 2017;10:90.
11 Meurn C. Beyond "Pilotitis": Three Critical Success Factors for National Digital Health Strategies, 2017. Available: https://nextbillion.net/beyond-pilotitis-three-critical-success-factors-for-national-digital-health-strategies/ [Accessed 2 Jul 2019].
12 HIMSS. *HIMSS dictionary of health information technology terms, acronyms and organizations*. 4th Edn, 2017. https://www.himss.org/himss-dictionary-health-information-technology-terms-acronyms-and-organizations-fourth-edition
13 McDonald CJ, Overhage JM, Barnes M, *et al*. The Indiana network for patient care: a working local health information infrastructure. *Health Aff* 2005;24:1214–20.
14 Nøhr C, Parv L, Kink P, *et al*. Nationwide citizen access to their health data: analysing and comparing experiences in Denmark, Estonia and Australia. *BMC Health Serv Res* 2017;17:534.
15 Ministry of Health. Kenya National eHealth Policy 2016 - 2030, 2016. Available: https://www.medbox.org/kenya-nation-e-health-policy-2016-2030/download.pdf
16 Ministry of Health. *Kenya health enterprise architecture (KHEA*, 2016.
17 Ministry of Health. *Kenya standards and guidelines for mHealth systems*, 2017.
18 MEASURE Evaluation. *Building a strong and Interoperable health information system for Ghana*. Chapel Hill, North Carolina, USA: University of North Carolina at Chapel Hill, 2018. https://www.measureevaluation.org/resources/publications/fs-18-275/at_download/document
19 MEASURE Evaluation. *Building a strong and Interoperable digital health information system for Uganda*. Chapel Hill, North Carolina, USA: University of North Carolina at Chapel Hill, 2018. https://www.measureevaluation.org/resources/publications/fs-18-296/at_download/document
20 Carvalho JV, Rocha Álvaro, Abreu A. Maturity models of healthcare information systems and technologies: a literature review. *J Med Syst* 2016;40:131.
21 Ministry of Health. *Kenya ehealth information systems interoperability standards*. Nairobi: Ministry of Health, 2015.
22 Ministry of Health. *Kenya National eHealth Strategy 2011 - 2017*, 2011.
23 JKUAT. Standard curriculum for health records management mooted, 2018. Available: http://www.jkuat.ac.ke/colleges/cohes/standard-curriculum-health-records-management-mooted/ [Accessed 13 May 2019].
24 ICT Authority. National optic fibre backbone (NOFBI) – ICT authority. Available: http://icta.go.ke/national-optic-fibre-backbone-nofbi/ [Accessed 2 Jul 2019].

BMJ Health & Care Informatics

Enhancing trust in clinical decision support systems: a framework for developers

Caroline Jones ,[1] James Thornton ,[2] Jeremy C Wyatt [3]

To cite: Jones C, Thornton J, Wyatt JC. Enhancing trust in clinical decision support systems: a framework for developers. *BMJ Health Care Inform* 2021;**28**:e100247. doi:10.1136/bmjhci-2020-100247

Received 02 October 2020
Revised 23 December 2020
Accepted 15 January 2021

[1]Hillary Rodham Clinton School of Law, Swansea University, Swansea, Wales, UK
[2]Law School, Nottingham Trent University, Nottingham, Nottinghamshire, UK
[3]Wessex Institute, University of Southampton, Southampton, UK

Correspondence to
Dr Caroline Jones;
caroline.jones@swansea.ac.uk

INTRODUCTION

Systematic reviews show that clinical decision support systems (CDSSs) can improve the quality of clinical decisions and healthcare processes[1] and patient outcomes[2]; although caution has been expressed as to balancing the risks of using CDSSs (eg, alert fatigue) when only small or moderate improvements to patient care have been shown.[3] Yet, despite the potential benefits, studies indicate that uptake of these tools in clinical practice is generally low due to a range of factors.[4–7] The well-funded National Health Service (NHS) PRODIGY programme is an example of a carefully developed CDSS - commissioned by the Department of Health to support GPs - which failed to influence clinical practice or patient outcomes, with low uptake by clinicians in a large-scale trial.[8] A subsequent qualitative study revealed that, among other issues—such as the timing of the advice—trust was an issue: 'I don't trust … practising medicine like that … I do not want to find myself in front of a defence meeting, in front of a service tribunal, a court, defending myself on the basis of a trial of computer guidelines' [quote from GP].[9]

Another qualitative study exploring factors hindering CDSSs' uptake in hospital settings found that clinicians perceive that CDSSs 'may reduce their professional autonomy or may be used against them in the event of medical-legal controversies'.[10] Thus, CDSSs may be 'perceived as limiting, rather than supplementing, physicians' competencies, expertise and critical thinking', as opposed to a working tool to augment professional competence and encourage interdisciplinary working in healthcare settings.[10] Similarly, a recent survey carried out by the Royal College of Physicians revealed that senior physicians had serious concerns about using CDSSs in clinical practice, with trust and trustworthiness being key issues (see examples below).[11]

Trust is an important foundation for relationships between the developers of information systems and users, and is a contemporary concern for policymakers. It has, for example, been highlighted in the House of Lords Select Committee on Artificial Intelligence (AI) report[12]; the Topol Review[13]; a number of European Commission communications,[14–16] reports[17–20] and most recently a White Paper on AI[21]; and investigated in the context of knowledge systems, for example, for Wikipedia.[22] Although it is an important concept, it is not always defined; rather, its meaning may be inferred. For example, the House of Lords Select Committee used the phrase 'public trust' eight times,[12] but the core concern appeared to be about confidence over the use of patient data, rather than patient perceptions regarding the efficacy (or otherwise) of the AI in question. Such documents appear to take an implicit or one-directional approach to what is meant by 'trust'.

Notably, the Guidelines of the High-Level Expert Group on AI outline seven key requirements that might make AI systems more trustworthy[17]; whereas the White Paper focuses on fostering an 'ecosystem of trust' through the development of a clear European regulatory framework with a risk-based approach.[21] Therefore, in keeping with the drive for promoting clinical adoption of AI and CDSSs while minimising the potential risks,[13] here we apply Onora O'Neill's[23 24] multidirectional trust and trustworthiness framework[25] to explore key issues underlying clinician (doctor, nurse or therapist) trust in and the use (or non-use) of AI and CDSS tools for advising them about patient management, and the implications for CDSS developers. In doing so, we do not seek to examine particular existing CDSSs' merits and flaws in-depth, nor do we address the merits of the deployment process itself. Rather, we focus on generic issues of trust that clinicians report having about CDSSs' properties, and on improving clinician trust in the use and outputs of CDSSs that have already been deployed.

Two points merit attention at this stage. First, O'Neill's[25] framework is favoured as—in the words of Karen Jones—O'Neill 'has done more than anyone to bring into theoretical focus the practical problem that would-be trusters face: how to align their trust with trustworthiness'.[26] Second, some nuance is required when determining who or what is being trusted. For example, Annette Baier makes clear that her own account of trust supposes:

> that the trusted is always something capable of good or ill will, and it is unclear that computers or their programs, as distinct from those who designed them, have any sort of will. But my account is easily extended to firms and professional bodies, whose human office-holders are capable of minimal goodwill, as well as of disregard and lack of concern for the human persons who trust them. It could also be extended to artificial minds, and to any human products, though there I would prefer to say that talk of trusting products like chairs is either metaphorical, or is shorthand for talk of trusting those who produced them.[27]

Similarly, Joshua James Hatherley 'reserve[es] the label of 'trust' for reciprocal relations between beings with agency'.[28] Accordingly, our focus is on the application of O'Neill's[25] framework to CDSS developers as 'trusted' agents, and measures they could adopt to become more trustworthy.

O'Neill's trust and trustworthiness framework: A summary

O'Neill notes that 'trust is valuable when placed in trustworthy agents and activities, but damaging or costly when (mis)placed in untrustworthy agents and activities'.[25] She usefully disaggregates trust into three core but related elements:

1. Trust in the truth claims made by others, such as claims about a CDSS's accuracy made by its developer. These claims are empirical, since their correctness can be tested by evaluating the CDSS.[29]

   Trust in others' commitments or reliability to do what they say they will, such as clinicians trusting a developer to maintain and update their CDSS products. This is normative: we use our understanding of the world and the actors in it to judge the plausibility of a specific commitment, such as our bank honouring its commitment to send us statements.

2. Trust in others' competence or practical expertise to meet those commitments. This is again normative: we use our knowledge of the agent in whom we place our trust and our past experience of their actions to judge their competence, such as trust in our dentist's ability to extract our tooth and the 'skill and good judgement she brings to the extraction'.[25]

This approach utilises two 'directions of fit': the empirical element (1) in one direction (does the claim 'fit' the world as it is?), and the two normative elements (2-3) in another (does the action 'fit' the claim?).[25] Relatedly, O'Neill has written on the concept of 'judgement'; drawing a distinction between judgement in terms of looking at the world and assessing how it measures up (or 'fits') against certain standards (normative), versus an initial factual judgement of what a situation is, which 'has to fit the world rather than to make the world fit or live up to' a principle (empirical).[30]

In deciding whether to trust and use a CDSS, a user is similarly also making judgements about it. O'Neill's threefold framework may therefore provide a helpful way to examine the issues in this context. In the following sections we discuss how CDSS developers can use each component of this framework to increase their trustworthiness, and conclude with suggestions on how informaticians might fruitfully apply this framework more widely to understand and improve user–developer relationships. Inevitably, this theoretical approach cannot address every potential issue, but it is used here as a means of organising diverse concerns around trust issues into a coherent framework.

TRUSTING THE TRUTH CLAIMS MADE BY DEVELOPERS

CDSS developers might assume that their users are interested in the innovative machine learning or knowledge representation method used, or how many lines of code the CDSS incorporates. However, Petkus *et al*'s[11] recent survey of the views and experience of 19 senior UK physicians representing the views of a variety of specialties provides some evidence of what a body of senior clinicians expect from CDSSs, that developers can use to shape their truth claims and build clinical trust. While this is not generalisable/representative of all clinicians it does provide a useful illustration of clinical concerns, and our intent is to demonstrate how applying O'Neill's trust/trustworthiness framework might help our understanding of how to mitigate these issues. Table 1 shows the six clinical concerns about CDSSs which scored highest in the analysis. The score combines both the participant-rated

Table 1 Concerns about CDSS quality in Petkus *et al* survey

Concerns about CDSS quality	Score
The accuracy of advice may be insufficient for clinical benefit	15.5
How extensively was clinical effectiveness of CDSS tested	15
Whether CDSSs are based on the latest evidence	14.5
CDSSs can interrupt clinical workflow or disrupt consultations	14.5
CDSSs can ignore patient preferences	12.5
Whether the CDSS output is worded clearly	11.5

CDSS, clinical decision support systems.

severity of the concern and its frequency in the responses; the maximum score on this scale was 19:[11]

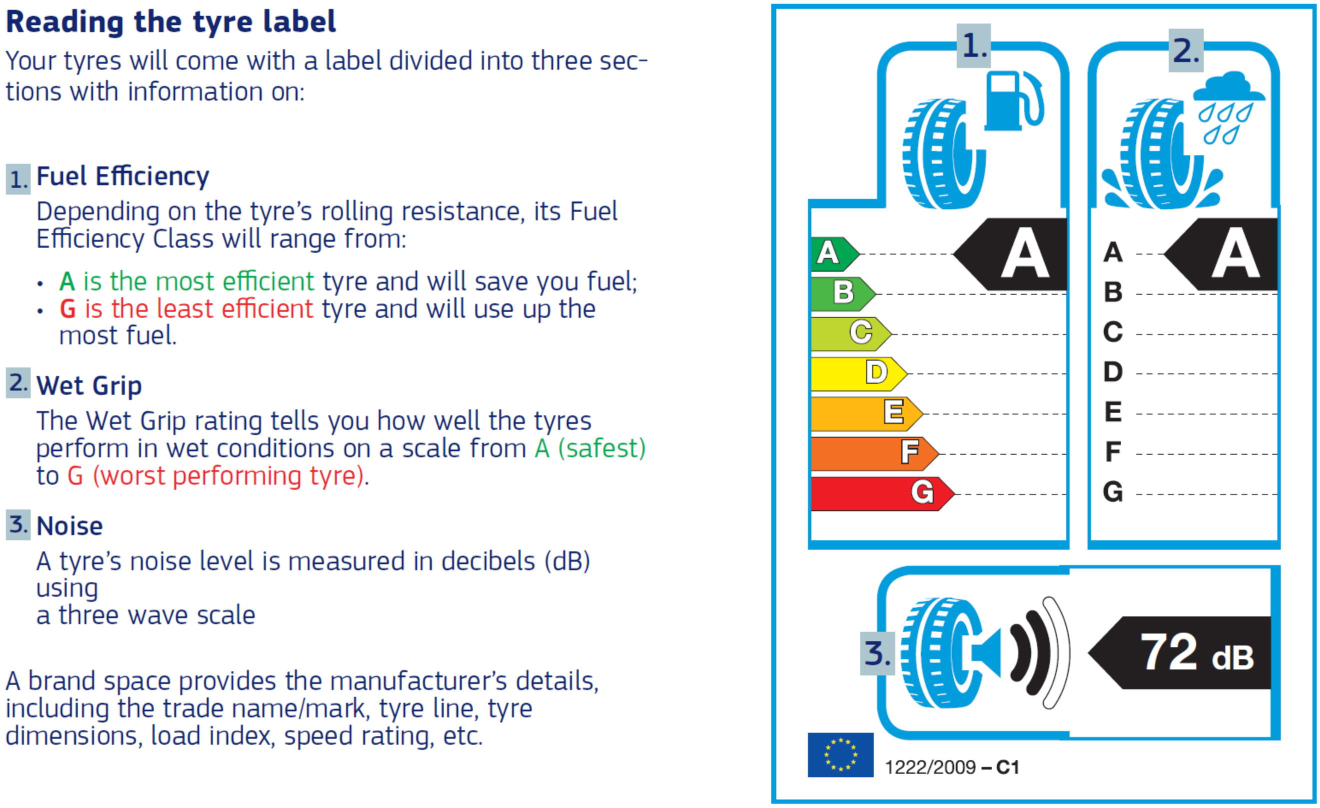
The greatest concerns here relate to O'Neill's concept (or direction of fit) of empirical trust. Whether the advice provided by a CDSS is correct, has strong evidence for its clinical effectiveness from testing etc. ultimately concerns whether its advice 'fits' or matches (eg, in diagnosis) the patient's actual condition. Can it (and/or the people that designed/made it) be trusted in an empirical sense of being factually correct?

What kind of truth claims may appeal to clinicians?

Drawing on the evidence in table 1, developers should report to clinicians: the accuracy of the advice or risk estimates; CDSS effectiveness (impact on patients, decisions and the NHS); whether the CDSS content matches current best evidence (see 'Guidelines: codes and standards frameworks' below); its usability and ease of use in clinical settings; whether its output is worded clearly, and if takes account of patient preferences. These claims should be phrased in professional language, avoiding the extravagant claims about AI often seen in the press.[31 32] Instead of different developers adopting a range of metrics for reporting study results there is a need for a standard CDSS performance reporting 'label' for these assessments, to help clinicians identify, compare and judge the empirical claims being made about competing CDSSs. This is by analogy with European Union (EU) consumer regulations dictating how, for example, tyre manufacturers report on road noise, braking performance and fuel economy for their tyres (figure 1),[33] and EU plans for a health app label.

Ensuring that the truth claims can be verified

First, CDSS developers should be aware of the 'evidence-based medicine' culture,[34] reflected in the top three concerns in table 1. This means that, before clinicians make decisions such as how to treat a patient or which CDSS to use, they look for well designed, carefully conducted empirical studies in typical clinical settings using widely accepted outcomes that answer well-structured questions. This entails a 'critical appraisal' process to identify and reject studies that are badly designed or conducted, or from settings or with patients that do not resemble those where the CDSS will be used.[34] So, it has long been established that empirical evaluation and the evidence it generates are crucial to generating trust.[29] However, a systematic review of empirical research has shown that, when CDSS developers themselves carried out the study, they were three times as likely to generate positive results as when an independent evaluator did so.[35] Therefore, studies that establish these truth claims should be carried out by independent persons or bodies. To counter suspicions of bias or selective reporting, the full study protocol and results should be made openly available, for example, by publication.[36 37] Again, there is



Reading the tyre label

Your tyres will come with a label divided into three sections with information on:

1. Fuel Efficiency

Depending on the tyre's rolling resistance, its Fuel Efficiency Class will range from:

- A is the most efficient tyre and will save you fuel;
- G is the least efficient tyre and will use up the most fuel.

2. Wet Grip

The Wet Grip rating tells you how well the tyres perform in wet conditions on a scale from A (safest) to G (worst performing tyre).

3. Noise

A tyre's noise level is measured in decibels (dB) using
a three wave scale

A brand space provides the manufacturer's details, including the trade name/mark, tyre line, tyre dimensions, load index, speed rating, etc.

Figure 1 Example of an EU tyre label and how to interpret it.[38]

Table 2 Concerns about professional practice, ethics and liability in Petkus *et al* survey

Concerns about professional practice, ethics and liability	Score
The legal liability of doctors who rely on CDSS advice is unclear	17.5
Some CDSSs act like a 'black box', with no insight possible for the user about how they arrived at their advice or conclusions	15
Doctors may follow incorrect CDSS advice, even if they would make correct decisions without it	13.5
CDSSs can embed unconscious bias, with some patient groups receiving unfair care as a result	12.5

CDSS, clinical decision support systems.

an opportunity to establish standard methods for carrying out performance or impact studies, so that clinicians can trust and compare study results on different CDSSs from different suppliers—as exemplified by EU tyre performance testing standards.[33 38]

Concerns that software developers raise about evaluating CDSS are that these studies are expensive and can take a lot of time,[37] so yield results that can be obsolete by the time they are available. However, choosing the right designs such as MOST (multiphase optimization strategy), SMART (sequential multiple assignment randomized trial), or A/B testing (randomized control experiment to compare two versions, A and B)[39] means that studies can be carried out rapidly and at low cost. Further, if the study not only meets the requirements of the EU Regulation on Medical Devices (see 'UK and EU Regulation on Medical Devices' below), but also provides strong foundations for clinical trust in the CDSS developers, then commissioned independent studies can show a very positive return on investment and could be justified as part of a CDSS's product marketing strategy.

TRUSTING OTHERS' COMMITMENTS

O'Neill asks whether we can trust what others say they will do.[23] Petkus *et al*'s[11] survey also asked clinicians about professional practice, ethics and liability matters, as table 2 shows:

The last two items in table 2 relate to issues of empirical trust (is the advice factually correct?), which can be addressed by following the suggestions in the section on 'Trusting the truth claims made by developers' above. However, the first two concerns (and those found by Liberati *et al*[10]) address not only whether the CDSS provides correct advice, but also whether it does what it claims to do. Clinicians are unable to evaluate concerns about a 'black-box' CDSS because they will likely have no idea about how answers have been arrived at: it demands faith from clinicians that the trust commitments will be met. Rather than a useful support to their practice, such a CDSS may be considered a hindrance to the exercise of clinicians' judgement and critical thinking—as in the trial of PRODIGY (a clinical decision support tool commissioned by the Department of Health to help GPs).[8] There are related concerns about legal liability. What if the clinician relies on the CDSS and this causes harm to a patient? The clinician must trust that a 'black-box' CDSS will do what it is supposed to, and not cause harm for which they may be held legally responsible. Harm could obviously be caused by the CDSS if it is not working as the developers intended (eg, due to software issues). However, even without such issues, if the CDSS utilises a deep learning method such as neural networks, the clinician still has to trust that the mechanism through which conclusions have been derived is sensible, and has only taken into account clinically relevant details, ignoring spurious information such as the patient's name or the presence of a ruler in images of a suspicious skin lesion.[40]

In terms of potential legal liability, the situation does indeed appear to be unclear. Searches we carried out in legal databases (Lexis Library, Westlaw, BAILII), and PubMed, for terms around CDSS (adviser, expert system, risk score, algorithm, flowchart, automated tool, etc) turned up blank; nor have other researchers been able to locate published decisions in the UK, Europe or USA.[41] However, it is well established that clinicians are legally responsible for the medical advice and treatment given to their patients, irrespective of the use of a CDSS.[42] They must still reach the standard of the reasonable clinician in the circumstances. This makes it all the more important, if clinical uptake is to be improved, that clinicians have reasons to trust the CDSS developers and in turn their products/services.[43]

How can CDSS developers facilitate this trust?

While developers cannot fix an uncertain legal framework, there are several steps they can take to nurture trust in this area. Most obviously, to ensure that the way the CDSS works and comes to its conclusions are made as clear as possible to users. It may not be realistic to do so completely, particularly as CDSS software becomes more complex via machine learning.[44] However, giving—where possible—some account of the mechanism for how decisions are arrived at; the quality, size and source of any data-sets relied on; and assurance that standard guidelines for training the algorithm were followed (as well as monitoring appropriate learning diagnostics) will probably assuage some clinicians' concerns.[44]

In addition, even if some 'black-box' elements are unavoidable, clinicians' anxieties regarding the dependability or commitment aspects of O'Neill's[23] trust framework may be alleviated by ensuring that frequent updates, fixes and support are all available. This should help clinicians feel more confident that the CDSS is likely to be reliable, and gives them something concrete to point to later to evidence their diligence and reasonableness, for example if they appear in court or at a professional conduct hearing.[9–11]

Box 1 Developer actions that suggest competence and commitment to producing high quality clinical decision support systems (CDSSs)

- Recruit and retain a good development team with the right skills.[58]
- Use the right set of programming tools and safety-critical software engineering processes and methods, for example, HAZOP (Hazard and Operability Analysis) to understand and limit the risks of CDSSs.[17 60]
- Carry out detailed user research for example, user-centred design workshops; establish an online user community and monitor it for useful insights; or form a multidisciplinary steering group of key stakeholders.[13 61]
- Obtain the best quality, unbiased data to train the algorithm; use the right training method and diagnostics to monitor the learning process.[46]
- Implement relevant technical standards, obtain a CE mark (Conformitè Europëenne: the EU's mandatory conformity mark by which manufacturers declare that their products comply with the legal requirements regulating goods sold in the European Economic Area) for their CDSS as a medical device.[45]
- Publish an open interface to their software; carry out interoperability testing.[62 63]
- Build on a prior track record of similar products that appeared safe.[58]
- Follow relevant codes of practice for artificial intelligence and data-based technologies.[46 47]
- Implement continuing quality improvement methods, for example, log and respond to user comments and concerns[60]; deliver updates to the CDSS regularly[61]; seek to become certified as ISO 9000 compliant.

TRUSTING OTHERS' COMPETENCE

O'Neill[23] suggests that we ask whether others' actions meet, or will meet, the relevant standards or norms of competence. Factors that may impact positively on improving clinician trust include, but are not limited to, those listed in box 1.

In this section, we focus on the technical standards,[45] and current codes of practice and development standards frameworks potentially applicable to CDSSs.[46 47] Much more could be said about these approval processes than space permits here. However, the point is not to analyse the merits of the approval processes, but to illustrate how O'Neill's framework helps to highlight their additional importance (beyond being strictly required) as a way to enhance (normative) trust.

UK and EU Regulation on Medical Devices

The initial question is whether CDSSs are medical devices? Classification as a medical device means that a CDSS will be subject to the EU Regulation on Medical Devices.[45] The European Medicines Agency (the agency responsible for the evaluation and safety monitoring of medicines in the EU) states that 'medical devices are products or equipment intended generally for a medical use'.[48] Article 1 stipulates that 'medical devices', manufactured for use in human beings for the purpose of, inter alia, diagnosis, prevention, monitoring, treatment or alleviation of disease, means: 'any instrument, apparatus, appliance, software, material or other article, whether used alone or in combination, including the software intended by its manufacturer to be used specifically for diagnostic and/or therapeutic purposes and necessary for its proper application'.[45]

In the UK, the Medicines and Healthcare Products Regulatory Agency (MHRA) has indicated that a CDSS is 'usually considered a medical device when it applies automated reasoning such as a simple calculation, an algorithm or a more complex series of calculations. For example, dose calculations, symptom tracking, clinicians (sic) guides to help when making decisions in healthcare'.[49] Hence, although some CDSSs may fall outside this definition (eg, by providing information only), our analysis is directed at those that do fall within the meaning of medical devices.

Accordingly, developers must adhere to the requirements of the EU Medical Devices Regulation[45] and post-Brexit, under domestic legislation, namely the Medicines and Medical Devices Act 2021.[50] These requirements include passing a conformity assessment carried out by an EU-recognised notified body (for medical devices for sale in both Northern Ireland and the EU), or a UK approved body (for products sold in England, Wales and Scotland)[51] to confirm that the CDSS meets the essential requirements (the precise assessment route depends on the classification of the device).[52] The focus of this testing is safety. Following confirmation that the device meets the essential requirements, a declaration of conformity must be made and a CE mark must be visibly applied to the device prior to it being placed on the market[53] (from 1 January 2021 the UKCA (UK Conformity Assessed) mark has been available for use in England, Wales and Scotland,[54] and the UKNI (UK Northern Ireland) mark for use in Northern Ireland).[55] The general obligations of manufacturers are provided in Article 10 of the EU Medical Devices Regulation, including risk management, clinical evaluation, postmarket surveillance and processes for reporting and addressing serious incidents[45]; see also the 'yellow card' scheme operated by the MHRA which allows clinicians or members of the public to report issues with medical devices.[56] Clinical users will rightly mistrust any CDSS developer who is unaware of these regulations, or fails to follow them carefully.

Nevertheless, NHSX (the organisation tasked with setting the overall strategy for digital transformation in the NHS) is seeking to 'streamline' the assurance process of digital health technologies.[57] Similarly, in the USA, the Food and Drug Administration (FDA) is piloting an approach where developers demonstrating 'a culture of quality and organisational excellence based on objective criteria' could be precertified.[58] Such 'trusted' developers could then benefit from less onerous FDA approval processes for their future products due to their demonstrable competence.[25]

Guidelines: codes and standards frameworks

In addition to the generic Technology Code of Practice[59] which should inform developers' practices, there are two sets of guidance specifically focused on the development and use of digital health tools, including data-derived AI tools for patient management – one issued by the Department of Health and Social Care (DHSC) and NHS

England,[46] and the second by the National Institute for Health and Care Excellence (NICE).[47]

The DHSC and NHS England code of conduct aims to complement existing frameworks, including the EU Regulation and CE mark process, to 'help to create a trusted environment',[46] supporting innovation that is safe, evidence based, ethical, legal, transparent and accountable. It refers to the 'Evidence standards framework for digital health technologies' developed by NICE in conjunction with NHS England, NHS Digital, Public Health England, MedCity and others.[47] The aim of this standards framework is to facilitate better understanding by developers (and others) as to what 'good levels of evidence for digital healthcare technologies look like',[47] and is applicable to technologies using AI with fixed algorithms; whereas those using adaptive algorithms are instead governed by the DHSC code (see Principle 7).[46]

Visible and/or certified compliance with these codes and standards would provide developers with normative objective standards to meet, and point clinical users of CDSSs to evidence of their competence.[25] Having confidence in the professionalism of the developers should go some way towards reassuring clinicians as to the safety, accuracy and efficacy of CDSSs, thus potentially fostering greater uptake in practice.

CONCLUSION

O'Neill's[25] approach to trust and trustworthiness, focusing on empirical trust in developers' truth claims and normative trust in their commitment and competence to meet those claims, has proved a useful framework to analyse and identify ways that developers can improve user trust in them, and in turn—it is suggested—the CDSSs they produce. That is, of course, not to suggest that developers are necessarily at fault in any way. It may be that they are unfairly distrusted by (potential) users. We suggest the application of O'Neill's framework has helped to identify ways to facilitate and enhance trust in developers, and by extension, their CDSSs.

In summary, developers should:

- Make relevant claims about system content, performance and impact framed in professional language, preferably structured to a standard that allows clinicians to compare claims about competing CDSSs. These claims need to be supported by well-designed empirical studies, conducted by independent evaluators.
- Minimise 'black box' elements, ensure that internal mechanisms are—so far as possible—explained to users, and that CDSS software comes with a comprehensive update and support package. This could help clinicians gain a sense of control over the CDSS, and thus perceive the technology as a valuable working tool that complements their own skills and expertise.
- Comply with all relevant legal and regulatory (codes and standards) frameworks. Having confidence in the professionalism and competence of the developers should go some way towards reassuring clinicians as to the safety, accuracy and efficacy of CDSSs, thus potentially fosteing greater uptake in their use.

The benefit of applying O'Neill's[23] framework is that it requires us to consider issues associated with different facets of both trust and trustworthiness, maximising the possibilities for enhancing trust and trustworthiness once such concerns or objections are overcome. An implicit or one-directional understanding of trust might result in a narrower conclusion, focused on just one element of O'Neill's framework.[25] For example, an understanding solely based on normative competence might focus on the importance of complying with the regulations (not only to avoid sanctions, but to enhance trust); this is important, but O'Neill's framework demands consideration of different, equally useful, elements of trustworthiness.

This analysis is focused on clinician use of decision support tools, but we believe that a similar analysis would generate useful insights had we looked at other users and information systems, such as the public use of risk assessment apps, or professional use of electronic referral or order communication system advisory tools. The principles of examining the empirical truth claims of the software and the evidence on which they are based, then the competence of the supplier to match these claims and their commitment to do so, seems to generate useful insights no matter who the users are or what digital service is being trusted. Thus, we suggest that O'Neill's[25] framework is considered by health and care informaticians—both those developing and evaluating digital services—as a useful tool to help them explore and expand user trust in these products and services.

Contributors All authors designed and cowrote the draft of the paper. CJ took primary responsibility for wider materials on trust and the section on 'Trusting others' competence'. JT carried out the literature review and led on the section on 'Trusting others' commitments'. JCW contributed points about the evidence base for CDSSs, as well as designing the survey which partly stimulated this work, and led on the section on 'Trusting the truth claims made by developers'. All authors critically reviewed and edited the final draft.

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 JW is receiving consultancy payments from NHSX AI Lab for advising on the validation and evaluation of AI systems for the NHS, but the views shared in this article are his own. The authors declare no other competing interest that might be relevant to the views expressed in this article.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; internally peer reviewed.

ORCID iDs

Caroline Jones http://orcid.org/0000-0001-7632-9468
James Thornton http://orcid.org/0000-0001-7847-5696
Jeremy C Wyatt http://orcid.org/0000-0001-7008-1473

REFERENCES

1 Roshanov PS, Fernandes N, Wilczynski JM, *et al*. Features of effective computerised clinical decision support systems: meta-regression of 162 randomised trials. *BMJ* 2013;346:f657.
2 Varghese J, Kleine M, Gessner SI, *et al*. Effects of computerized decision support system implementations on patient outcomes in inpatient care: a systematic review. *J Am Med Inform Assoc* 2018;25:593–602.
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 Moxey A, Robertson J, Newby D, *et al*. Computerized clinical decision support for prescribing: provision does not guarantee uptake. *J Am Med Inform Assoc* 2010;17:25–33.
5 Kortteisto T, Komulainen J, Mäkelä M, *et al*. Clinical decision support must be useful, functional is not enough: a qualitative study of computer-based clinical decision support in primary care. *BMC Health Serv Res* 2012;12:349.
6 Patterson ES, Doebbeling BN, Fung CH, *et al*. Identifying barriers to the effective use of clinical reminders: bootstrapping multiple methods. *J Biomed Inform* 2005;38:189–99.
7 Pope C, Halford S, Turnbull J, *et al*. Using computer decision support systems in NHS emergency and urgent care: ethnographic study using normalisation process theory. *BMC Health Serv Res* 2013;13:111.
8 Eccles M, McColl E, Steen N, *et al*. Effect of computerised evidence based guidelines on management of asthma and angina in adults in primary care: cluster randomised controlled trial. *BMJ* 2002;325:941.
9 Rousseau N, McColl E, Newton J, *et al*. Practice based, longitudinal, qualitative interview study of computerised evidence based guidelines in primary care. *BMJ* 2003;326:314.
10 Liberati EG, Ruggiero F, Galuppo L, *et al*. What hinders the uptake of computerized decision support systems in hospitals? A qualitative study and framework for implementation. *Implement Sci* 2017;12:113.
11 Petkus H, Hoogewerf J, Wyatt JC. AI in the NHS– are physicians ready? A survey of the use of AI & decision support by specialist societies, and their concerns. *Clinical Medicine* 2020;20:324–8.
12 House of Lords Select Committee on AI. *AI in the UK: ready, willing and able?* London: UK Parliament, 2018.
13 The Topol Review: preparing the healthcare workforce to deliver the digital future. NHS 2019.
14 Commission. Artificial Intelligence for Europe. COM (2018) 237 final.
15 Commission. Liability for emerging digital technologies. SWD(2018) 137 final.
16 Commission. Building Trust in Human-Centric Artificial Intelligence. COM (2019) 168 final.
17 Independent High Level Expert Group on AI. Ethics Guidelines for Trustworthy AI. *European Commission* 2019.
18 Independent High Level Expert Group on AI. Policy and Investment Recommendations for Trustworthy AI. *European Commission* 2019.
19 Expert Group on Liability and New Technologies - New Technologies Formation, Liability for Artificial Intelligence and other emerging digital technologies. EU 2019.
20 Commission. 'Report on the safety and liability implications of Artificial Intelligence, the Internet of Things and robotics' COM (2020) 64 final.
21 European Commission. 'White Paper on AI: A European approach to excellence and trust' COM (2020) 65 final.
22 Adams CE, Montgomery AA, Aburrow T, *et al*. Adding evidence of the effects of treatments into relevant Wikipedia Pages: a randomised trial. *BMJ Open* 2020;10:e033655.
23 O'Neill O. *A question of trust*. Cambridge University Press, 2002.
24 O'Neill O. *Autonomy and trust in bioethics*. Cambridge University Press, 2002.
25 O'Neill O. Linking trust to Trustworthiness. *International Journal of Philosophical Studies* 2018;26:293–300.
26 Jones K. Chapter 11, at 186. In: Archard D, ed. *Distrusting the trustworthy. in reading Onora O'Neill*. Taylor & Francis Group, 2013.
27 Baier A. Chapter 10, at 178. In: Archard D, ed. *What is trust?. in reading Onora O'Neill*. Taylor & Francis Group, 2013.
28 Hatherley JJ. Limits of trust in medical AI. *J Med Ethics* 2020;46:478–81.
29 Wyatt J, Spiegelhalter D. Evaluating medical expert systems: what to test and how? *Med Inform* 1990;15:205–17.
30 O'Neill O. Experts, practitioners, and practical judgement. *J Moral Philos* 2007;4:154–66.
31 Sample I. "It's going to create a revolution": how AI is transforming the NHS. The Guardian, 2018. Available: https://www.theguardian.com/technology/2018/jul/04/its-going-create-revolution-how-ai-transforming-nhs
32 Copestake J. Babylon claims its chatbot beats GPs at medical exam. BBC, 2018. Available: https://www.bbc.co.uk/news/technology-44635134
33 Thimbleby H. *Fix IT: Stories from Healthcare IT*. Oxford: Oxford University Press, 2020.
34 Sackett DL, Rosenberg WM, Gray JA, *et al*. Evidence based medicine: what it is and what it isn't. *BMJ* 1996;312:71–2.
35 Garg AX, Adhikari NKJ, McDonald H, *et al*. Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review. *JAMA* 2005;293:1223–38.
36 Liu X, Cruz Rivera S, Moher D, *et al*. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat Med* 2020;26:1364–74.
37 Liu JLY, Wyatt JC. The case for randomized controlled trials to assess the impact of clinical information systems. *J Am Med Inform Assoc* 2011;18:173–80.
38 . Available: https://ec.europa.eu/info/energy-climate-change-environment/standards-tools-and-labels/products-labelling-rules-and-requirements/energy-label-and-ecodesign/energy-efficient-products/tyres_en
39 Murray E, Hekler EB, Andersson G, *et al*. Evaluating digital health interventions: key questions and approaches. *Am J Prev Med* 2016;51:843–51.
40 Narla A, Kuprel B, Sarin K, *et al*. Automated classification of skin lesions: from Pixels to practice. *J Invest Dermatol* 2018;138:2108–10.
41 Fox J, Thomson R. Clinical decision support systems: a discussion of quality, safety and legal liability issues. *Proc AMIA Symp* 2002:1–7.
42 Brahams D, Wyatt J. Decision AIDS and the law. *Lancet* 1989;2:632–4.
43 Cohen IG, Cops GH. Docs, and code: a dialogue between big data in health care and predictive policing. *UC Davis Law Review* 2017;51:437–74.
44 Hart A, Wyatt J. Evaluating black-boxes as medical decision AIDS: issues arising from a study of neural networks. *Med Inform* 1990;15:229–36.
45 EU Regulation on Medical Devices 2017/745.
46 UK Government Department of Health and Social Care. Code of conduct of AI and other data driven technologies. London, 2019. Available: https://www.gov.uk/government/publications/code-of-conduct-for-data-driven-health-and-care-technology
47 National Institute for Health and Care Excellence. Evidence standards framework for digital health technologies, 2019. Available: https://www.nice.org.uk/about/what-we-do/our-programmes/evidence-standards-framework-for-digital-health-technologies
48 European Medicines Agency. Medical devices, 2019. Available: https://www.ema.europa.eu/en/human-regulatory/overview/medical-devices
49 Medicines and Healthcare Products Regulatory Agency. Guidance: medical device stand-alone software including apps (including IVDMDs), 2018. Available: https://www.gov.uk/government/publications/medical-devices-software-applications-apps
50 . Available: https://www.legislation.gov.uk/ukpga/2021/3/contents/enacted/data.htm
51 Medicines and Healthcare Products Regulatory Agency. Medical devices UK Approved bodies, 2021. Available: https://www.gov.uk/government/publications/medical-devices-uk-approved-bodies/
52 European Commission. Guidance document - Classification of Medical Devices - MEDDEV 2.4/1 rev.9, 2015. Available: http://ec.europa.eu/DocsRoom/documents/10337/attachments/1/translations
53 UK Government Department for Business, Energy & Industrial Strategy. Guidance: CE marking, 2012. Available:https://www.gov.uk/guidance/ce-marking
54 MHRA. Medical devices: conformity assessment and the UKCA mark. Available: https://www.gov.uk/guidance/medical-devices-conformity-assessment-and-the-ukca-mark
55 Guidance using the UKNI marking, 2021Department for Business, Energy and Industrial Strategy. Available: https://www.gov.uk/guidance/using-the-ukni-marking
56 Yellow Card. Medicines and Healthcare Products Regulatory Agency, 2020. Available: https://yellowcard.mhra.gov.uk/
57 Joshi I, Joyce R. NHSX is streamlining the assurance of digital health technologies, 2019. Available: https://healthtech.blog.gov.uk/2019/11/01/nhsx-is-streamlining-the-assurance-of-digital-health-technologies/
58 FDA. Digital health software Precertification (Pre-Cert) program. from, 2020. Available: https://www.fda.gov/medical-devices/digital-health-

center-excellence/digital-health-software-precertification-pre-cert-program

59 UK Government Digital Service. Technology code of practice, 2019. Available: https://www.gov.uk/government/publications/technology-code-of-practice

60 NHS Digital Clinical Safety team. DCB0129: clinical risk management: its application in the manufacture of health it systems and DCB0160: clinical risk management: its application in the deployment and use of health it systems. from, 2018. Available: https://digital.nhs.uk/data-and-information/information-standards/information-standards-and-data-collections-including-extractions/publications-and-notifications/standards-and-collections/dcb0160-clinical-risk-management-its-application-in-the-deployment-and-use-of-health-it-systems

61 Mahadevaiah G, RV P, Bermejo I, *et al*. Artificial intelligence-based clinical decision support in modern medical physics: selection, acceptance, commissioning, and quality assurance. *Med Phys* 2020;47:8.

62 NHS Digital. Interoperability toolkit. Available: https://digital.nhs.uk/services/interoperability-toolkit

63 Walsh K, Wroe C. Mobilising computable biomedical knowledge: challenges for clinical decision support from a medical knowledge provider. *BMJ Health Care Inform* 2020;27:e100121.

BMJ Health & Care Informatics

Designing COVID-19 mortality predictions to advance clinical outcomes: Evidence from the Department of Veterans Affairs

Christos A Makridis ,[1,2] Tim Strebel,[3] Vincent Marconi,[4] Gil Alterovitz [1,5]

To cite: Makridis CA, Strebel T, Marconi V, *et al*. Designing COVID-19 mortality predictions to advance clinical outcomes: Evidence from the Department of Veterans Affairs. *BMJ Health Care Inform* 2021;**28**:e100312. doi:10.1136/bmjhci-2020-100312

Received 23 December 2020
Revised 17 March 2021
Accepted 31 March 2021

[1]National Artificial Intelligence Institute at the Department of Veterans Affairs, US Department of Veterans Affairs, Washington, District of Columbia, USA
[2]Digital Economy Lab, Stanford University, Stanford University, Stanford, California, USA
[3]Washington D.C. VA Medical Center, Department of Veterans Affairs, Washington, District of Columbia, USA
[4]Rollins School of Public Health, Emory University, Atlanta, Georgia, USA
[5]Harvard Medical School, Boston, Massachusetts, USA

Correspondence to
Dr Christos A Makridis;
christos.makridis@va.gov

ABSTRACT

Using administrative data on all Veterans who enter Department of Veterans Affairs (VA) medical centres throughout the USA, this paper uses artificial intelligence (AI) to predict mortality rates for patients with COVID-19 between March and August 2020. First, using comprehensive data on over 10 000 Veterans' medical history, demographics and lab results, we estimate five AI models. Our XGBoost model performs the best, producing an area under the receive operator characteristics curve (AUROC) and area under the precision-recall curve of 0.87 and 0.41, respectively. We show how focusing on the performance of the AUROC alone can lead to unreliable models. Second, through a unique collaboration with the Washington D.C. VA medical centre, we develop a dashboard that incorporates these risk factors and the contributing sources of risk, which we deploy across local VA medical centres throughout the country. Our results provide a concrete example of how AI recommendations can be made explainable and practical for clinicians and their interactions with patients.

Summary box

- We build a model using artificial intelligence (AI) and machine learning (ML) techniques to predict mortality among all Veterans that have been in the Department of Veterans local medical centres between March and August 2020.
- Our preferred model achieves a 0.87 area under the the receiver operator characteristics curve and an area under the precision-recall curve of 0.41.
- In addition to age, our model reveals that an individual's labs and vitals are significant predictors of mortality, followed by medical history.
- We pilot our predictive model by creating a platform for clinicians across local VA centres that produces individual-specific risk scores for their patients, thereby allowing clinicians to offer more tailored treatment plans for patients.
- Our paper suggests that artificial intelligence has the potential to substantially improve clinical experiences and patient outcomes, but the artificial intelligence-driven results must be accessible, interpretable and actionable.

INTRODUCTION

The recent COVID-19 pandemic represents the largest global shock to health and economic systems in at least a century, leading to significant declines in economic activity,[1 2] mortality[3] and well-being.[4] These patterns and the resulting aftershock have led to a surge in research activity to generate risk profiles to understand how individuals and communities might be heterogeneously exposed to the virus.[5 6] However, researchers have struggled to obtain bias-free, reliable, and externally-valid predictions on representative datasets.[7]

The primary contribution of this paper is to develop a reliable predictive model for understanding mortality rates among Veterans and to take these predictions to practice by creating an accessible and informative dashboard that clinicians can use to improve their treatment of patients. Motivated by an increasing recognition that socio-economic factors are important for understanding health and well-being[8–10] and race,[11] we draw on administrative data from the Department of Veterans Affairs (VA) and estimate a series of artificial intelligence (AI) models that incorporate medical history, demographics, and lab results for over 10 000 Veterans. Others have emphasised the role of other comorbidities, like asthma, as risk factors for COVID-19,[12] but none have integrated all these factors together, particularly in a representative sample or full population.

We obtain an area under the receive operator characteristics curve (AUROC) and area under the precision-recall curve (AUPRC) of 0.87 and 0.41, as well as F1 and recall scores of 0.40 and 0.76. We decompose the contribution of each feature, identifying a handful of vital signs and lab indicators that matter even more than age in predicting mortality.

While age alone helps obtain 'reasonable' AUROC scores, we show that these results are an artefact of the nature of an imbalanced dataset where mortality rates are low. Furthermore, we find that models with age alone produce high AUROC scores, but low AUPRC scores. The inclusion of chronic and acute medical conditions helps, but the F1 and recall scores do not rise to much until we introduce vital and lab indicators. Through a unique partnership with the Washington D.C. VA medical centre, we subsequently create a dashboard that uses our preferred predictive model to provide clinicians with personal risk scores for each patient and the leading indicators that are driving the score. Importantly, these risk scores enumerate the primary contributing factors so that clinicians are provided with not only actionable information, but also context over the logic behind the score. We are piloting the dashboard and making it available across local VA medical centres, which is a general contribution that extends even beyond the Veterans context.

Our paper contributes to a timely research agenda on the effects of COVID-19 and the identification of individuals who are more exposed to it than others. For example, age has emerged as one of the most important comorbidities.[13 14] However, we show that age alone does a poor job in producing robust predictions. Because COVID-19 mortality rates are low to begin with, and most datasets are fairly imbalanced, it is easy to obtain a reasonable AUROC with a weak predictive model simply by producing many true negatives. Moreover, we show that there is a lot of heterogeneity even within age brackets, which could be a function of social capital within the local community or other preventative health measures.[15]

We also join a broader literature that embeds AI into tools for clinicians, including predictive tools for viral pneumonia and even secure analytics platforms, as in the case of OpenSAFELY that covers over 17 million adults in the UK to estimate hazard models as a function of comorbidities and other demographic characteristics.[16 12] The VA has been a pioneer in creating COVID-19 models. For example, Osborne *et al*[17] construct a care assessment need (CAN) score that is correlated with COVID-19 outcomes, showing that patients with a higher CAN also had a higher risk of COVID-19 infection and death. Similarly, King *et al*[18] estimate the probability of mortality as a function of demographic and medical characteristics. We use AI to estimate the risk factors and optimizing for multiple performance metrics. We also include variables from operational services that are typically available to clinicians. In addition, we create a dashboard to facilitate trustworthy AI by making the risk factor easily accessible and interpretable for clinicians, among others, consistent with the recent principles around trustworthy AI.[19]

To our knowledge, we are the first to create and deploy an *AI*-driven tool to enhance clinicians' treatment of patients. To the extent that clinicians can obtain reliable predictions of individual health risks, then they can provide more tailored treatments and better monitoring of patients during their visits in the hospital. We are working to deploy these predictions across medical centres, together with a simple heuristic that flags patients as low, medium and high risk based on whether our classifier predicts a probability of death in the top, middle or bottom percentile of the mortality distribution. While our focus is on Veterans, our results generalise to broader contexts since there is overlap in the distribution of covariates between Veterans and non-Veterans (eg, age, education, race).

Traditional measures of health among Veterans focus on physical conditions obtained from, for example, a combination of medical history and demographic factors.[20] These factors are important since they may influence individuals' predisposition to certain ailments.[21] For example, especially with the recent COVID-19 pandemic, age has emerged as one of the most important individual-level predictors of infection risk and mortality.[5 6] However, researchers have struggled to obtain bias-free, reliable and externally-valid predictions on representative datasets.[7]

On top of these individual-level characteristics that serve as important mediating characteristics in the ongoing pandemic, there is also an increasing recognition that geographic factors matter for understanding variation in healthcare utilisation. For example, differences in life expectancy vary significantly across commuting zones, although the dispersion is smaller in higher income areas.[22] Moreover, confidence in healthcare systems and their ability to care for the needs of their communities varies across metropolitan areas.[23]

However, while there is a general understanding that demographics play a role in understanding differences in physical and mental health among individuals, including Veterans, there is also an increasing recognition that social determinants are potentially even more important.[24 25 26] This comes at a time when new data is becoming available. For example, recent work provides a methodology for mining electronic health record (EHR) textual data to detect the presence of homelessness and adverse childhood experiences as predictive factors behind individual health.[10] Unstructured data can provide valuable information about Veteran experiences, allowing researchers to map qualitative information about experiences into comparable indices.

There is also substantial evidence of geographic differences in life expectancy and mortality outcomes. For example, life expectancy is closely related with individual income and these outcomes also vary across geographies with different average incomes, suggesting that local health- care resources may play a role for explaining differences in mortality across space.[22] Moreover, specifically for Veterans, there are large differences in utilisation rates of healthcare services across space, at least in part because of the composition of practices among VA medical professionals at a local level.[27] Additional research also explores how sociodemographic factors help explain differences in COVID-19 deaths across local VA medical centres.[28]

METHODS

The data we use for model training and evaluation come from the EHR at the Department of Veterans Affairs Health Administration (VHA). To develop an ML algorithm capable of predicting mortality within a 30-day window of infection, we analyse patient data from the EHR in the VA Corporate Data Warehouse (CDW). Specifically, we analyse data consisting of of patient demographics, International Classification of Diseases (ICD) Diagnosis codes, blood work and vital signs of patients infected with SARS CoV-2. Our training sample consisted of 11 097 (1294 deceased) treated for COVID-19 from 2 March through 3 August 2020. Before dropping observations with over 25% missing, we have 129 station and 32 706 patients whereas when we drop those with over 25% missing, we have 124 patients and and 11 962 patients. A second validation sample consisting of 1634 (128 deceased) patients treated from 4 August through 24 August 2020 was held out to assess model performance on data that is unbiased from the model training process. Laboratory results indicating positive detection of SARS CoV-2 were used as criteria for infection.

In an effort to create the most predictive model possible, we use the date of positive SARS CoV-2 PCR specimen collection as our chronological reference point for analysis and model training. Variables analysed fall within the following broad categories: patient demographics, comorbidities, chronic acute conditions, laboratory pathology and vital sign values. Several comorbidities are indicative of the mortality window for with patients SARS CoV-2. One distinguishable characteristic among patients that experienced mortality was a higher number of comorbidities. To summarise the level of multimorbidity in patients, we used the Quan-Elixhauser Mortality Index as a variable.[29 30]

We also experiment with data from the Census Bureau's 5-year American Community Survey from 2014 to 2018. The Census provides a wide array of demographic characteristics at county or state level, including: the race distribution, the population density, the share male, the age distribution (the share under age 18, age 25–44, age 45–64 and 65+), the share married, the education distribution (the share with less than a high school degree, some college, and college or more), the income distribution (the share with less than US$15 000, US$15–29 000, US$30–39 000, US$40–49 000, US$50–59 000, US$60–99 000, US$100–1 49 000, over US$150 000), and the poverty rate (the share of people living in poverty under age 18, age 18–64 and 65+). However, after controlling for our individual characteristics, these location characteristics do not improve the model performance. While our prior work has found that these characteristics matter for predicting cross-sectional differences in mortality and infections,[31] our individual-level characteristics in the VA data subsume the zipcode characteristics since they are more granular.

We use the following variables in our predictive models:

- Patient demographics: the latest available observations up until the point of SARS CoV-2 lab specimen collection, including: age, race, ethnicity and marital status.
- Comorbidities: Elixhauser Mortality Score was derived from patient ICD 10 diagnosis codes. These codes were derived from clinical encounters, active problems, inpatient and outpatient billing records ranging back 7 years from date of the patients first positive SARS CoV-2 laboratory test.
- Chronic and other disease history: comprehensive groups were formed using the same set of ICD 10 diagnosis codes for comorbidities to represent certain diseases: dementia, gait and mobility issues, atherosclerosis, prostate problems, hypertension, hyperlipidaemia, anaemia, diabetes and chronic obstructive pulmonary disease (COPD).
- Acute conditions: a second set of ICD 10 codes extrapolated from active problems and encounters was used to code for acute conditions 3 days prior and up to the date of first positive SARS CoV-2 lab: encounter for palliative care, do not resuscitate, hypoxia, pneumonia, respiratory failure, kidney failure, acute respiratory distress syndrome, cardiac arrest and sepsis.
- Lab work: pathology components from the date of the patients first positive SARS CoV- 2 Lab were analysed: erythrocyte mean corpuscular volume fL, erythrocyte sedimentation rate mm/hour, lactate mmol/L, bilirubin—total mg/dL, D-dimer ng/mL, white blood cell count K/cmm, platelets 10*9/L, lactate dehydrogenase U/L, lymphocytes, C reactive protein mg/dL, CO2—partial pressure mm Hg, PO_2 mm Hg, red blood cell count M/cmm, lymphocytes, ferritin ng/mL, urea nitrogen mg/dL and albumin g/dL.
- Vital signs: vital signs from the date of the patients first positive SARS CoV-2 lab were analysed: blood pressure, pulse, temperature, respiration, height, weight, body mass index, pulse oximetry and fraction of inspired oxygen (FIO_2).

Table 1 documents the summary statistics for these characteristics separately for patients who recovered and those who died. Consistent with prior literature, we see stark differences in age between those who recovered and those who died: a mean (median) of 62 (64) years old versus 77 (75), respectively. We see greater dispersion in age among those who recovered (SD of 15 vs 10). We also observe substantial differences among a handful of other lab results, including: lymphocytes, urea nitrogen, platelets, D-dimer, and, perhaps most importantly, the Elix Mortality Score. For example, given that lymphocytes are the B and T cells that help fight infection, it is not surprising that we find that patients who recovered have roughly 43% higher counts than those who died.

For model calibration, we use five-fold cross validation AUPRC mean scores for hyper-parameter optimisation. We also bootstrap the training dataset using five-fold cross validation AUROC, F1 and recall mean scores. After model

Table 1 Descriptive statistics for recovered and deceased patients

	Convalesced mean std		25%	50%	75%	Mortality mean	Std	25%	50%	75%
Age	62.41	15.25	52.78	64.38	72.99	77.01	10.86	70.45	75.85	86.02
Lymphocytes %	21.78	11.21	13.50	20.20	28.40	15.14	11.58	7.70	12.70	19.42
C-reactive protein mg/dL	6.35	6.87	1.16	3.85	9.48	11.33	8.63	4.24	9.47	16.08
Urea nitrogen mg/dL	19.92	15.41	12.00	15.00	22.00	35.29	25.37	18.00	27.00	44.00
Platelets 10*9/L	207.80	80.16	154.00	194.00	246.00	189.76	88.74	135.00	170.00	230.00
CO_2—partial pressure mm Hg	38.76	9.25	32.60	37.40	43.60	39.16	12.06	31.10	36.80	44.90
Erythrocyte mean corpuscular volume fL	88.34	6.26	84.90	88.60	92.20	90.10	6.91	86.10	90.20	94.40
Red blood cell count M/cmm	4.56	0.72	4.15	4.62	5.04	4.12	0.83	3.54	4.15	4.68
D-Dimer ng/mL	616.93	3155.99	70.00	175.00	408.00	1332.36	5836.19	139.25	328.00	774.75
Elix Mortality Score	5.20	14.67	–5.00	2.00	14.00	16.30	15.88	4.00	16.00	28.00
Bilirubin—total mg/dL	0.67	0.55	0.40	0.60	0.80	0.83	1.66	0.40	0.60	0.90
Albumin g/dL	3.66	0.61	3.30	3.70	4.10	3.21	0.64	2.80	3.20	3.70
Pulse	87.23	17.07	75.00	86.00	98.00	89.29	18.93	76.00	88.00	101.00
Systolic	133.03	20.79	119.00	132.00	146.00	129.94	23.87	114.00	128.00	145.00
Diastolic	78.28	12.64	70.00	78.00	86.00	72.97	13.72	64.00	72.00	81.00
Pulse oximetry	96.01	3.42	95.00	96.00	98.00	94.31	5.17	93.00	95.00	97.00
FIO_2	30.55	19.58	21.00	24.00	28.00	41.98	27.43	24.00	28.00	50.00
Respiration	18.81	3.68	17.00	18.00	20.00	20.53	5.13	18.00	20.00	22.00
Temperature	99.05	1.45	98.10	98.70	99.90	99.11	1.68	98.00	98.80	100.10

Sources: Department of Veterans Affairs. The table reports the mean, SD and percentiles of key variables used in the predictive models.

calibration, we evaluate performance on the validation dataset using four metrics: AUROC, AUPRC, F1 and recall scores. We include recall as a primary evaluation metric to see how well the classifier identifies the positive class, that is, mortality in concordance with other metrics that assess overall classification performance.

Our selection of these models was based off of two priorities. First, we require a probabilistic model—that is, one that produces predicted probabilities when fed a vector of 0/1 values. This is useful from an operational standpoint. Users of the model using can adjust the probability threshold for the outcome of mortality to meet their operational needs. For example, consider a Primary Care clinic that uses the model to decide which patients require additional follow-up after diagnosis. If the clinic wants to be more cautious, clinicians can lower the probability threshold. Second, we desire explainability—that is, results that are interpretable and actionable for clinicians. We limit our pool of prospective algorithms to those that could be explained with weights given to each input, allowing us to rank the importance of different features for clinicians. There is a growing recognition that AI must be explainable for it to have the greatest impact and adoption across organisations.[32]

For all evaluated models excluding XGBoost, missing values were imputed using a *K*-nearest neighbours (KNN) method. To mitigate the effects of data sparsity biasing our models, observations missing less than 25% of their dependent variables were dropped from both training and evaluation datasets. While there is no perfect way to deal with missing data, one of the desirable features of XGBoost is its built-in support for sparsity. When decision tree nodes are constructed during the training process, optimal traversal pathways are decided for both for non-missing and missing values.[33] Other models require dense datasets, forcing users to either to drop observations or impute missing values. To compare the performance of XGBoost native support for data sparsity, we evaluate two XGBoost models: one trained and evaluated using KNN imputation and one without.

RESULTS

There has been a proliferation of studies evaluating risk factors behind COVID-19 infections and mortality.[7] Many of these studies have assessed their performance based only on the AUROC. However, looking solely at the AUROC can lead to misleading inferences and weak

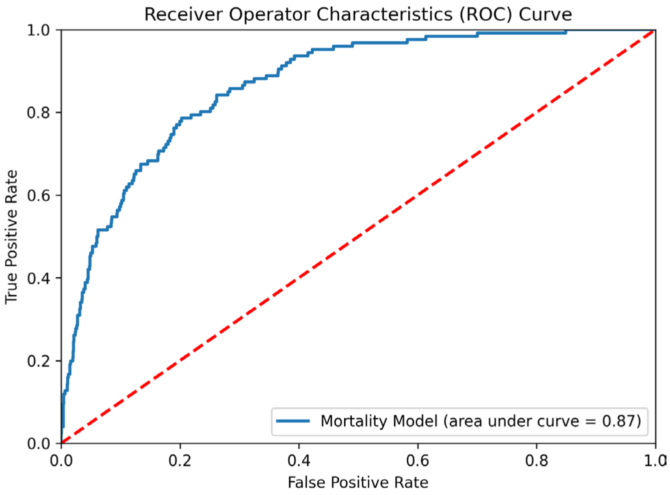


Figure 1 Department of Veterans Affairs. The figure plots the area under the receiver operator characteristics curve for mortality as the outcome variable using XGBoost.

predictive models since infection, as well as mortality, is so rare, meaning that over predicts negative rates will actually boost the AUROC.

In particular, we found that using the AUROC as a primary evaluation metric on imbalanced class datasets produced models with low sensitivity at the default probability rate (0.5). Furthermore, lowering the probability threshold revealed that these models performed very poorly along both sensitivity and specificity. We discovered that, in order to develop a model that is both accurate and captures a greater number of true positives, we applied a broader set of metrics, namely the AUPRC. Nonetheless, figure 1 reports the AUROC, which is 0.87—a score in line with many prior studies.

Of all the models analysed, the XGBoost decision tree ensemble using sparse datasets performed best. Using bootstrapping and five-fold cross validation this model achieved a mean AUROC score of 0.87 (0.86 to 0.88 95% CI), a mean F1 score of 0.49 (0.48 to 0.59 95% CI) and a mean recall score of 0.73 (0.7 to 0.76 95% CI). On the validation dataset, the XGBoost model achieved a 0.87 AUROC score, a 0.41 AUPRC, an F1 score of 0.40 and recall score of 0.11. Figure 2 presents these performance metrics. Part of the reason the performance does not differ much across the different models stems from the fact that we are working with a small sample. A growing literature from computer science suggests that the gains of sophisticated AI models are realised in larger datasets.

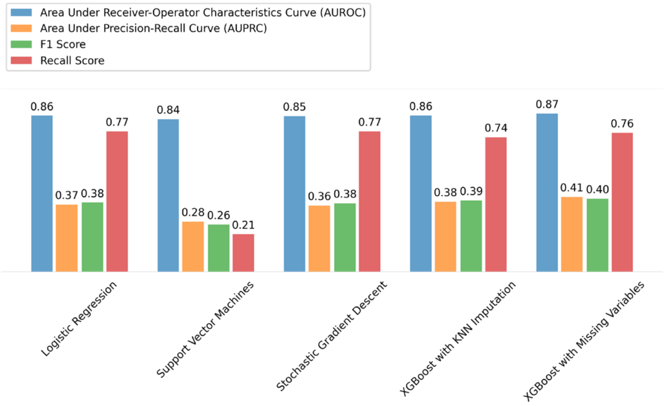


Figure 2 Department of Veterans Affairs. The figure reports the area under the receiver operator characteristics curve (AUROC), area under the precision recall curve (AUPRC), the F1 score, and the recallscore all using different modeling strategies. Recall is equal to the ratio of true positives to the sum of true positives and false negatives. Precision is equal to the ratio of true positives to the sum of true positives and false positives. The F1 score is equal to 2*(Recall * Precision) / (Recall + Precision).

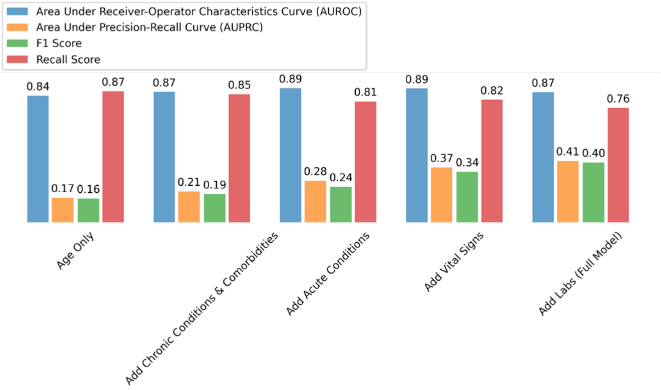


Figure 3 Department of Veterans Affairs. The figure reports the area under the receiver operator characteristics curve (AUROC), area under the precision recall curve (AUPRC), the F1 score, and the recallscore all using different features as predictive characteristics. Recall is equal to the ratio of true positives tothe sum of true positives and false negatives. Precision is equal to the ratio of true positives to the sum oftrue positives and false positives. The F1 score is equal to 2*(Recall *Precision) / (Recall + Precision).

Given that the specific algorithm that we use to predict mortality does not have a large quantitative effect on model quality, we now explore the role of different features as predictive characteristics in figure 3. While the AUROC is highly similar across specifications, the other performance metrics, such as F1 and recall scores, differ significantly. Importantly, since a high AUROC can be obtained in an unbalanced dataset whenever the algorithm produces low probabilities, then we might find an artificially high AUROC. In other words, we may produce a lot of true negatives, which lead to high sensitivity scores, but at the expense of true positives.

While some models yielded slightly higher recall scores at the default probability threshold (0.5), XGBoost performed better on all other metrics. Figure 3 summarises the ROC at various probability thresholds. If users of this model wish to be more cautious, they can simply choose a lower probability threshold at the expense of a higher false-positive rate. At each probability threshold, the table displays the sensitivity (true-positive rate) and specificity (true-negative rate) achieved on the validation dataset. To provide greater insight into the results from our XGBoost model, Figure 4 plots the decision tree and the resulting probabilities at each node. This algorithm is of the family

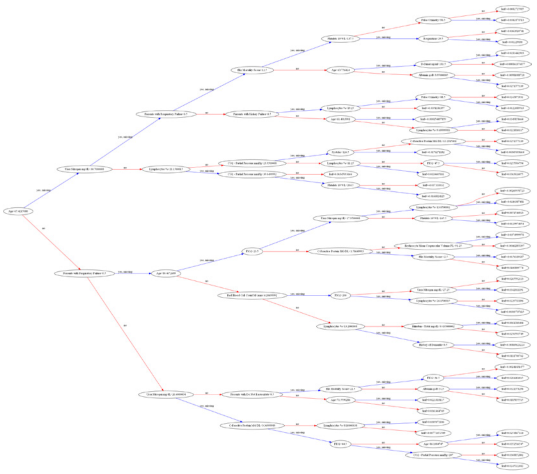


Figure 4 Department of Veterans Affairs. The figure plots the tree for our mortality outcomes using all the variables that were embedded in the model.

of ensemble learning techniques and is based on the famous Random Forest algorithm. The term ensemble learning is used to describe a powerful machine learning method in which multiple machine learning models are used for prediction.

Furthermore, figure 5 ranks the features, by importance, as predictors of mortality outcomes using the F score. Consistent with prior literature, age ranks as the top comorbidity, followed by lymphocytes, C-reactive protein, urea nitrogen, platelets, FIO_2, red blood cell count, enthrocyte mean corpuscular, and D-dimer. These are all intuitive characteristics that would enter into the risk factor. For example, since lymphocytes are the B and T cells that help fight infection, they can decrease during viral diseases. Similarly, platelets allow blood to clot and can decrease with viral infection.

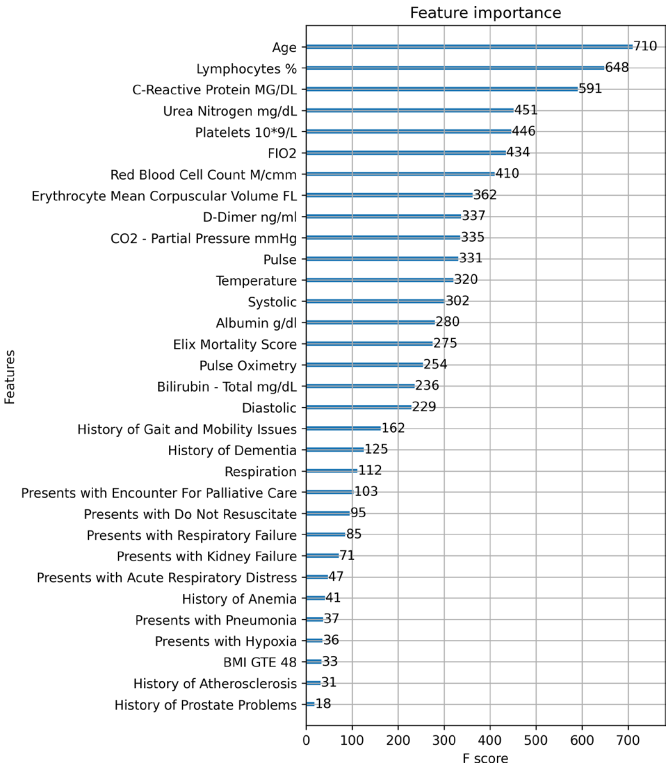


Figure 5 Department of Veterans Affairs. The figure reports the most important features from the estimation of XG Boost using the F score as the metric. BMI, body mass index.

Consider, for example, the AUROC with only age vs the full model, which contains medical conditions, vital signs, and labs. While the AUROC between the two are nearly identical (0.84 vs 0.87), the full model has a substantially higher AUPRC, F1 score, and recall score. For example, the AUPRC and F1 score grow from 0.17 and 0.16 to 0.41 and 0.40, respectively, which is over a two-times order of magnitude increase. We focus on not only who dies (ie, sensitivity=true positives / (true positives+false negatives), but also who recovers (ie, true negatives=true negatives / (true negatives+false positives). The inclusion of chronic conditions, and to a larger extent acute conditions, helps increase the performance of the model, the inclusion of vital signs and labs are the features that improve the model the most. Given that many of the studies in this emerging literature on COVID-19 have focused on AUROC as a metric for evaluating model performance, we view our broader set of metrics as not only a form of model validation, but also a contribution in and of itself for obtaining more reliable predictions.

While there is no strict AUROC and AUPRC threshold for defining reliable models, it is important to focus on the AUPRC in settings with an imbalanced dataset.[34] For example, here we have a small share of patients who died from COVID-19, which puts the AUPRC in perspective, since they show the number of true positives among positive predictions. In this sense, given a mortality rate of 0.043, the baseline AUPRC is 4.43%, so our actual AUPRC of 0.41 is well above what a classifier would predict randomly. Moreover, to better understand the quality of our predictions, figure 6 plots the distribution of the risk factors (eg, convalescence and mortality) across patients with the associated CI. Although we see significant dispersion in the risk factors, the CIs are still fairly narrow, suggesting that these predictions have been reliably estimated.

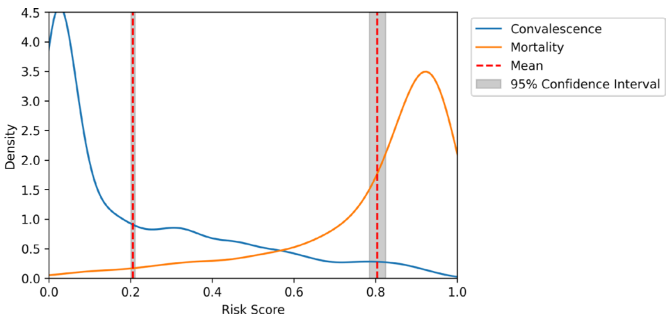
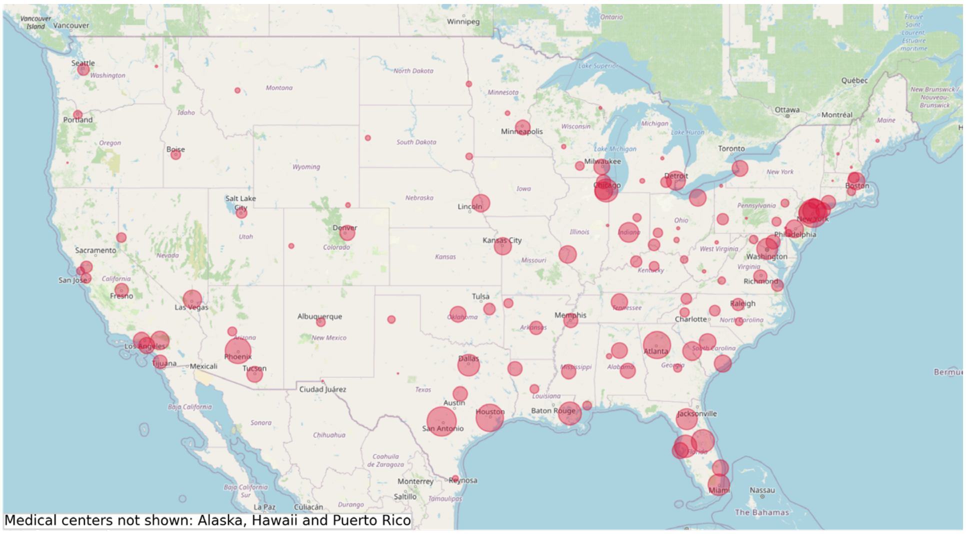


Figure 6 Department of Veterans Affairs. The figure reports the distribution of our predicted risk factor and convalescence with their associated confidence intervals.



Medical centers not shown: Alaska, Hawaii and Puerto Rico

Figure 7 VA medical center facilities in the USA.

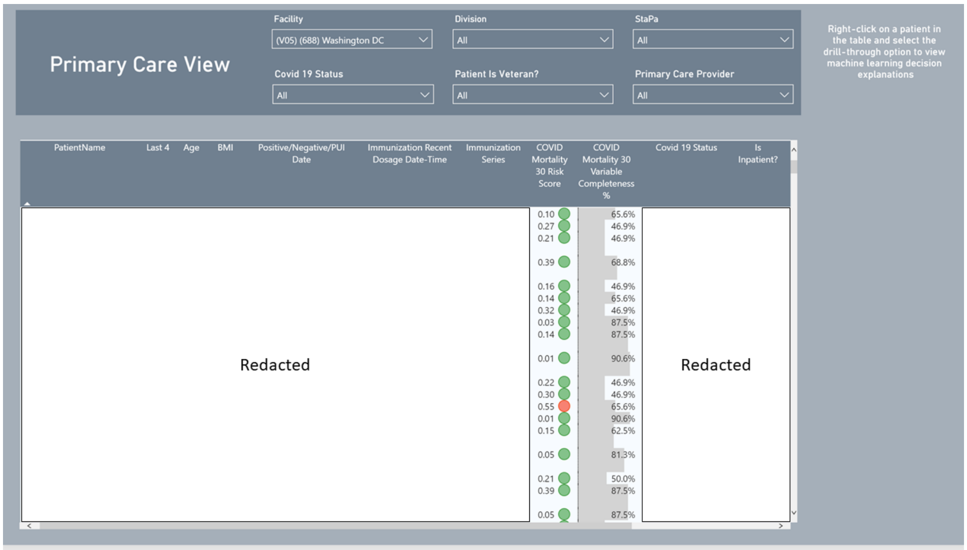
DISCUSSION WITH CLINICAL APPLICATIONS

In addition to creating a predictive model for understanding the role of different comorbidities and obtaining predicted probabilities for mortality, we also create an operational tool that aids in point of care decision-making for treating patients afflicted with SARS CoV-2. We pilot our 30-day mortality model in a PowerBI dashboard available to VA clinicians, built using data from the VA CDW. The dashboard is refreshed daily and uses well-established security practices to keep patient data safe and ensure that information is limited to users' local VA facility. Figure 7 provides a spatial illustration of the VA medical facilities, weighted by the number of patients, across the USA.

The dashboard has two views: one for primary care and another for inpatient care providers. The primary care view allows primary care teams to filter the datasets by patient provider, track COVID-19 testing and view mortality risk scores which are the probabilities generated by the model. For in-patient providers, they can filter the inpatient dataset by specialty and hospital location. These features are embedded so that the AI-driven tool adheres to the principles of trustworthy AI, particularly as they apply to Veterans,[19] namely with a clear purpose (i.e., informing clinicians about the mortality risk of patients), with reliability and accuracy (i.e., reporting performance metrics), and with understandable and actionable analytics (i.e., enumerating the primary factors behind the patient's risk factor).

Figure 8 presents visuals of these dashboards.

One of the most useful features of our dashboard is that providers do not have to take the risk scores at face value. They can search for a view that presents model inputs, variable weights, as well as a list of missing inputs. If they are want to learn more about a patient, they can order labs and/or obtain vital signs from the missing values list to obtain more accurate mortality risk assessments. Model weights are Shapley's Additive Explanations

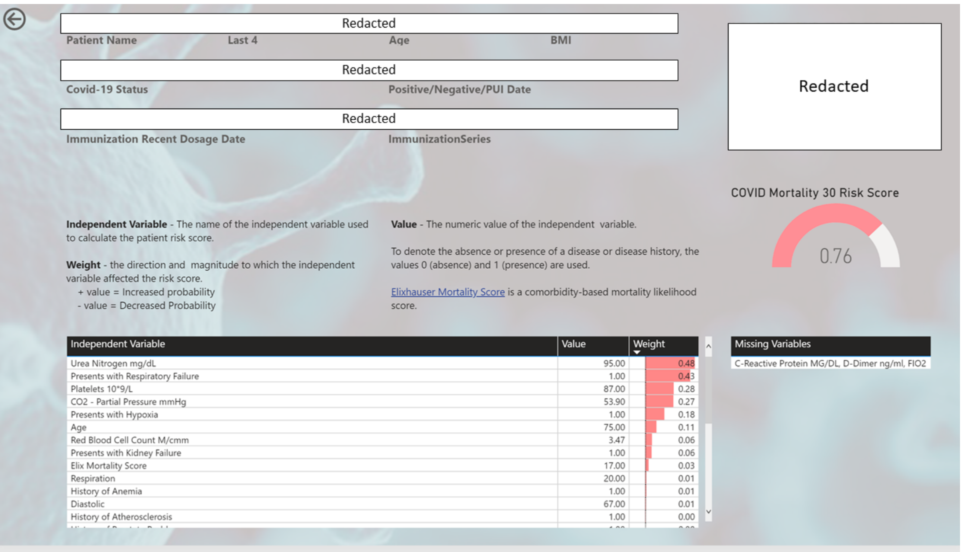

Primary Care View

Facility: (V05) (688) Washington DC
Division: All
StaPa: All
Covid 19 Status: All
Patient Is Veteran?: All
Primary Care Provider: All

Right-click on a patient in the table and select the drill-through option to view machine learning decision explanations

PatientName	Last 4	Age	BMI	Positive/Negative/PUI Date	Immunization Recent Dosage Date-Time	Immunization Series	COVID Mortality 30 Risk Score	COVID Mortality 30 Variable Completeness %	Covid 19 Status	Is Inpatient?
Redacted							0.10 🟢	65.6%	Redacted	
							0.27 🟢	46.9%		
							0.21 🟢	46.9%		
							0.39 🟢	68.8%		
							0.16 🟢	46.9%		
							0.14 🟢	65.6%		
							0.32 🟢	46.9%		
							0.03 🟢	87.5%		
							0.14 🟢	87.5%		
							0.01 🟢	90.6%		
							0.22 🟢	46.9%		
							0.30 🟢	46.9%		
							0.55 🔴	65.6%		
							0.01 🟢	90.6%		
							0.15 🟢	62.5%		
							0.05 🟢	81.3%		
							0.21 🟢	50.0%		
							0.39 🟢	87.5%		
							0.05 🟢	87.5%		

Figure 8 Primary care and in-patient views for mortality predictions.



Redacted

Patient Name Last 4 Age BMI

Redacted

Covid-19 Status Positive/Negative/PUI Date

Redacted

Immunization Recent Dosage Date ImmunizationSeries

Redacted

COVID Mortality 30 Risk Score

0.76

Independent Variable - The name of the independent variable used to calculate the patient risk score.

Weight - the direction and magnitude to which the independent variable affected the risk score.
+ value = Increased probability
- value = Decreased Probability

Value - The numeric value of the independent variable.

To denote the absence or presence of a disease or disease history, the values 0 (absence) and 1 (presence) are used.

Elixhauser Mortality Score is a comorbidity-based mortality likelihood score.

Independent Variable	Value	Weight
Urea Nitrogen mg/dL	95.00	0.48
Presents with Respiratory Failure	1.00	0.43
Platelets 10*9/L	87.00	0.28
CO2 - Partial Pressure mmHg	53.90	0.27
Presents with Hypoxia	1.00	0.18
Age	75.00	0.11
Red Blood Cell Count M/cmm	3.47	0.06
Presents with Kidney Failure	1.00	0.06
Elix Mortality Score	17.00	0.03
Respiration	20.00	0.01
History of Anemia	1.00	0.01
Diastolic	67.00	0.01
History of Atherosclerosis	1.00	0.00

Missing Variables
C-Reactive Protein MG/DL, D-Dimer ng/ml, FIO2

Figure 9 Risk factors for mortality predictions.

(SHAP) values. SHAP is a game theoretical approach to explain the output of machine learning models.[35] SHAP values allow users of our dashboard to see the direction and magnitude to which each variable input affects the patient's risk score. Figure 9 plots a visual for the risk factor layout of the dashboard.

This view provides explains how the model arrived at its concluded risk score to clinicians. The table displays each dependent variable input used by the XGBoost model to derive the individual's risk score. The 'Feature' column is the dependent variable name, the 'Explanation' column is the weight that is, the direction and magnitude that the input effected the risk score, and the 'Value column is the numeric value of the dependent variable. Positive explanation values imply that the input increased the risk score and negative values imply the inverse.

While our tool helps clinicians improve their treatment of patients and guide them to the most pressing risk factors, we recognise that the tool has at least two limitations. First, it is not meant to tell clinicians what to do: our AI is designed to augment clinician responsibilities, not replace them. Second, since the tool provides a list of important determinants of the risk factor, the clinician is called to think about potential explanations behind the phenomena that they observe with the patient. In this sense, the AI is designed to help consolidate data and draw out the clinician's knowledge and expertise to drive better patient outcomes.

CONCLUSION

While there is already a large literature exploring the contributions of demographic factors and pre-existing conditions to COVID-19, there is little empirical evidence on the role that sociodemographic factors play within a community. This paper draws on administrative data from the Department of VA and each of their medical centres to estimate predictive models for mortality as a function of individual demographic characteristics, medical history, and labs and vitals for every Veteran under the VA's care.

Our model performs well on not only the conventional AUROC metric, but also other metrics, such as the AUPRC, F1 score and recall score. We show that these metrics are important for producing reliable predictive models since the mortality rate for COVID-19 is so low, meaning that models tuned to maximise the AUROC are likely to produce many false positives.

Using our new predictive model, we develop and implement a dashboard for clinical application in the District of Columbia VA medical centre. Our dashboard provides clinicians with not only the medical history and demographic characteristics of patients, but also risk factors that incorporate the results of our predictive models. In particular, we use our estimated models, together with the individual-level characteristics, to generate personalised predicted probabilities that the individual will experience acute hospitalisation and mortality, which we flag for the clinicians to help them maximise the odds for a successful recovery by the patient.

Our results open up a number of interesting avenues. Most importantly, we are in the process of piloting our clinical diagnostic tool with more medical centres with an intent in gauging the effectiveness of the instrument and identifying ways of improving it. We are also interested in extending the tool into other conditions and viruses; COVID-19 is simply on specific application. Moreover, we believe that there is significant value in a 'learning healthcare system' where medical centres prototype different tools, pool their combined knowledge, and iterate over quality improvements for the purpose of driving better health outcomes for their patients.

OTHER INFORMATION

This work was supported by the Department of VA Office of Research & Development. Apart from the data on

demographic characteristics, which are publicly available from the Census Bureau, the administrative data on Veterans and their medical information is restricted to the Department of VA.

Christos A. Makridis contributed to the design, writing and editing of the paper. Tim Strebel contributed to the analysis. Vince Marconi contributed to the editing of the paper. Gil Alterovitz contributed to the design and editing of the paper.

Contributors CM contribute to the design, writing and editing of the paper. TS contributed to the analysis. VM contributed to the editing of the paper. GA contributed to the design and editing of the paper.

Funding This work was supported by the Department of Veterans Affairs. Apart from the data on demographic characteristics, which are publicly available from the Census Bureau, is restricted to the Department of Veterans Affairs. Christos Makridis contributed to the design, writing, and editing of the paper. Tim Strebel contributed to the analysis. Vince Marconi contributed to the editing of the paper. Gil Alterovitz contributed to the design and editing of the paper.

Map disclaimer The depiction of boundaries on this map does not imply the expression of any opinion whatsoever on the part of BMJ (or any member of its group) concerning the legal status of any country, territory, jurisdiction or area or of its authorities. This map is provided without any warranty of any kind, either express or implied.

Competing interests None declared.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data comes almost exclusively from the internal VA Corporate Data Warehouse (CDW) database and none of the information leaves the VA premises. This project has been completed in partnership between the National Artificial Intelligence Institute at the Department of Veterans Affairs and the local Washington DC VA medical center.

ORCID iDs

Christos A Makridis http://orcid.org/0000-0002-6547-5897
Gil Alterovitz http://orcid.org/0000-0002-0495-7059

REFERENCES

1 Makridis CA, Hartley J. The cost of COVID-19: a rough estimate of the 2020 GDP impact.Mercatus center, policy brief special edition 2020.
2 Cajner T, Crane L, Decker RA, *et al*. The U.S. labor market during the beginning of the pandemic recession. BFI working paper. 2020.
3 Banerjee A, Pasea L, Harris S, *et al*. Estimating excess 1-year mortality associated with the COVID-19 pandemic according to underlying conditions and age: a population-based cohort study. *The Lancet* 2020;395:1715–25.
4 Witters D, Harter J. Worry and stress fuel record drop in U.S. life satisfaction. Gallup 2020.
5 Britton T, Ball F, Trapman P. A mathematical model reveals the influence of population heterogeneity on herd immunity to SARS-CoV-2. *Science* 2020;369:846–9.
6 Martin CA, Jenkins DR, Minhas JS, *et al*. Socio-Demographic heterogeneity in the prevalence of COVID-19 during lockdown is associated with ethnicity and household size: results from an observational cohort study. *EClinicalMedicine* 2020;25:100466.
7 Wynants L, Van Calster B, Collins GS, *et al*. Prediction models for diagnosis and prognosis of covid-19 infection: systematic review and critical appraisal. *BMJ* 2020;369:369.
8 Amarasingham R, Moore BJ, Tabak YP, *et al*. An automated model to identify heart failure patients at risk for 30-day readmission or death using electronic medical record data. *Med Care* 2010;48:981–8.
9 Navathe AS, Zhong F, Lei VJ, *et al*. Hospital readmission and social risk factors identified from physician notes. *Health Serv Res* 2018;53:1110–36.
10 Bejan CA, Angiolillo J, Conway D, *et al*. Mining 100 million notes to find homelessness and adverse childhood experiences: 2 case studies of rare and severe social determinants of health in electronic health records. *J Am Med Inform Assoc* 2018;25:61–71.
11 Rentsch CT, Kidwai-Khan F, Tate JP, *et al*. Patterns of COVID-19 testing and mortality by race and ethnicity among United States veterans: A nationwide cohort study. *PLoS Med* 2020;17:e1003379.
12 Williamson EJ, Walker AJ, Bhaskaran K, *et al*. Factors associated with COVID-19-related death using OpenSAFELY. *Nature* 2020;584:430–6.
13 Zhou F, Yu T, Du R, *et al*. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. *Lancet* 2020;395:1054–62.
14 Richardson S, Hirsch JS, Narasimhan M, *et al*. Presenting characteristics, comorbidities, and outcomes among 5700 patients hospitalized with COVID-19 in the new York City area. *JAMA* 2020;323:2052–9.
15 Makridis CA, Wu C. Ties that bind (and social distance): how social capital helps Com- munities weather the COVID-19 pandemic. *PLoS One* 2021;16:1.
16 Guo L, Wei D, Zhang X. Clinical features predicting mortality risk in patients with viral pnemonia: the MuLBSTA score. *Frontiers in Microbiology* 2019:10.
17 Osborne TF, Veigulis ZP, Arreola DM, *et al*. Automated EHR score to predict COVID-19 outcomes at US department of Veterans Affairs. *PLoS One* 2020;15:e0236554.
18 King JT, Yoon JS, Rentsch CT, *et al*. Development and validation of a 30-day mortality index based on pre-existing medical administrative data from 13,323 COVID-19 patients: the Veterans health administration COVID-19 (VACO) index. *PLoS One* 2020;15:e0241825.
19 Makridis C, Hurley S, Klote M, *et al*. Ethical applications of artificial intelligence: evidence from health research on veterans. *JMIR Med Inform* 2021;9:e28921 https://medinform.jmir.org/2021/6/e28921
20 Haibach JP, Haibach MA, Hall KS, *et al*. Military and veteran health behavior research and practice: challenges and opportunities. *J Behav Med* 2017;40:175–93.
21 Kazis LE, Ren XS, Lee A, *et al*. Health status in Va patients: results from the Veterans health study. *Am J Med Qual* 1999;14:28–38.
22 Chetty R, Stepner M, Abraham S, *et al*. The association between income and life expectancy in the United States, 2001-2014. *JAMA* 2016;315:1750–66.
23 Ahern MM, Hendryx MS. Social capital and trust in providers. *Soc Sci Med* 2003;57:1195–203.
24 Mokdad AH, Marks JS, Stroup DF, *et al*. Actual causes of death in the United States, 2000. *JAMA* 2004;291:1238–45.
25 Holt-Lunstad J, Smith TB, Layton JB. Social relationships and mortality risk: a meta-analytic review. *PLoS Med* 2010;7:7.
26 Makridis CA, Zhao DY, Bejan CA, *et al*. Leveraging machine learning to characterize the role of socio-economic determinants on physical health and well-being among Veterans. *Comput Biol Med* 2021;133:104354.
27 Ashton CM, Petersen NJ, Souchek J, *et al*. Geographic variations in utilization rates in Veterans Affairs hospitals and clinics. *N Engl J Med* 1999;340:32–9.
28 Makridis CA, Mudide A, Alterovitz G. How much does the (social) environment matter? using artificial intelligence to predict COVID-19 outcomes with socio-demographic data. *Pacific Symposium on Biocomputing* 2020 https://psb.stanford.edu/psb-online/proceedings/psb21/makridis.pdf
29 Elixhauser A, Steiner C, Harris DR, *et al*. Comorbidity measures for use with administrative data. *Med Care* 1998;36:8–27.
30 Quan H, Sundararajan V, Halfon P, *et al*. Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. *Med Care* 2005;43:1130–9.
31 Makridis CA, Mudibe A, Alterovitz G. How much does the (social) environment matter? using artificial intelligence to predict COVID-19 outcomes with socio-demographic data. Proceedings for the Pacific Symposium on biocomputing 2021.
32 Jiménez-Luna J, Grisoni F, Schneider G. Drug discovery with explainable artificial intelligence. *Nature Machine Intelligence* 2020;2:573–84.
33 Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. KDD '16 2016.
34 Saito T, Rehmsmeier M. The Precision-Recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One* 2015;10:e0118432.
35 Lundberg SM, Si L. A unified approach to interpreting model predictions. advances in neural information processing systems 30 (NIPS 2017) 2017.