BMJ Health & Care Informatics

Exploring stakeholder attitudes towards AI in clinical practice

Ian A Scott ,[1,2] Stacy M Carter,[3] Enrico Coiera [4]

To cite: Scott IA, Carter SM, Coiera E. Exploring stakeholder attitudes towards AI in clinical practice. *BMJ Health Care Inform* 2021;**28**:e100450. doi:10.1136/bmjhci-2021-100450

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/bmjhci-2021-100450).

Received 22 July 2021
Accepted 14 November 2021

[1]Internal Medicine and Clinical Epidemiology, Princess Alexandra Hospital, Woolloongabba, Queensland, Australia
[2]School of Clinical Medicine, University of Queensland, Brisbane, Queensland, Australia
[3]Australian Centre for Health Engagement Evidence and Values, School of Health and Society, University of Wollongong, Wollongong, New South Wales, Australia
[4]Centre for Clinical Informatics, Macquarie University, Sydney, New South Wales, Australia

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

ABSTRACT

Objectives Different stakeholders may hold varying attitudes towards artificial intelligence (AI) applications in healthcare, which may constrain their acceptance if AI developers fail to take them into account. We set out to ascertain evidence of the attitudes of clinicians, consumers, managers, researchers, regulators and industry towards AI applications in healthcare.

Methods We undertook an exploratory analysis of articles whose titles or abstracts contained the terms 'artificial intelligence' or 'AI' and 'medical' or 'healthcare' and 'attitudes', 'perceptions', 'opinions', 'views', 'expectations'. Using a snowballing strategy, we searched PubMed and Google Scholar for articles published 1 January 2010 through 31 May 2021. We selected articles relating to non-robotic clinician-facing AI applications used to support healthcare-related tasks or decision-making.

Results Across 27 studies, attitudes towards AI applications in healthcare, in general, were positive, more so for those with direct experience of AI, but provided certain safeguards were met. AI applications which automated data interpretation and synthesis were regarded more favourably by clinicians and consumers than those that directly influenced clinical decisions or potentially impacted clinician–patient relationships. Privacy breaches and personal liability for AI-related error worried clinicians, while loss of clinician oversight and inability to fully share in decision-making worried consumers. Both clinicians and consumers wanted AI-generated advice to be trustworthy, while industry groups emphasised AI benefits and wanted more data, funding and regulatory certainty.

Discussion Certain expectations of AI applications were common to many stakeholder groups from which a set of dependencies can be defined.

Conclusion Stakeholders differ in some but not all of their attitudes towards AI. Those developing and implementing applications should consider policies and processes that bridge attitudinal disconnects between different stakeholders.

Summary

What is already known?

- Very little is known about the attitudes of different stakeholders towards artificial intelligence (AI) applications in healthcare.
- While the AI industry see their applications as promising for improving healthcare, the views of clinicians, patients and other groups directly involved in delivering or receiving care may not be so favourable.

What does this paper add?

- This paper provides an exploratory analysis of published reports of the attitudes and perceptions of different stakeholder groups towards AI applications in healthcare.
- Stakeholder groups hold similar attitudes towards AI on some attributes but differ in their attitudes towards others.
- In general, attitudes towards AI in healthcare were positive, more so for those with direct experience of AI in care delivery, but with the proviso that certain safeguards were met.
- Those developing and implementing AI applications should consider policies and processes that bridge attitudinal disconnects between different stakeholders.

INTRODUCTION

Artificial intelligence (AI) refers to advanced computer programs that mimic intelligent human behaviours and assist humans with different tasks. Medical AI applications span a spectrum, from diagnosis and disease screening to treatment selection and prognostication,[1] and aim to optimise care, improve efficiency and enhance clinician and consumer experience. Despite scores of AI applications having received regulatory approval for use in clinical settings in recent years, and many more having passed the proof-of-concept stage, relatively few that purport to directly assist decision-making have been adopted at scale into clinical practice.[2] This limited uptake may be due, at least partly, to misperceptions of what the term AI actually means and negative attitudes towards AI held by key players in the healthcare ecosystem. Multiple stakeholders share interest in the performance and outcomes of AI applications, comprising clinicians, consumers, managers, researchers, regulators and industry. Their perceptions and expectations of AI may differ, and need to be understood and considered by AI developers and implementers if AI applications

are to be designed and operationalised in ways acceptable to all parties.

METHODS

We undertook an exploratory analysis of articles whose titles or abstracts contained the terms 'artificial intelligence' or 'AI' and 'medical' or 'healthcare' and 'attitudes', 'perceptions', 'opinions', 'views', 'expectations'. Using a snowballing strategy, we searched PubMed and Google Scholar for articles published 1 January 2010 through 31 May 2021. Reference lists of retrieved articles were perused for additional studies. We excluded articles that did not employ a formal survey or interview tool and/or did not report quantified response measures for individual questions among respondents. We only selected articles dealing with non-robotic AI applications used to support clinician-mediated care-related tasks or decision-making, and excluded mobile or wearable applications that were exclusively consumer facing. Key findings were extracted and summarised in narrative form according to four categories of participants. We used these results to derive a thematic synthesis of stakeholder expectations and corresponding requirements (or dependencies) for developers of AI applications to consider.

RESULTS

A total of 27 articles were included[3–29] of which most (16, 59%) targeted clinicians,[3–18] 8 (30%) focused on consumers (including patients),[19–26] 1 (4%) on health executives[27] and 2 (7%) on industry stakeholders comprising AI vendors, researchers and regulators.[28 29] Detailed study descriptions are provided in the online supplemental appendix and summary results are listed in table 1. Most studies (23; 85%) used online surveys,[3–20 22–24 27 28] of which only three (11%)[15 17 24] were designed using the Checklist for Reporting Results of Internet E-Surveys.[30] Three (11%) studies undertook face to face interviews,[25 26 29] and one used a paper-based questionnaire.[21] A specific definition or example of AI was provided to participants in only 10 (37%) studies,[3 8 17 19 22–27] with generic descriptors (eg, 'computers' or 'machines') used in 6 (22%)[5 13 14 16 28 29] and none in 11 (41%).[4 6 7 9–12 15 18 20 21] Survey response rates were reported in 11 (41%) studies,[5 6 9 12 13 15 17 18 21 23 28] ranging from <0.1% to 66%, with 6 (22%)[7 8 10 11 14 16] reporting no response rates and the remainder using convenience samples[19 20 22–27 29] of which one calculated a required sample size.[19]

Clinicians

Clinicians practising in imaging-based disciplines, where deep machine learning is most advanced, featured in several surveys. In an Australian survey of 632 specialists (ophthalmology (n=305), radiology/radiation oncology (n=230), dermatology (n=97)),[3] most had never actually used any AI application in practice (81%), but predicted AI would improve their field (71%) and impact future workforce needs (86%). Most considered AI had to perform better than specialists for disease screening

Table 1 Stakeholder perceptions of clinical AI applications

Positive perceptions	Negative perceptions
Clinicians	
Improved diagnostic accuracy; fewer errors[3 5] More efficient work flows[4 5 17 18] Less time spent on administrative and other mundane tasks[3 13] Synthesis of clinical information[15 18] Updating of clinical records[14] More time spent with patients[5] Improved access to care[3]	Liability for AI-mediated errors[3] Insufficient training and continuing professional development in AI[3 5 7 8 12] Reputational loss and reduced demand for specialist opinion[9 18] Potential erosion of empathetic communication with patients[13 18] Risk of privacy breaches and loss of confidentiality of patient information[17] Lack of proof of efficacy of AI applications in clinical settings[3 29] Lack of explainability[16]
Consumers	
Second opinions to clinicians[21 22 25] Improved access to care[23]	Dehumanisation of the clinician–patient relationship[18 19] Threat to shared decision-making involving patients[22] Low trustworthiness of AI advice[19 20 23] Insufficient clinician and regulatory oversight[21] Uncertainty around fairness and equity in treatment allocation[26]
Healthcare executives	
Improved operational efficiency, cybersecurity, analytic capacity, cost savings[27]	Uncertainty around patient satisfaction, access to care, improved patient outcomes[27]
Industry professionals	
Shared many of the positive attitudes listed above[27–29]	Limited access to high quality data for model development[29] Unresolved legal liability question[29] Lack of explicit and robust regulatory frameworks[29] Low levels of funding for independent, investigator-led research in AI[29]

AI, artificial intelligence.

(64%) and diagnosis (80%). The top three perceived AI benefits were improved patient access to screening, greater diagnostic confidence and reduced specialist time spent on mundane tasks. The top three concerns were outsourcing application development to large commercial AI companies, clinician liability due to AI errors and decreased reliance on specialists ('do-it-yourself' medicine). Most respondents (86%) felt their professional colleges were ill prepared for introducing AI into practice, citing need for training curricula, guidelines and working groups with AI expertise.

Radiologist attitudes towards AI were mostly positive. Most surveyed Italian radiologists (n=1032) favoured adopting AI (77%), did not fear job loss due to AI (89%) and anticipated fewer diagnostic errors (73%) and optimised workflows (68%), although at the expense of some reputational loss and decreased demand for their services (60%).[4] Among 270 French radiologists, most anticipated fewer errors (81%), reduced time spent on image interpretation (74%) and more time spent with patients (52%), with most wanting ongoing education in AI (69%).[5]

Trainees and medical students with an interest in radiology expressed more mixed views, with a third of 69 US radiology residents stating, with hindsight, they may have chosen a different career because of AI.[6] Among 484 UK medical students, half (49%) were disinclined towards a radiology career, despite most (89%) seeing expertise in AI as benefitting them (89%) and wanting AI education included in medical degrees (78%).[7] In Germany, 263 medical students thought AI will improve radiology (86%), not replace radiologists (83%), and desired further training in AI (71%).[8] Canadian students (n=322) expressed similar views, but also voiced concerns about reduced radiologist demand (67%).[9]

Clinicians in pathology and dermatology also tended to view AI positively. Among 487 survey respondents in pathology from 59 countries, 73% expressed interest or excitement in AI as a diagnostic tool for improving workflow efficiency and quality assurance.[10] Fewer than 20% feared displacement or negative career impacts, with most (73%) stating diagnostic decision-making should remain a predominantly human task or one shared equally with AI. While only 25% were concerned about AI errors, opinions about medico-legal responsibility were split, with 44% believing the AI vendor and pathologist should be held equally liable and 50% believing the pathologist should bear prime responsibility. Most (93%) pathologists supported AI if it resulted in more time being spent on academic or research efforts in answering questions previously not possible. Similarly, among 1271 dermatologists from 92 countries, 77% saw AI as improving diagnostic accuracy, particularly in regards to dermatoscopic images, and 80% thought AI should be part of medical training.[11] Less than 6% saw dermatologists being replaced by AI, although 18% held non-specified fears of negative impacts. In contrast, being replaced by AI was of great concern to 27% of laboratory workers and non-clinical technicians in a survey of 1721 subjects, although most (64%) expressed support for AI projects within their organisation and 40% believed AI could reduce errors and save time in their routine work.[12]

Clinicians from non-imaging-based disciplines considered the potential of AI to be more limited. Among 720 UK general practitioners, most (>70%) thought human empathy and communication could not be emulated by AI, that value-based care required clinician judgement, and that benefits of AI would centre on reducing workflow inefficiencies, particularly administrative burdens.[13] Similarly, most psychiatrist respondents (n=791) from 22 countries felt AI was best suited to documenting and updating medical records (75%) and synthesising information to reach a diagnosis (54%).[14] Among 669 Korean doctors, most (83%) considered AI useful in analysing vast amounts of clinical data in real time, while more than a quarter (29%) thought AI would fail in dealing with uncommon scenarios owing to inadequate data.[15] Respondents felt responsibility for AI-induced errors lay with doctors (49%), patients consenting to use of AI (31%) or AI companies that created the tools (19%). Most Chinese clinicians (82% of 191) were disinclined to use an AI diagnostic tool they did not trust or could not understand how it would improve care.[16] Among 98 UK clinicians (including 34 doctors, 23 nurses, 30 allied health professionals), 80% expressed privacy concerns and 40% considered AI potentially dangerous (indeed as bad as nuclear weapons, although this response was primed by reference to a film in which Elon Musk expressed similar sentiments).[17] However, 79% also believed AI could assist their field of work and 90% had no fear of job loss. In a survey of 250 hospital employees from four hospitals in Riyadh, Saudi Arabia (nurses=121; doctors=70; technicians=59), the majority stated AI could reduce errors (67%), speed up care processes (70%) and deliver large amounts of high-quality, clinically relevant data in real time (65%).[18] However, most thought AI could replace them in their job (78%) despite AI limitations in being unable to provide opinions in every patient (66%) or in unexpected situations (64%), unable to sympathise with patients (67%) and developed by computer specialists with little clinical experience (68%).

Consumers

Consumer surveys of AI in healthcare are few and yield mixed views depending on who was surveyed and what AI functions were considered. Most clinical trials of AI tools also omit assessment of patient attitudes.[31] In general, patients view AI more favourably than non-patients, but only if AI is highly trustworthy and associated with clinician oversight.

An online US survey of 50 individuals revealed dehumanisation of clinician–patient relations, low trustworthiness of AI advice and lack of regulatory oversight as significant risks which predominated over potential benefits, although privacy breaches or algorithm bias were not expressed as major concerns.[19] In an online survey

of 6000 adults from various countries, only 27% respondents expressed comfort with doctors using AI to influence clinical decisions.[20]

In a survey of 229 German patients, most (≥60%) favoured physicians over AI for history taking, diagnosis and treatment plans, but simultaneously acknowledged AI could help integrate the most recent scientific evidence into clinician decision-making.[21] Most (>60%) preferred physician opinion to AI where the two disagreed, and were less accepting (≤45%) of AI use in cases of severe versus less severe disease. In a UK case-based questionnaire study involving 107 neurosurgery patients, most accepted using AI for image interpretation (66%), operative planning (76%) and real-time alert of potential complications (73%), provided the neurosurgeon was in control at all times.[22] Among 1183 mostly female patients with various chronic conditions who were considering biometric monitoring devices and AI, only 20% considered benefits (such as improved access to care, better follow-up, reduced treatment burden) greatly outweighed risks and 35% would decline the use of AI-based tools in their care.[23] The majority (>70%) of parents of paediatric patients (n=804) reported openness to AI-driven tools if accuracy was proven, privacy and shared decision-making were protected and care using AI was convenient, of low cost, and not in any way dehumanised.[24] Among 48 US dermatology patients, most (60%) anticipated earlier diagnosis and better care access, while 94% saw the main function of AI as offering second opinions to physicians, and perceived AI as having both strengths (69% believed AI to be very accurate most of the time) and weaknesses (85% expected rare but serious misdiagnoses).[25] A small study found 18 patients with meningioma wanted assurance that use of AI to allocate treatment was fair and equitable, that AI-mediated mistakes would be disclosed and reparations to patients forthcoming and that patient consent was obtained for any sharing of health data.[26]

Healthcare executives

In a global survey of 180 healthcare executives, 40% of respondents overall favoured increased use of AI applications, although this figure varied according to jurisdiction, with Australian executives (23%) being least in favour.[27] Perceived AI benefits comprised improved cybersecurity (56%) operational efficiency (56%), analytics capacity (50%) and cost savings (43%). However, fewer respondents thought there would necessarily be improvements in patient satisfaction (13%), access to care (10%) or clinical outcomes (6%). Respondents cited success factors for AI implementation as comprising adequate staff training and expertise (73%), explicit regulator legislation (64%) and mature digital infrastructures (62%).

Industry professionals

Information technology (IT) specialists, technology and software vendors, researchers and regulators—the 'insiders' of AI—may harbour attitudes different to those of AI users such as clinicians, consumers and healthcare executives.

In one German survey (n=123; 42 radiologists, 55 IT specialists, 26 vendors), all three groups mostly agreed (>75%) that AI could improve efficiency of care, provided AI applications had been validated in clinical studies, were capable of being understood by clinicians and were referenced in medical education.[28] However, only 25% of participants would advocate sole reliance on AI results, only 14% felt AI would render care more human and 93% required confirmation of high levels of accuracy. In interviews involving 40 French subjects (13 physicians, 7 industry representatives, 5 researchers, 7 regulators, 8 independent observers), all agreed reliable AI required access to large quantities of patient data, but such access had to be coupled with confidentiality safeguards and greater transparency in how data were gathered and processed to protect the integrity of physician–patient relationships.[29] On other matters there were notable differences. Physicians highlighted many tools lacked proof of efficacy in clinical settings and they would not assume criminal liability if a tool they could not understand produced errors. Industry representatives wanted greater access to more high-quality data, while wanting to avoid injury liability as they believed this would hinder tool development. Regulators were urgently searching for robust procedures for assessing safety of constantly evolving AI tools, and resolving liability for AI error which would otherwise discourage clinicians and patients from using AI. Researchers with no commercial sponsors wanted more funding and more rapid translation of their findings into practice.

Expectations and dependencies

Our analysis identified certain stakeholder expectations of AI (table 2), with the most frequently cited being a need for accurate and trustworthy applications that improve clinical decision-making, workflow efficiencies and patient outcomes, but which do not diminish professional roles. These expectations, which varied in strength of expression across studies, reflect the dominance of clinician surveys in existing studies. The corresponding self-explanatory dependencies were extrapolated by the authors, and are aligned with those expressed in authoritative reports from the National Academy of Medicine[32] and the WHO.[33] According to these bodies, understanding stakeholder views is essential in formulating clinical AI policy and that AI designers should focus on education, communication and collaboration in bridging attitudinal disconnects between different stakeholders.

DISCUSSION

Overview of findings

The diversity in attitudes towards AI of different stakeholders and the cautionary sentiments expressed by many suggest AI applications should be seen as complex sociotechnical systems with many interacting components.[34]

Table 2 Expectations and dependencies

Expectations	Dependencies
Ensuring accuracy, freedom from bias, trustworthiness.[3–5 8 19 20 23 24 29]	AI applications should be based on models that, in their development, have involved domain experts and have minimised bias related to under-representation of patient groups or contextually inappropriate outcome measures, and have been shown to produce accurate results in the populations for which they are to be used.
Improving efficiency and reduced administrative burden.[3–5 10 13–15 17 18,]	AI applications must be fitted to, and complement, routine clinical workflows and, where possible, self-populate the required data with minimal clinician input.
Improving clinical decision-making and outcomes.[3 11 18 21 22 25 27 29]	AI applications must be shown to be as or more effective in improving clinical decision-making and patient experiences and outcomes than current care, not just efficacious in controlled research settings, and be accompanied with clinician oversight.
Maintaining the integrity of clinician-patient relationships.[5 13 18 19 24]	AI applications should not distract from, or degrade, human to human interaction and shared decision-making.
Ensuring explainability and transparency.[16 19 20 23]	AI applications must be developed and assessed with an eye to maximising explainability and transparency in regards to their inner workings, while acknowledging limits to the extent this can be achieved. As much as possible, important features underpinning AI predictions should be identified, and outputs should be presented in ways easily interpretable to clinicians and patients.
Preserving professional status.[3–9 11 12 18]	AI applications must be implemented with care regarding potential loss of jobs or professional reputation, highlighting the potential of AI to remove the tedious aspects of work, improve job satisfaction and provide new skills. This must be coupled with careful attention to clinicians' training needs and career development.
Obtaining regulatory approval.[3 19 21 27 29]	AI applications should be subject to regulatory standards that are robust, transparent and responsive to updates of existing applications.
Determining liability for error.[3 10 19 21 29]	AI applications should be associated with clear lines of responsibility regarding liability for error, including no-fault provisions when, despite good evidence of efficacy and safety, errors occur as a result of technical failures involving applications whose workings are beyond the comprehension and control of the human user.
Ensuring data privacy, confidentiality and security.[17 24 27]	AI developers must ensure they adhere to legal and community expectations regarding privacy, confidentiality and security of health and medical data.
Ensuring access and equity.[24–26]	AI applications shown to be effective must be equitably accessible to low income, remote or other disadvantaged populations, and not be concentrated in already well-served populations with well-structured digital and data infrastructures.

AI, artificial intelligence.

However, stated positive or negative perceptions of AI may not consistently translate into adoption or resistance, or necessarily track what is possible or even probable in a still-developing technology. The failure of many survey studies to cite concrete examples of AI applications in the prelude to questionnaires (some justifying this as a way of avoiding the conjuring up of negative 'Terminator' or 'cyborg' images) may have caused confusion among respondents as to what they were being asked to conceptualise and respond to. Response rates were either low (<50%) or incalculable, with respondents more likely than non-respondents to hold strong attitudes. Priming effects in how AI was introduced and questions were worded may have biased some responses. Finally, responses in some studies appeared internally inconsistent in that, for example, radiology residents and students acknowledged AI would improve their discipline and wanted more AI training, but, at the same time, feared loss of professional status and held concerns about career choice.

Individuals without direct experience of AI who perceived it in the abstract tended to be more guarded in their views compared with the more optimistic views of direct users or recipients of AI. However, this optimism was more often grounded in views of workflow improvements and error minimisation, rather than perceptions of improved clinical outcomes, greater fairness of access or less risk to patient autonomy compared with current clinical practice. All stakeholders voiced concern about potential harm to patients from AI that lacks human oversight in its design, development and deployment, that the expected benefits of AI were by no means guaranteed, and that explicit regulatory standards must be formulated.

Applications which automate image interpretation and data synthesis were regarded more favourably by clinicians than those directly influencing clinical decisions or having potential to negatively impact clinician–patient relationships or clinician autonomy. Repetitive tasks using digitised data, such as radiological or dermatological diagnosis, are seen as more amenable to being performed by AI applications than interactive or procedural tasks such as consultations or surgical operations.[35] Privacy breaches and inability to understand or control AI applications worried clinicians, while loss of clinician oversight and inability to properly share in decision-making worried consumers. There was a common desire to ensure humans remained at the centre of decision-making and preserve

empathetic, contextualised communication in clinical encounters.[36] Case studies have confirmed consumers prefer human advisers who can appreciate their unique circumstances, and see AI assisting, rather than replacing, clinician advice.[37]

All stakeholders wanted reassurance that AI-generated advice was trustworthy, and that this level of trust was context-dependent, with clinician opinion trumping AI advice where the two were discordant or where decisions relating to serious illness were being made. As others have also shown,[38] stakeholders tend to be less forgiving towards error made by AI than error made by humans. Who should bear liability for error was much more contentious, both between and within stakeholder groups, and subject to considerable ongoing debate.[39] In a very recent US survey study of 750 physicians and 1007 members of the public, the majority of both groups believed the physician should be held responsible for AI error, although more of the public held this view than did physicians (66% vs 57%; p=0.02).[40] In contrast, more physicians believed the AI vendor (44% vs 33%; p=0.004) should share liability, while equal proportions of both groups conferred liability on regulatory authorities (23% vs 23%) or healthcare organisations purchasing the application (29% vs 23%).

Despite their reservations, clinicians overall were keen to receive further education in AI in recognition of its potential to increase diagnostic accuracy and workflow efficiencies, and this need is increasingly recognised.[41] While some clinicians in imaging specialties were worried about potential negative impacts on job prospects and professional status, most clinicians felt AI could enhance professional satisfaction.

Perceptions and expectations

Understanding what drives stakeholder perceptions of AI is important as they critically influence predisposition towards accepting AI.[42] Further in-depth research into why differing views of AI are held should assist in formulating operational solutions that accommodate such diversity of views. We note few studies considered the extent to which age, sex, clinical setting, level of expertise in computing or mathematics, personal beliefs and values, or other attributes of individuals impacted on their perceptions of AI in healthcare, which some investigators suggest as being important.[43]

Notwithstanding these considerations, certain expectations were inherent to many studies from which dependencies can be defined. While these dependencies are not necessarily unique to AI applications, being relevant to other computer-based technologies, the rapid evolution and potentially huge scope of AI magnifies the imperative for these dependencies to be enshrined in governance and ethics policies of government and industry.

CONCLUSION

A wide range of stakeholders have interest in how AI applications can be used in delivering better healthcare. In general, attitudes towards AI are positive, provided certain safeguards are met. While some concerns about AI are common to most groups, others are unique to a more select few. The challenge for AI developers and implementers is to understand these various concerns and respond appropriately if their applications are to be adopted at scale.

Contributors IAS conceived the study design, undertook initial literature search and data analysis and wrote first draft of the manuscript. SC critically appraised the manuscript and contributed to additional text and table 2. EC critically appraised the manuscript and provided further text under Methods.

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

Competing interests None declared.

Patient consent for publication Not applicable.

Provenance and peer review Not commissioned; internally peer reviewed.

Data availability statement All data relevant to the study are included in the article or uploaded as supplementary information. Not applicable.

ORCID iDs
Ian A Scott http://orcid.org/0000-0002-7596-0837
Enrico Coiera http://orcid.org/0000-0002-6444-6584

REFERENCES

1 Rajkomar A, Dean J, Kohane I. Machine learning in medicine. *N Engl J Med* 2019;380:1347–58.
2 Shaw J, Rudzicz F, Jamieson T, *et al*. Artificial intelligence and the implementation challenge. *J Med Internet Res* 2019;21:e13659.
3 Scheetz J, Rothschild P, McGuinness M, *et al*. A survey of clinicians on the use of artificial intelligence in ophthalmology, dermatology, radiology and radiation oncology. *Sci Rep* 2021;11:5193.
4 Coppola F, Faggioni L, Regge D, *et al*. Artificial intelligence: radiologists' expectations and opinions gleaned from a nationwide online survey. *Radiol Med* 2021;126:63–71.
5 Waymel Q, Badr S, Demondion X, *et al*. Impact of the rise of artificial intelligence in radiology: what do radiologists think? *Diagn Interv Imaging* 2019;100:327–36.
6 Collado-Mesa F, Alvarez E, Arheart K. The role of artificial intelligence in diagnostic radiology: a survey at a single radiology residency training program. *J Am Coll Radiol* 2018;15:1753–7.
7 Sit C, Srinivasan R, Amlani A, *et al*. Attitudes and perceptions of UK medical students towards artificial intelligence and radiology: a multicentre survey. *Insights Imaging* 2020;11:14.
8 Pinto Dos Santos D, Giese D, Brodehl S, *et al*. Medical students' attitude towards artificial intelligence: a multicentre survey. *Eur Radiol* 2019;29:1640–6.
9 Gong B, Nugent JP, Guest W, *et al*. Influence of artificial intelligence on Canadian medical students' preference for radiology specialty: ANational survey study. *Acad Radiol* 2019;26:566–77.
10 Sarwar S, Dent A, Faust K, *et al*. Physician perspectives on integration of artificial intelligence into diagnostic pathology. *NPJ Digit Med* 2019;2:28.
11 Polesie S, Gillstedt M, Kittler H, *et al*. Attitudes towards artificial intelligence within dermatology: an international online survey. *Br J Dermatol* 2020;183:159–61.

12 Ardon O, Schmidt RL. Clinical laboratory employees' attitudes toward artificial intelligence. *Lab Med* 2020;51:649–54.
13 Blease C, Kaptchuk TJ, Bernstein MH, *et al*. Artificial intelligence and the future of primary care: exploratory qualitative study of UK general practitioners' views. *J Med Internet Res* 2019;21:e12802.
14 Doraiswamy PM, Blease C, Bodner K. Artificial intelligence and the future of psychiatry: insights from a global physician survey. *Artif Intell Med* 2020;102:101753.
15 Oh S, Kim JH, Choi S-W, *et al*. Physician confidence in artificial intelligence: an online mobile survey. *J Med Internet Res* 2019;21:e12422.
16 Fan W, Liu J, Zhu S, *et al*. Investigating the impacting factors for the healthcare professionals to adopt artificial intelligence-based medical diagnosis support system (AIMDSS). *Ann Oper Res* 2020;294:567–92.
17 Castagno S, Khalifa M. Perceptions of artificial intelligence among healthcare staff: a qualitative survey study. *Front Artif Intell* 2020;3:578983.
18 Abdullah R, Fakieh B. Health care employees' perceptions of the use of artificial intelligence applications: survey study. *J Med Internet Res* 2020;22:e17620.
19 Esmaeilzadeh P. Use of AI-based tools for healthcare purposes: a survey study from consumers' perspectives. *BMC Med Inform Decis Mak* 2020;20:170.
20 Pega Inc. What consumers really think about AI: a global study. Insights into the minds of consumers to help Businesses reshape their customer engagement strategies. Cambridge, MA: Pega, 2017. Available: https://www.ciosummits.com/what-consumers-really-think-about-ai.pdf
21 Lennartz S, Dratsch T, Zopfs D, *et al*. Use and control of artificial intelligence in patients across the medical workflow: single-center questionnaire study of patient perspectives. *J Med Internet Res* 2021;23:e24221.
22 Palmisciano P, Jamjoom AAB, Taylor D, *et al*. Attitudes of patients and their relatives toward artificial intelligence in neurosurgery. *World Neurosurg* 2020;138:e627–33.
23 Tran V-T, Riveros C, Ravaud P. Patients' views of wearable devices and AI in healthcare: findings from the ComPaRe e-cohort. *NPJ Digit Med* 2019;2:1–8.
24 Sisk BA, Antes AL, Burrous S, *et al*. Parental attitudes toward artificial intelligence-driven precision medicine technologies in pediatric healthcare. *Children* 2020;7:145.
25 Nelson CA, Pérez-Chada LM, Creadore A, *et al*. Patient perspectives on the use of artificial intelligence for skin cancer screening: a qualitative study. *JAMA Dermatol* 2020;156:501–12.
26 McCradden MD, Baba A, Saha A, *et al*. Ethical concerns around use of artificial intelligence in health care research from the perspective of patients with meningioma, caregivers and health care providers: a qualitative study. *CMAJ Open* 2020;8:E90–5.
27 Consulting A. *Injecting intelligence into healthcare. Accenture executive survey on AI in healthcare Australia*. Australia: Accenture, 2018. https://www.accenture.com/_acnmedia/PDF-93/Accenture-AI-Health-Survey-Au.pdf#zoom=50
28 Jungmann F, Jorg T, Hahn F, *et al*. Attitudes toward artificial intelligence among radiologists, IT specialists, and industry. *Acad Radiol* 2021;28:834–40.
29 Laï M-C, Brian M, Mamzer M-F. Perceptions of artificial intelligence in healthcare: findings from a qualitative survey study among actors in France. *J Transl Med* 2020;18:14.
30 Eysenbach G. Improving the quality of web surveys: the checklist for reporting results of Internet E-Surveys (cherries). *J Med Internet Res* 2004;6:e34.
31 Perlmutter A, Benchoufi M, Ravaud P, *et al*. Identification of patient perceptions that can affect the uptake of interventions using biometric monitoring devices: systematic review of randomized controlled trials. *J Med Internet Res* 2020;22:e18986.
32 Matheny M, Israni ST, Ahmed M. Artificial intelligence in health care: the hope, the hype, the promise, the peril. Available: https://nam.edu/wp-content/uploads/2019/12/AI-in-Health-Care-PREPUB-FINAL.pdf [Accessed 10 Jul 2021].
33 WHO. *Ethics and governance of artificial intelligence for health: WHO guidance*. Geneva: World Health Organisation, 2021.
34 Gerke S, Babic B, Evgeniou T, *et al*. The need for a system view to regulate artificial intelligence/machine learning-based software as medical device. *NPJ Digit Med* 2020;3:53.
35 Mesko B. The impact of digital health technologies on the future of medical specialties in one infographic. the medical futurist. Available: https:// medicalfuturist.com/towards-creativity-in-healthcare-the-impact-of-dig ital-technologies-on-medical-specialties-in-an-infographic?utm_source=The%20Medical%20 Futurist %20Newsletter&utm_campaign=b07238 ce29-EMAIL_ CAMPAIGN_2019_08_12&utm_medium=email&utm _term=0_ efd6a3cd08-b07238ce29-420645589 [Accessed 10 Jul 2021].
36 Mittelman M, Markham S, Taylor M. Patient commentary: stop hyping artificial intelligence-patients will always need human doctors. *BMJ* 2018;363:k4669.
37 Longoni C, Bonezzi A, Morewedge CK. Resistance to medical artificial intelligence. *J Consum Res* 2019;46:629–50.
38 Dietvorst BJ, Simmons JP, Massey C. Algorithm aversion: people erroneously avoid algorithms after seeing them err. *J Exp Psychol Gen* 2015;144:114–26.
39 Price WN, Gerke S, Cohen IG. Potential liability for physicians using artificial intelligence. *JAMA* 2019;322:1765–6.
40 Khullar D, Casalino LP, Qian Y, *et al*. Public vs physician views of liability for artificial intelligence in health care. *J Am Med Inform Assoc* 2021;28:1574–7.
41 Rampton V, Mittelman M, Goldhahn J. Implications of artificial intelligence for medical education. *Lancet Digit Health* 2020;2:e111–2.
42 Lee MK. Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data Soc* 2018;5:205395171875668.
43 Thurman N, Moeller J, Helberger N. My friends, editors, algorithms, and I. *Digit J*2018.

BMJ Health & Care Informatics

Driving digital health transformation in hospitals: a formative qualitative evaluation of the English Global Digital Exemplar programme

Marta Krasuska,[1] Robin Williams,[2] Aziz Sheikh,[1] Bryony Franklin,[3] Susan Hinder,[2] Hung TheNguyen,[2] Wendy Lane,[4] Hajar Mozaffar,[5] Kathy Mason,[4] Sally Eason,[4] Henry Potts,[6] Kathrin Cresswell[1]

To cite: Krasuska M, Williams R, Sheikh A, *et al*. Driving digital health transformation in hospitals: a formative qualitative evaluation of the English Global Digital Exemplar programme. *BMJ Health Care Inform* 2021;**28**:e100429. doi:10.1136/bmjhci-2021-100429

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/bmjhci-2021-100429).

Received 11 June 2021
Accepted 22 November 2021

[1]Usher Institute, University of Edinburgh College of Medicine and Veterinary Medicine, Edinburgh, UK
[2]Institute for the Study of Science, Technology and Innovation, The University of Edinburgh School of Social and Political Science, Edinburgh, UK
[3]School of Pharmacy, University College London, London, UK
[4]NHS Arden and GEM CSU, Warwick, UK
[5]Business School, The University of Edinburgh, Edinburgh, UK
[6]Institute of Health Informatics, University College London, London, UK

Correspondence to
Dr Kathrin Cresswell;
kathrin.cresswell@ed.ac.uk

ABSTRACT

Background There is currently a strong drive internationally towards creating digitally advanced healthcare systems through coordinated efforts at a national level. The English Global Digital Exemplar (GDE) programme is a large-scale national health information technology change programme aiming to promote digitally-enabled transformation in secondary healthcare provider organisations by supporting relatively digitally mature provider organisations to become international centres of excellence.

Aim To qualitatively evaluate the impact of the GDE programme in promoting digital transformation in provider organisations that took part in the programme.

Methods We conducted a series of in-depth case studies in 12 purposively selected provider organisations and a further 24 wider case studies of the remaining organisations participating in the GDE programme. Data collected included 628 interviews, non-participant observations of 190 meetings and workshops and analysis of 9 documents. We used thematic analysis aided by NVivo software and drew on sociotechnical theory to analyse the data.

Results We found the GDE programme accelerated digital transformation within participating provider organisations. This acceleration was triggered by: (1) dedicated funding and the associated requirement for matched internal funding, which in turn helped to prioritise digital transformation locally; (2) governance requirements put in place by the programme that helped strengthen existing local governance and project management structures and supported the emergence of a cadre of clinical health informatics leaders locally; and (3) reputational benefits associated with being recognised as a centre of digital excellence, which facilitated organisational buy-in for digital transformation and increased negotiating power with vendors.

Conclusion The GDE programme has been successful in accelerating digital transformation in participating provider organisations. Large-scale digital transformation programmes in healthcare can stimulate local progress through protected funding, putting in place governance structures and leveraging reputational benefits for participating provider organisations, around a coherent vision of transformation.

Summary

What is already known?

► There is currently a strong drive internationally towards creating digitally advanced healthcare systems through coordinated efforts at a national level but there is lack of knowledge on how to stimulate large-scale digitalisation.

What does this paper add?

► Large-scale digital transformation programmes in healthcare can stimulate local progress through protected funding, putting in place governance structures, and leveraging reputational benefits for participating provider organisations, around a coherent vision of transformation.

INTRODUCTION

There is currently a strong international drive towards creating digitally-enabled health systems and settings, with governments embarking on large-scale health information technology (HIT) change initiatives to improve quality, safety and efficiency of health and care.[1 2] For example, in the USA, the Health Information Technology for Economic and Clinical Health (HITECH) initiative launched in 2009 combined over US$25.9 billion of central funding with development of a national set of standards for implementation of electronic health records (EHRs) to stimulate digital transformation of provider organisations.[3] The German federal government's 2020 Hospital Future Act committed over €3 billion across a 2-year period to stimulate digital transformation of hospitals.[4] Government-led, large-scale HIT change programmes have also recently been initiated in Canada,[5] Australia[6] and New Zealand.[7] However, historically, such national programmes often have failed

Box 1 Overview of the Global Digital Exemplar programme

The Global Digital Exemplar (GDE) programme is a large-scale health information technology (HIT) change programme launched by National Health Service (NHS) England aiming to stimulate the digital transformation of the English healthcare system. It had a total budget of over £385 million of central funding, a 5-year duration (2017–2021) and involvement of 51 individual provider organisations.

The GDE programme was introduced in the aftermath of the English National Programme for IT—the largest national digitalisation programme worldwide with a budget of over £9.8 billion,[8] which was discontinued in 2012 following a brief period of relatively uncoordinated digital transformation attempts across the healthcare system.

The key strategy of the GDE programme, led by NHS England, was to stimulate digital transformation across English NHS healthcare providers and to form a central point for facilitating knowledge creation by creating 'Global Digital Exemplars' (GDEs)—local centres of digital excellence that could serve as examples of best practice.

Provider organisations were selected to become GDEs, based on their relatively high levels of digital maturity (the extent to which organisations had digitally-enabled processes) and capability to undertake an innovative digital transformation programme. Each GDE provider organisation signed a funding agreement with NHS England to implement a detailed portfolio of HIT change projects over a period of 2–3.5 years and received £5–10 million of central funding (which had to be matched with the same level of internal funding). Additionally, GDE provider organisations were paired with one (and in two cases two) partner providers—referred to as Fast Followers (FFs). The FFs were not expected to be as digitally mature as their partner GDEs but to be sufficiently mature to be able to rapidly accelerate their digital transformation through knowledge transfer from their partner. The FFs were also asked to prepare a portfolio of digital transformation projects to be carried out during this period. FFs received half of the central funding that the GDE organisations received (ie, £5 million), which again had to be locally matched with the same amount. Twenty-three provider organisations took part as GDEs and 25 as FFs. All participating organisations were asked to establish a senior clinical digital leadership role in the form of a Chief Clinical Information Officer ahead of the start of the programme.

The Healthcare Information and Management Systems Society (HIMSS) Electronic Medical Record Adoption Model classification tracking hospitals' levels of digital maturity on a scale from Level 0 to 7,[14] was used as a benchmark for digital excellence in the programme. Acute GDEs were expected to achieve HIMSS Level 6 with a view to 7 and mental health GDEs and FFs Level 5 by the end of the programme.

In addition, the GDE programme supported coordinated learning including setting up learning networks for staff in participating organisations, organising networking events and other knowledge transfer activities including the production/circulation of Blueprints (documents capturing learning in implementing particular changes).[15]

to realise their ambitious digitalisation goals. For example, in England, the National Programme for Information Technology (NPfIT)—the largest ever national digitalisation programme with an initial budget of over £9.8 billion[8 9]—was discontinued in 2012 as it was perceived to not sufficiently cater for the needs of implementing organisations.[10] The relative lack of success of many nationally-led, large-scale HIT change programmes may be attributed to limited current understanding of how such programmes work to help promote digital transformation locally.[11] There is therefore now a growing need for evidence on how best to stimulate digital transformation of healthcare systems and settings through these kinds of initiatives.

To address this gap, we here present findings from an independent, formative evaluation of the Global Digital Exemplar (GDE) programme—a flagship, national HIT change initiative aiming to stimulate digitalisation of English hospitals through creating a cohort of provider organisations that would act as exemplars of digital excellence (box 1).[12] The programme was developed in response to an independent review that drew lessons from previous digital transformation initiatives in the UK and the USA.[13] Given that funding available was not sufficient to allow all provider organisations to fully digitalise, this strategy adopted a phased approach with funding initially allocated to relatively digitally mature organisations. These were paired up with less mature partner organisations, with whom they were encouraged to share knowledge and thereby accelerate digitalisation. We aimed to address the following research question: How did the GDE programme promote digital transformation in participating provider organisations?

METHODS

We undertook a longitudinal qualitative study of the GDE programme that aimed to explore digital transformation in participating provider organisations and the wider healthcare system.[16] Our work had both formative and summative elements, but its defining characteristic was its formative nature, feeding back emerging findings to decision-makers and thereby shaping delivery of the programme.

The detailed methodology is described in a separate published research protocol and in Appendix 1.[17] The evaluation took place between January 2018 and March 2021. We followed the Consolidated Criteria for Reporting Qualitative Research in this paper.[18] Formative, qualitative evaluations, conducted in real-time alongside change programmes, can help to explore the processes involved in seeking to stimulate digital transformation and can thereby inform future initiatives.[19 20] This type of evaluation collects evidence on the processes involved in stimulating digital transformation through HIT change initiatives and on an array of emerging outcomes including consequences not anticipated/intended by programme architects. Such formative evaluations are well placed to inform decision-makers during the programme that is being evaluated.

We conducted 628 interviews, observed 190 meetings and analysed 499 documents (see box 2 for an overview of the data set and Appendix 2 for a detailed description). This included an additional round of interviews performed in autumn 2020 in relation to the impact of COVID-19 on digital transformation. Interviews lasted 1 hour on average.

Box 2 Description of sample (GDE, Global Digital Exemplar; FF, Fast Follower)

In-depth case study sites (12 provider organisations; 8 GDEs: 6 acute, 2 mental health; 4 FFs: 3 acute, 1 specialist)

- 309 interviews (39 senior managers; 65 clinical digital leaders; 47 non-clinical digital leaders; 46 GDE programme staff; 112 operational staff).
- 104 documents.
- 67 meetings observed.

Interview periods:

- Pilot interview: March 2018 (1 interview)
  - 1 GDE programme staff.
- First round: May 2018 – February 2019 (137 interviews)
  - 16 senior managers.
  - 20 clinical digital leaders.
  - 11 non-clinical digital leaders.
  - 14 GDE programme staff.
  - 76 operational staff.
- Second round: March 2019 – May 2019 (34 interviews)
  - 6 senior managers.
  - 10 clinical digital leaders.
  - 3 non-clinical digital leaders.
  - 12 GDE programme staff.
  - 3 operational staff.
- Third round: June 2019 – March 2020 (101 interviews)
  - 11 senior managers.
  - 27 clinical digital leaders.
  - 26 non-clinical digital leaders.
  - 10 GDE programme staff.
  - 27 operational staff.
- Fourth round: August 2020 – December 2020 (post-lockdown) (36 interviews)
  - 6 senior managers.
  - 8 clinical digital leaders.
  - 7 non-clinical digital leaders.
  - 9 GDE programme staff.
  - 6 operational staff.

Broader case study sites (24 provider organisations; 15 GDEs: 10 acute, 5 mental health; 9 acute FFs)

- 247 interviews (32 senior managers; 78 clinical digital leaders; 65 non-clinical digital leaders; 44 GDE programme staff; 28 operational staff).
- 283 documents.
- 19 meetings observed.

Interview periods:

- First round: 2018 (95 interviews).
- Second round: 2019 (69 interviews).
- Third round: 2020 (83 interviews).

Other data

- 72 interviews (61 policymakers; 3 vendors; 4 engagement leads and 4 other stakeholders).
- Non-participant observations of 104 national meetings, workshops and conferences.
- 112 documents.

Interview periods:

- First round: March 2018 – December 2018 (31 interviews).
- Second round: January 2019 – November 2019 (20 interviews).
- Third round: January 2020 – April 2020 (3 interviews).
- Fourth round: July 2020 – February 2021 (18 interviews).

RESULTS

Our analysis identified several sociotechnical dimensions associated with digital transformation. Many of these have already been extensively discussed in the literature (table 1) and we therefore focus here on exploring novel findings surrounding the wider macro-environmental dimensions associated with the GDE programme.

The impact of the GDE programme in stimulating digital transformation locally is described in Appendix 3.

Earmarked funding stimulated digital transformation locally

Dedicated funding over a multiyear period, comprising both external funding (allocated from a central national budget) and matched funding from the provider organisation's internal budget, was perceived to play a key role in accelerating digital transformation. Funding was used to support and bring forward major upgrades in digital information infrastructures (including renewing core EHR systems) together with a range of smaller-scale digital change projects such as implementation of electronic clinical observations systems or projects to support staff working remotely in the community. Many organisations reported that plans for these changes were already in place prior to the launch of the programme.

> It enabled us to do things, because of the money, it enabled us to do things, that we would have done anyway, at twice the speed, (…) but there is something about scale and there is something about speed, which brings a value that is greater than achieving it in twice the time. (Site D, GDE, in-depth case study, GDE programme staff)

Table 1 Findings associated with sociotechnical dimensions of change confirming previous findings in the empirical literature[20 28–44]

	Dimensions
Technological factors	System usability, system performance, adaptability and flexibility, system dependability, availability of data, integrity and confidentiality, data accuracy, sustainability.
Social factors	User satisfaction, complete/correct use, attitudes and expectations, user engagement, experiences of Information Technology use, workload implications and benefits of system use, impact of system on existing work processes, user input in design.
Organisational factors	Leadership and management, communication with stakeholders, implementation timelines, vision associated with system, training and user support, system champions implementation/optimisation resources, monitoring of progress and system optimisation.

The scope to secure external funding combined with a requirement for matched funding, also helped to secure local leadership buy-in and support.

> [Central NHS funding through the GDE Programme] was enough money to make a case to our finance director and the acting chief executives that we should do it [GDE Programme], because it was money we wouldn't get otherwise, for a thing we wanted to do anyway. (Site G, GDE, in-depth case study, clinical digital leader)

Protected funding was especially important in driving digital transformation for smaller provider organisations with correspondingly smaller internal budgets. For the largest organisations, external GDE programme funding was modest in relation to their overall digital investments. In particular, some of the large provider organisations had substantial development capabilities and large technology budgets that had allowed them, in some cases, to begin planning and implementing comprehensive digital change, meaning that they had already achieved a certain momentum ahead of the programme. As a result, participating in the programme strengthened but did not per se transform the digital strategies and capabilities of these organisations in the dramatic way that could be observed in smaller and less digitally mature providers (which in many cases included Fast Followers (FFs)). Provider organisations described this support as accelerating the rate of change but not radically changing the direction of their prior digital journey. They were able to achieve more because of these additional resources.

> My reflection on the GDE process is that I don't think we would have done this without it. I think we always wanted to do it and it gave us the opportunity to do what we wanted to do anyway but we would not have been able to employ this people, we would not have been able to pay [Supplier] to deliver the extra functionality, we wound not have been able to pay me for two years to provide some clinical input. (Site G, GDE, in-depth case study, clinical digital leader)

This momentum and ambition for change grew as a result and continued beyond the end of the programme.

> So it has focused…just by the injection of money rather than anything else, the money has enabled us to buy products which when you start delivering them, you then can't really stop, so although the £10m isn't enough, it's now made it an issue that we benefit from this if we did a bit more and we spent a bit more. (Site I, GDE, in-depth case study, clinical digital leader)

Provider organisations perceived that the provision of national support primarily through capital funding, as opposed to revenue funding, affected local digital transformation initiatives, as it promoted investment in purchasing hardware and software. The administrative complexity of converting capital funding into revenue streams meant that investing in staff and third-party services to maintain, service, support, upgrade and optimise systems was somewhat inhibited.

Prestige and reputational benefits helped to secure organisational buy-in and to negotiate with suppliers

The prestige and reputational benefits obtained through taking part in a flagship national HIT change programme and competing for the status of being a 'Global Digital Exemplar', were instrumental in securing leadership buy-in and also helped to secure wider organisational support for digital transformation efforts. Although some of the organisations participating in the programme already considered themselves as national leaders, being a GDE involved projecting a claim not only of being nationally excellent but also of attaining internationally recognised standards of excellence. Other national programmes had not specifically targeted this already high-achieving segment of provider organisations. Those who were already 'high-achievers' were keen to be seen as international leaders and others saw this as putting their organisations into the limelight.

In many cases, the 'Global Digital Exemplar' badge had been used to communicate the upcoming HIT change projects (eg, EHR upgrades, or implementation of electronic observations) across the organisation, for example, through posters and newsletters.

> [The GDE Programme and its agenda] was helpful both from a reputation and to badge it all in a concept of…it gave people a…rallying cry around our direction of travel. (Site 12, FF, broader case study, clinical digital leader)

The benefits of enhanced national visibility and status from participation in the Programme were less evident for organisations with a strong prior national or international profile (including many FFs). Smaller provider organisations with modest local profiles reported that taking part in the programme allowed them to be more visible and recognised locally.

> Reputationally, we're considered regionally as digitally mature, and that's quite a battle to fight. Not necessarily with other mental health or community trusts but certainly with the larger acutes [acute care provider organisations],… you kind of have to earn your place. You do have to earn your place around the table and some of the things that we've done in GDE have enabled us, to use a very common expression at the moment, a more sort of level playing field. (Site E, FF, in-depth case study, GDE programme staff)

Provider organisations further noted that the status associated with the programme increased their negotiating power with vendors. Large provider organisations (mainly GDEs) that were recognised nationally and internationally as leading centres were often invited to become reference sites for certain product implementations and thereby secured allocation of additional resources from vendors. Smaller, less prestigious provider organisations

Box 3 Limitations of our work

Our findings on the digital transformation outcomes associated with the Global Digital Exemplar (GDE) programme should be interpreted with caution. Intended and unanticipated consequences were still emerging at the end of our evaluation work. Attribution of outcomes in large-scale digital transformation initiatives is not straightforward, as interventions are often multifaceted, stimulating digitalisation through a combination of enhancements in technological systems and organisational processes—as a result, outcomes take a long time to materialise and may not then be directly attributable to HIT.[45] In addition, large-scale change programmes are situated within evolving wider policy and economic settings that may influence outcomes. Various local factors are also likely to have an impact. To address these complexities, our evaluation used a combination of in-depth case studies that allowed for detailed understanding of how the programme unfolded in a range of specific settings and wider case studies of other providers that involved capturing broader patterns and verifying findings from the in-depth case studies. Further, each participating provider organisation proposed a portfolio of digital innovations as part of the programme. Unfortunately, our methodology did not allow us to systematically appraise individual innovations and outcomes. However, a wide range outcomes were reported including many not initially anticipated improvements that were coming to the fore at the end of the evaluation period, sometimes in areas that were not directly related to the original area of implementation (eg, in shared care records across settings).

We focused largely on the perspective of provider organisations, particularly local GDE programme managers and implementers. As a result, perspectives of individual healthcare staff within provider organisations are underrepresented.

(including many mental health providers and FFs) in contrast often found themselves competing over vendor resources with other customers including other provider organisations taking part in the programme.

> I think if you speak to our finance director… he would say it's the [vendor] relationship that's the most valuable part of the GDE… being part of the GDE process, he thinks, gives him much more leverage with [vendor] to actually deliver what they've promised. Cause quite frankly, if they don't deliver it with us, then they won't be able to sell to other organisations, 'cause we will be their site, where everyone will come and see all their solutions together. (Site I, GDE, in-depth case study, GDE programme staff)

Being labelled a 'Fast Follower' offered lower perceived status benefits than GDE. Some FF organisations felt that they were in some respects more advanced than their GDE and should therefore be labelled 'partners' instead of 'followers'.

Governance requirements supported establishment of project management structures, secured executive buy-in and strengthened clinical digital transformation leadership

The funding agreement between provider organisations and the central funding body contained contractual obligations, which included the organisations' digital strategy and an outline of HIT projects to be undertaken with timescales, funding milestones and a Statement of Planned Benefits. Provider organisations were thus required to prepare and then execute a portfolio of HIT change projects in a relatively short period. Further, although not a formal obligation, there was also an expectation for the provider organisations to set up a local GDE Programme Board to oversee deployment of the programme locally. These in turn supported the creation and expansion of change management and engagement structures within provider organisations to support the implementation of the HIT change projects outlined in the funding agreement. The requirement to meet the milestones set out in the funding agreement, combined with well-depicted digital transformation goals, helped to secure executive support and helped to make the transformation agenda more salient at the executive level.

> I think one of the main parts that was really effective is the pace-setting element of the GDE. […] The pace setting as part of the Programme was a massive part of achievements. And I think the reason for that is it really focuses the board. Because you have essentially money attached to a deadline to achieve something, that's extremely motivating. And in trusts where you have so many competing priorities […] I thought was very effective actually that we had to hit certain milestones with good quality and that then funding would be achieved. And I think that really helped focus the board. And because of that, we had a really, I think, strong functioning Digital Oversight Committee through the Programme and that's one of the things that kept the momentum going. (Site 10, GDE, broader study, clinical digital leader)

Provider organisations were required to report regularly on implementation progress and benefits achieved to the central funding body. However the reporting methods were perceived as burdensome, particularly as these reports were not always aligned with the reports that provider organisations had to submit to their own boards and for other health service reporting systems. Provider organisations reported that the burden of reporting diverted efforts from other key activities related to digital transformation. Although there was an attempt to simplify central reporting procedures as the programme progressed, with the adoption and refinement of a computerised reporting tool, little progress was made in harmonising reporting requirements among different parts of the health service (which had different established report requirements, deadlines and reporting periods). Another issue was that, although the funding agreements laid out a timetable of contractual commitments, over time as the programme progressed, context, technologies and local priorities changed. Some provider organisations had trouble in meeting the contractual obligations and milestones, given the dynamism and uncertainties surrounding digital transformation, and highlighted the rigidity of funding agreements. Although

Box 4 Organisational characteristics associated with digital maturity

- Leadership focus on digitally-enabled transformation of services (rather than merely Information Technology deployment).
- Digital transformation expertise at Board level.
- Clinical engagement and dedicated intermediary roles between clinical and digital areas.
- Activity surrounding envisioning benefits/targets and measuring progress.
- Demonstrating benefits for individual users early on in the process.
- Strong and experienced project management structures dedicated to digital transformation.
- Willingness to share experiences and learn from others.
- Open and transparent decision-making and communication across the organisation.
- A conceptualisation of digital maturity as a continuous quality improvement process.

it was possible to renegotiate funding agreements, this process was seen as slow and time-consuming.

> Yes, we can set milestones for 6 months or 12 months but trying to set a milestone for three years' time when IT changes, the organisation changes so quickly. (Site D, GDE, in-depth case study, non-clinical digital leader)

Another aspect of centrally introduced governance requirements was a mandatory requirement to appoint a Chief Clinical Information Officer (CCIO) ahead of the programme—a senior leadership role within provider organisations combining clinical and digital transformation expertise. This requirement was critical in helping organisations to build capacity to manage and lead digital transformation projects. The CCIOs also had a major role in securing and enhancing clinical engagement in the digital transformation process and in selecting and configuring the technological systems to ensure they would be fit for purpose in the clinical context. Further, they contributed to raising the awareness and priority of the digital transformation agenda within senior leadership. The appointment of a CCIO further promoted the creation of a number of related senior leadership positions combing clinical and digital expertise such as a Chief Nursing Information Officer (CNIO), Chief Medical Information Officer (CMIO) and deputy CCIOs responsible for specific subdisciplines (eg, cardiology, oncology).

> We wouldn't have had CCIOs [Chief Clinical Information Officers] if we weren't a GDE really, I think the GDE opportunity coalesced in the IT department which was very IT-driven to actually, well, we need to engage clinicians in this, otherwise we won't get this money [from the GDE Programme], we've got to show that we've got clinical involvement. (Site I, GDE, in-depth case study, clinical digital leader)

The strengthening of digital informatics capabilities was reinforced by related changes in the whole sector including the establishment of the NHS Digital Academy—an NHS training programme that aimed to develop a new generation of clinical digital leaders to drive digital transformation.

> '…going through… the Digital Academy has really helped in this kind of difficult phase where you're looking at projects, programmes, organising, whole organisations around it. I mean I'm falling back on some of the stuff we did there quite a bit now actually and I was, I realise how inexperienced we were when we started.' (Site E, GDE, in-depth case study, clinical digital leader)

Finally, as part of the GDE requirements, participating provider organisations were expected to achieve high levels of performance under the Healthcare Information and Management Systems Society (HIMSS) Electronic Medical Record Adoption Model (EMRAM). Their ability to meet these ambitious targets within the relatively short timeframes of the GDE programme was greatly influenced by their choice of supplier. Some (US) vendors that had recently entered the UK market offered comprehensive 'mega-suites' already well-aligned with the wide range of functionality required to meet the HIMSS EMRAM accreditation criteria. Many GDE providers turned to these solutions in order to meet the ambitious aims of the programme. Other EHR adopters that stayed with their existing EHR supplier sought to bridge the gap by asking their vendor to extend their range of functionality or by procuring and integrating modules from other suppliers (a strategy labelled 'Best-Of-Breed'). These provider organisations and their suppliers thereby embarked on an unpredictable journey that posed challenges for both sides. Some suppliers struggled to deliver the new functionalities required within the timeframe of the GDE programme. In addition, the growth in demand due to the programme was such that even some large suppliers were unable to provide the level of development support expected by individual provider organisations.

DISCUSSION

Summary of key findings

The GDE programme strategy of supporting relatively digitally mature healthcare provider organisations to become exemplars of digitally enabled transformation has resulted in rapid acceleration of transformation and promoted the visibility and priority of digital transformation plans in those organisations. The programme also contributed to the promotion of clinically focused digital change management capability and the emergence and strengthening of local clinical change leaders (ie, those planning and implementing local programmes, including CIOs, GDE programme managers, CNIOs, CMIOs and CCIOs). This has driven a visible culture shift among clinicians and leaders to a proactive expectation that

Table 2 Lessons for running digital transformation programmes

Reconciling national, regional and local priorities and functions	There is a need for strategic national goals while allowing local ownership and flexibility to tailor efforts to local needs. There is an ongoing discussion on which functions should be conducted regionally and which centrally and there are trade-offs with each approach that need to be considered. Some specialist functions may best be undertaken centrally (eg, oversight of markets), while some kinds of specialism may best be maintained by a system wide division of labour (eg, procurement) but could be done through a matrix of regionally located stakeholders. Other kinds of functions that require knowledge of local organisations and population demographics may best be done locally (eg, population health).
Digital transformation requires a long-term vision and support	In the GDE Programme, the long-term stable national vision was not clearly articulated from the start. It was unclear what defined a 'successful' GDE and what would happen when GDE status is achieved.
Digital transformation requires an understanding of the existing policy and organisational landscape (a birds eye perspective)	Clear understanding of the policy landscape and existing incentives and risks/costs and how these impact on different stakeholder groups is important when implementing digital change initiatives. Digitally enabled transformation requires a clear understanding is needed so that the change initiatives/programmes can make use of incentives and manage risks.
Digital transformation requires long-term funding and flexibility	Annualised budgets complicate long-term strategy. Additional funding for digital transformation is often only available for a year. There is an urgent need to address the problems of revenue funding. All digital projects have revenue implications in terms of both depreciation of the system purchased and in maintaining it. Many provider organisations find capital funding, traditionally available for 'equipment', constraining with the increasing salience of licencing and per user charges (software as a service model) thus digitalisation is essentially a revenue commitment. Changes in policy and priorities, and associated shifts in direction, were disruptive to those on the ground. A balance needs to be achieved between developing new initiatives and continuing earlier ones. National programme managers are acutely aware of this, but see these features as part of the political landscape that are unable to change, and therefore develop strategies/workarounds to manage and mitigate these instabilities.
Addressing the digital divide	The GDE programme has created beacons of excellence, but there is now a policy focus on levelling up digital maturity across organisations. There may be scope in twinning organisations (especially on the basis of co-location or common platforms) in a more structured way going forward building on the success of GDE/Fast Follower partnerships.

GDE, Global Digital Exemplar.

digital solutions underpin care delivery and enable transformation. There has also been a concomitant increase in engagement and capability in the general workforce as organisations increasingly digitalised their organisational processes. Earmarked funding, the strengthening of local governance structures, digital project management capability and the reputational benefits associated with being included in the GDE programme have helped to ensure buy-in for digital transformation plans from both senior managers and frontline staff. This ensured that what was delivered was digital transformation rather than simply a technology implementation programme. However, it is important to keep in mind that while the GDE programme support imparted momentum and direction, some provider organisations were already on this trajectory of change and during the programme followed local digital transformation strategies that were already planned.

Strengths and limitations

We conducted a national, longitudinal, formative evaluation of a first-of-a-kind large-scale HIT change initiative to advance digital transformation in the English National Health Service. We collected a large, qualitative data set from participating provider organisations and from national actors over extended timescales. This allowed gaining comprehensive insights into the mechanisms of change promoted through the GDE programme and associated outcomes. Detailed limitations of our work can be found in box 3.

Integration of the findings with existing literature

Previous findings surrounding the importance of sociotechnical dimensions of digitally-enabled change in the empirical literature have been confirmed in our work,[20 28–44] but we have uncovered some important issues surrounding macro-environmental dimensions of change

and how these can impact on technological, social and organisational dimensions. These include the role of wider incentives, prestige and governance requirements to stimulate local digitalisation efforts.

We found that the GDE programme, as a large-scale digital transformation initiative, accelerated digital transformation in selected digitally advanced sites. Key to success was a combination of dedicated resources, governance frameworks, local ownership and vision. It began with a national review that took stock of previous national experiences and sought to learn from them, actively involving national and international experts, and laying out a vision and steps towards achieving digitally-enabled transformation.[12] This stands in stark contrast to previous experiences in the NPfIT, which was, from the start, driven by an arguably unrealistic vision based on centralised procurement which created problems around technology choice and lack of organisational and clinical buy-in.[46]

The GDE programme allowed a new digital vision and we observed changes in staff attitudes towards digitalisation. This in turn facilitated staff engagement with digitally-enabled transformation activities. The impact of the programme was affected by the COVID-19 crisis that impeded organisational progress towards achieving HIMSS targets but which, by demonstrating the value of digital capabilities (notably in remote consultations), also encouraged more rapid uptake and acceptance and helped to accelerate digital transformation locally.

The GDE programme has also helped to reconcile tensions surrounding local input in decision-making with national direction. Key here was setting national goals and monitoring progress, while allowing a degree of local freedom over how to achieve these goals.[45] Experiences with other national initiatives reinforce the effectiveness of balancing goal-setting with local choice, a perspective that is supported by the notion of loose coupling where organisational subsystems function well if they can maintain their own identity and autonomy.[47–49]

The US HITECH initiative reinforces the important role of centrally allocated funding and goal setting in facilitating adoption.[50 51] However, although resulting in dramatic increased computerisation of healthcare, HITECH has also illustrated that rapid adoption and mandating use without the cultural changes needed to support transformation can create unrealistic expectations and disillusion frontline clinical staff, a consequence that only became apparent after the programme had concluded.[52] The emergence and strengthening of local clinical change leaders helping to promote clinical engagement and leadership-buy-in might help to mitigate risk.[53] Throughout this journey, HIMSS served as a roadmap, allowing implementing organisations to plan changes in small steps and allowing national programme managers to benchmark and monitor progress.[54] However, requiring providers to rapidly achieve particular benchmarks may restrict markets (favouring existing vendors whose products are already aligned with HIMSS EMRAM) and limit innovation as it leaves little room for experimentation and innovation around local priorities.[55]

Although characterising digital maturity was not the focus of this paper, these results, building also on existing literature and our previous work surrounding the definition of technological characteristics of digital excellence in hospitals,[56] serve as a starting point to identify organisational characteristics of digital excellence in hospitals (box 4).

Implications for policy and practice

In contrast to recent heavily-funded technology procurement programmes that failed to deliver,[2 8 52] the GDE programme has succeeded in promoting digital transformation across a significant tranche (20%) of provider organisations. The experience highlights how a coordinated approach with relatively modest funding can catalyse rapid and significant improvements in digital maturity in healthcare. At the time of writing (August 2021), four provider organisations had achieved HIMSS Level 6 and two had achieved HIMSS Level 7.[57]

Programme managers recognised that the most mature provider organisations (eg, those expected to meet targets in 2 years) had already begun their digital journey. Although a few organisations struggled to meet the ambitious programme goals, most achieved a substantial boost in terms of the pace and strategic direction of their digital transformation. In this sense the programme seems to have successfully targeted what welfare policymakers have described as the 'Goldilocks zone', minimising (wasteful) over servicing and (ineffective) underservicing.[58]

As this programme ends, there is a risk that the momentum created through the programme is lost. It is imperative to build on lessons learnt and exploit the valuable experience acquired in the programme through follow-on initiatives. Its immediate successor, the Digital Aspirant (DA) programme, currently underway in NHS England, addresses concerns that less mature providers might be left behind.[59] Less digitally mature organisations are likely to require more support.[60] Questions arise as to whether the DA programme will deliver similar successes to those seen in the GDE programme. The key drivers identified in this paper are somewhat weakened under the DA programme: Organisations participating in DA programme start with lower levels of digital maturity and will receive less funding than those that participated in the GDE programme. Programme governance arrangements are more limited than in the GDE programme, and some of the successful mechanisms to facilitate learning have not been carried forward (notably GDE/FF partnerships). The prestige associated with being a Digital Aspirant may also be lower. The policy agenda is however evolving. Having demonstrated an ability to create islands of excellence, the 2019 NHS Long Term Plan requires all providers to achieve a core level of digitalisation by 2024 to allow information exchange across regional ecosystems.[61] Future efforts should focus on strengthening learning networks in order to ensure that lessons learnt are effectively and widely disseminated

across the wider NHS. We have summarised the lessons for running digital transformation programmes emerging from our work in table 2.

CONCLUSIONS

The GDE programme helped to accelerate digital transformation in participating provider organisations and to establish the foundations for a digital health learning ecosystem. It appears to have achieved this through protected funding, putting in place governance structures and through harnessing reputational benefits for participating provider organisations. The GDE programme provides a template for successful digital transformation that was lacking after the failure of recent high profile heavily funded technology procurement programmes. It is now important that learning from this initiative is maximised in efforts to bridge the digital divide across provider organisations.

Acknowledgements We gratefully acknowledge the input of the participants and the Steering Group of this evaluation.

Contributors KC and RW conceived this paper. KC, MK and RW led the drafting of the manuscript and all authors commented on drafts of the manuscript. KC acts as the guarantor.

Funding This article has drawn on a programme of independent research supported by NHS England. The views expressed are those of the author(s) and not necessarily those of the NHS, NHS England or NHS Digital. This work was also supported by the National Institute for Health Research (NIHR) Imperial Patient Safety Translational Research Centre.

Disclaimer The views expressed in this publication are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health and Care.

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval The work received institutional ethical approval from the School of Social and Political Science at The University of Edinburgh, UK (no reference number provided). Participants gave informed consent to participate in the study before taking part.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available upon reasonable request. The raw data are available upon reasonable request from the corresponding author.

REFERENCES

1 Sheikh A, Sood HS, Bates DW. Leveraging health information technology to achieve the "triple aim" of healthcare reform. *J Am Med Inform Assoc* 2015;22:849–56.
2 Jha AK, Doolan D, Grandt D, *et al*. The use of health information technology in seven nations. *Int J Med Inform* 2008;77:848–54.
3 Washington V, DeSalvo K, Mostashari F, *et al*. The HITECH era and the path forward. *N Engl J Med* 2017;377:904–6.
4 Lovell T. German hospitals to get €3 billion funding boost for digitalisation: The Hospital Future Act aims to modernise the country's hospital system. In: *Healthcare IT News*, 2020.
5 Infoway CH. *A new day in healhcare is coming: annual report 2018-2019*, 2019.
6 Burton-Jones A, Akhlaghpour S, Ayre S. Changing the conversation on evaluating digital transformation in healthcare: insights from an institutional analysis. *Information and Organization* 2019;100255.
7 Bowden T, Coiera E. Comparing New Zealand's 'Middle Out' health information technology strategy with other OECD nations. *Int J Med Inform* 2013;82:e87–95.
8 Review of the final benefits statement for programmes previously managed under the National programme for it in the NHS. Available: https://www.nao.org.uk/wp-content/uploads/2013/06/10171-001_NPfiT_Review.pdf [Accessed 16 May 2021].
9 The challenges in implementing digital change. Available: https://www.nao.org.uk/wp-content/uploads/2021/07/The-challenges-in-implementing-digital-change.pdf [Accessed 14 Aug 2021].
10 Coiera EW. Lessons from the NHS National Programme for IT. *Med J Aust* 2007;186:3–4.
11 Greenhalgh T, Russell J, Ashcroft RE, *et al*. Why national eHealth programs need dead philosophers: Wittgensteinian reflections on policymakers' reluctance to learn from history. *Milbank Q* 2011;89:533–63.
12 NHS England. Global digital exemplars. Available: https://www.england.nhs.uk/digitaltechnology/connecteddigitalsystems/exemplars/ [Accessed 16 May 2021].
13 Wachter RM. *Making it work: harnessing the power of health information technology to improve care in England, report of the National Advisory group on health information technology in England*, 2016.
14 HIMSS continuity of care maturity model (CcmM). Available: https://www.himssanalytics.org/ccmm [Accessed 16 May 2021].
15 Williams R, Sheikh A, Franklin BD, *et al*. Using blueprints to promote interorganizational knowledge transfer in digital health initiatives-a qualitative exploration of a national change program in English hospitals. *J Am Med Inform Assoc* 2021;28:1431–9.
16 Fereday J, Muir-Cochrane E. Demonstrating rigor using thematic analysis: a hybrid approach of inductive and deductive coding and theme development. *Int J Qual Methods* 2006;5:80–92.
17 Cresswell K, Sheikh A, Dean Franklin B, *et al*. Formative independent evaluation of a digital change programme in the English National health service: study protocol for a longitudinal qualitative study. *BMJ Open* 2020;10:e041275.
18 Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. *Int J Qual Health Care* 2007;19:349–57.
19 Catwell L, Sheikh A. Evaluating eHealth interventions: the need for continuous systemic evaluation. *PLoS Med* 2009;6:e1000126.
20 Cresswell K, Williams R, Sheikh A. Developing and applying a formative evaluation framework for health information technology implementations: qualitative investigation. *J Med Internet Res* 2020;22:e15068.
22 Sittig DF, Singh H. A new sociotechnical model for studying health information technology in complex adaptive healthcare systems. *Qual Saf Health Care* 2010;19 Suppl 3:i68–74.
28 Greenhalgh T, Wherton J, Papoutsi C, *et al*. Beyond adoption: a new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. *J Med Internet Res* 2017;19:e367.
29 Shaw J, Shaw S, Wherton J, *et al*. Studying scale-up and spread as social practice: theoretical introduction and empirical case study. *J Med Internet Res* 2017;19:e244.
30 Sittig DF, Ash JS, Singh H. The safer guides: empowering organizations to improve the safety and effectiveness of electronic health records. *Am J Manag Care* 2014;20:418–23.
31 Greenhalgh T, Stramer K, Bratan T, *et al*. Introduction of shared electronic records: multi-site case study using diffusion of innovation theory. *BMJ* 2008;337:.
32 Doebbeling BN, Chou AF, Tierney WM. Priorities and strategies for the implementation of integrated informatics and communications technology to improve evidence-based practice. *J Gen Intern Med* 2006;21 Suppl 2:S50–7.
33 Beuscart-Zéphir MC, Anceaux F, Crinquette V, *et al*. Integrating users' activity modeling in the design and assessment of hospital electronic patient records: the example of anesthesia. *Int J Med Inform* 2001;64:157–71.

34 Keshavjee K, Bosomworth J, Copen J, *et al*. Best practices in EMR implementation: a systematic review. *AMIA Annu Symp Proc* 2006:982.
35 Rose AF, Schnipper JL, Park ER, *et al*. Using qualitative studies to improve the usability of an EMR. *J Biomed Inform* 2005;38:51–60.
36 Granlien MF, Hertzum M, Gudmundsen J. The gap between actual and mandated use of an electronic medication record three years after deployment. *Stud Health Technol Inform* 2008;136:419–24.
37 Bates DW, Ebell M, Gotlieb E, *et al*. A proposal for electronic medical records in U.S. primary care. *J Am Med Inform Assoc* 2003;10:1–10.
38 Aarts J, Doorewaard H, Berg M. Understanding implementation: the case of a computerized physician order entry system in a large Dutch University medical center. *J Am Med Inform Assoc* 2004;11:207–16.
39 Jaspers MWM, Peute LWP, Lauteslager A, *et al*. Pre-Post evaluation of physicians' satisfaction with a redesigned electronic medical record system. *Stud Health Technol Inform* 2008;136:303–8.
40 Singh H, Sittig DF. Measuring and improving patient safety through health information technology: the health it safety framework. *BMJ Qual Saf* 2016;25:226–32.
41 Clemmer TP. Computers in the ICU: where we started and where we are now. J Critical Care 2004;(4):201-207. Pendergast DK, Buchda VL. Charting the course. A quality journey. *Nurs Admin Q* 2003;27:330–5.
42 Hendy J, Reeves BC, Fulop N, *et al*. Challenges to implementing the National programme for information technology (NPfIT): a qualitative study. *BMJ* 2005;331:331–6.
43 Morrison C, Jones M, Blackwell A, *et al*. Electronic patient record use during ward rounds: a qualitative study of interaction between medical staff. *Crit Care* 2008;12:R148.
44 McGowan JJ, Cusack CM, Poon EG. Formative evaluation: a critical component in EHR implementation. *J Am Med Inform Assoc* 2008;15:297–301.
45 Cresswell K, Sheikh A, Franklin BD, *et al*. Theoretical and methodological considerations in evaluating large-scale health information technology change programmes. *BMC Health Serv Res* 2020;20:477.
46 Toynbee P. NHS: the Blair years. *BMJ* 2007;334:1030–1.
47 Abraham C, Nishihara E, Akiyama M. Transforming healthcare with information technology in Japan: a review of policy, people, and progress. *Int J Med Inform* 2011;80:157–70.
48 Weick KE. Educational organizations as loosely coupled systems. *Adm Sci Q* 1976;21:1–9.
49 Estonia's digital transformation. Available: https://library.oapen.org/bitstream/handle/20.500.12657/23594/9780198843719.pdf?sequence=1#page=158 [Accessed 16 May 2021].
50 Blumenthal D, Tavenner M. The "meaningful use" regulation for electronic health records. *N Engl J Med* 2010;363:501–4.
51 Adler-Milstein J, Jha AK. HITECH act drove large gains in hospital electronic health record adoption. *Health Aff* 2017;36:1416–22.
52 Halamka JD, Tripathi M. The HITECH era in retrospect. *N Engl J Med* 2017;377:907–9.
53 Ingebrigtsen T, Georgiou A, Clay-Williams R, *et al*. The impact of clinical leadership on health information technology adoption: systematic review. *Int J Med Inform* 2014;83:393–405.
54 Cresswell K, Sheikh A, Krasuska M, *et al*. Reconceptualising the digital maturity of health systems. *Lancet Digit Health* 2019;1:e200–1.
55 Achieving a digital NHS lessons for national policy from the acute sector. Available: https://www.nuffieldtrust.org.uk/files/2019-05/digital-report-br1902-final.pdf [Accessed 16 May 2021].
56 Krasuska M, Williams R, Sheikh A, *et al*. Technological capabilities to assess digital excellence in hospitals in high performing health care systems: international eDelphi exercise. *J Med Internet Res* 2020;22:e17022.
57 HIMSS. STAGE 6 & 7 Achievement. Available: https://www.himssanalytics.org/europe/stage-6-7-achievement [Accessed 16 Aug 2021].
58 Leon N, Schneider H, Daviaud E. Applying a framework for assessing the health system challenges to scaling up mHealth in South Africa. *BMC Med Inform Decis Mak* 2012;12:1–2.
59 Digital Aspirants. Available: https://www.nhsx.nhs.uk/key-tools-and-info/digital-aspirants/ [Accessed 16 May 2021].
60 Cresswell K, Williams R, Sheikh A. Bridging the growing digital divide between NHS England's hospitals. *J R Soc Med* 2020;1:141076820974998.
61 Nhs long term plan. Available: https://www.longtermplan.nhs.uk/ [Accessed 16 Aug 2021].

BMJ Health & Care Informatics

Development and validation pathways of artificial intelligence tools evaluated in randomised clinical trials

George C M Siontis ,[1] Romy Sweda,[1] Peter A Noseworthy,[2] Paul A Friedman,[2] Konstantinos C Siontis,[2] Chirag J Patel[3]

To cite: Siontis GCM, Sweda R, Noseworthy PA, *et al*. Development and validation pathways of artificial intelligence tools evaluated in randomised clinical trials. *BMJ Health Care Inform* 2021;**28**:e100466. doi:10.1136/bmjhci-2021-100466

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/bmjhci-2021-100466).

Received 04 August 2021
Accepted 04 December 2021

[1]Department of Cardiology, Inselspital, University Hospital of Bern, Bern, Switzerland
[2]Department of Cardiovascular Medicine, Mayo Clinic, Rochester, Minnesota, USA
[3]Department of Biomedical Informatics, Harvard Medical School, Boston, Massachusetts, USA

Correspondence to
Dr George C M Siontis; Georgios.Siontis@insel.ch

ABSTRACT

Objective Given the complexities of testing the translational capability of new artificial intelligence (AI) tools, we aimed to map the pathways of training/validation/testing in development process and external validation of AI tools evaluated in dedicated randomised controlled trials (AI-RCTs).

Methods We searched for peer-reviewed protocols and completed AI-RCTs evaluating the clinical effectiveness of AI tools and identified development and validation studies of AI tools. We collected detailed information, and evaluated patterns of development and external validation of AI tools.

Results We found 23 AI-RCTs evaluating the clinical impact of 18 unique AI tools (2009–2021). Standard-of-care interventions were used in the control arms in all but one AI-RCT. Investigators did not provide access to the software code of the AI tool in any of the studies. Considering the primary outcome, the results were in favour of the AI intervention in 82% of the completed AI-RCTs (14 out of 17). We identified significant variation in the patterns of development, external validation and clinical evaluation approaches among different AI tools. A published development study was found only for 10 of the 18 AI tools. Median time from the publication of a development study to the respective AI-RCT was 1.4 years (IQR 0.2–2.2).

Conclusions We found significant variation in the patterns of development and validation for AI tools before their evaluation in dedicated AI-RCTs. Published peer-reviewed protocols and completed AI-RCTs were also heterogeneous in design and reporting. Upcoming guidelines providing guidance for the development and clinical translation process aim to improve these aspects.

Summary box

What is already known?

- Randomised controlled trials generating the highest grade of evidence are starting to emerge for AI tools in medicine (AI-RCTs).
- Even though distinct steps for the development process of clinical diagnostic and prognostic tools are established, there is no specific guidance for AI-based tools and for the conduct of AI-RCTs.

What does this paper add?

- A limited number of AI-RCTs have been completed and reported.
- AI-RCTs are characterised by heterogenous design and reporting.
- There is significant variation in the patterns of development and validation for AI tools before their evaluation in AI-RCTs.
- Data that would allow independent replication and implementation of the AI tools are usually not provided in the AI-RCTs.

INTRODUCTION

Artificial intelligence (AI) methods are playing an increasingly important role in digital healthcare transformation and precision medicine, particularly because of breakthroughs in diagnostic and prognostic applications developed with deep learning and other complex machine learning approaches. Numerous AI tools have been developed for diverse conditions and settings, demonstrating favourable diagnostic and prognostic performance.[1–3] However, similarly to any other clinical intervention,[4–6] adoption of AI tools in patient care requires careful evaluation of their external validity and their impact on downstream interventions and clinical outcomes, beyond performance metrics during development and external validation. The most robust evaluation of any diagnostic or therapeutic intervention may be performed in the setting of randomised controlled trials (RCTs), which are now slowly emerging in the AI space.

Even though distinct steps of training, validation and testing for the development of AI tools have been described, there are no standardised recommendations for AI-based diagnostic and predictive modelling in biomedicine.[7–10] In addition, overfitting, or the phenomenon of training an AI model that is too closely aligned with a limited training dataset such that it has no generalisation ability, is often of concern in highly parameterised AI models. External validation of AI tools aiming to verify a hyperparameterised model is therefore a critical step in the

evaluation process. Furthermore, the extrapolation of model performance from one setting and patient population to others is not guaranteed.[11 12] Moreover, concerns have been raised about the transparency of reporting in the AI literature to facilitate independent replication of AI tools.[13]

Given the complexities of testing the translational capability of new AI tools and the lack of coherent recommendations, we aimed to map the current pathways of training/validation/testing in development process of AI tools in any medical field and identify external validation patterns of AI tools considered for evaluation in dedicated RCTs (here mentioned as AI-RCTs).

MATERIALS AND METHODS

Data source and study selection process

We identified protocols of ongoing AI-RCTs and reports of completed AI-RCTs that evaluated AI tools compared with control strategies in a randomised fashion for any clinical purpose and medical condition. We searched PubMed for publications in peer-review journals in the last 20 years (last search on 31 December 2020) using the following search terms: "artificial intelligence", "machine learning", "neural network", "deep learning", "cognitive computing", "computer vision" and "natural language processing". We did not search for protocols of AI-RCTs published only in protocol registries since the compliance with reporting and the provided information has been shown to be poor compared with peer-reviewed protocols or published reports of clinical trials.[14–18] We considered only peer-reviewed reports of protocols of AI-RCTs which provided detailed information on the trial design of our interest. We considered clinical trials in which the AI tool (algorithm) was either previously developed or was planned to be developed (trained) as part of the trial before being evaluated in the RCT. Clinical trial protocols were included irrespectively of their status (ongoing or completed). The listed references of eligible studies were also searched for additional potentially eligible studies. The detailed search algorithm is provided in online supplemental box.

Mapping of AI tool development: citation content analysis

For each eligible protocol and report of AI-RCT, we scrutinised the cited articles to identify any previous published study reporting on AI tool development (including training, validation or testing) or claiming external validation in an independent population than the one where the AI tool was initial developed. Each potentially eligible study identified above, was subsequently evaluated in full-text to determine whether it describes the development and/or independent evaluation (external validation) of the AI tool of interest. Finally, we searched Google Scholar for articles citing the index development study of the AI tool or its external validation (if any) in order to trace other studies of external validation (onnline supplemental box).

Data collection

A detailed list of information was gathered from each eligible protocol and report of completed AI-RCT using a standardised form which was built and modified, as required, in an iterative process. We extracted relevant information from the main manuscript and any online supplemental material. From each report, we extracted trial and population characteristics which include: single versus multicentre trial, geographical location of the contributing centres, number of arms of randomisation, level of randomisation (patient or clinicians), total sample size, power calculation approach, type of control intervention, underlying medical condition, period of recruitment, funding source (industry related, non-industry related, both, none, none reported), follow-up duration or duration of the intervention, patient-level data collection through dedicated study personnel or from electronic health records, strategies for dealing with missing data; details on the primary outcome(s) of interest which include: single or composite, continuous or binary, outcome adjudication method(s); considering the primary outcome. Among the unique AI tools, we classified the primary outcomes as therapeutic, diagnostic or feasibility outcomes. We documented whether the results of the completed AI-RCT are in favour to intervention based on the AI tool. We extracted information on whether researchers provide access to the code based on which the AI tool was built. We finally assessed the risk of bias (RoB) in the results of completed AI-RCTs that compared the effect of the AI tool compared with other intervention(s) by using the revised Cochrane risk-of-bias tool for randomised trials RoB 2.[19]

For each study describing the development or external validation of an index AI tool, we extracted the following information: year of publication, recruitment period, geographic area of study population, sample size, clinical field, and whether the authors provided any information that would allow the replication of applied coding. We considered as external validation studies those which fulfilled at least one the following conditions compared with the corresponding development study: different study population, different geographic area, different recruitment period or different group of investigators validating the AI tool.

Statistical analysis

We descriptively analysed the protocols and reports of completed AI-RCTs as a whole and separately. We considered the protocols of already published AI-RCTs as a single report with the index trial. The extracted data were summarised into narrative synthesis and presented in summary tables in the level of AI tools and in the level of AI-RCTs. For illustration purposes, we graphically summarised interconnections of the available development (training/validation/testing) studies, external validation studies and the respective AI-RCTs (either protocols of reports) for each AI tool of interest. We visually evaluated the diversity of the distributions of

peer-reviewed development, external validation studies and the ongoing/published reports of AI-RCTs among the unique AI tools. We also illustrated the time lags and differences in sample sizes between different steps of development (whenever applicable) of an AI tool to subsequent evaluation in dedicated AI-RCTs. Illustrations were conducted in R (V.3.4.1; R-Project for Statistical Computing).

RESULTS

Protocols and completed AI-RCTs

The selection process of eligible protocols and reports of AI-RCTs is summarised in online supplemental figure 1. Overall, we identified 23 unique AI-RCTs[20–45] (6 protocols and 17 reports of completed AI-RCTs) evaluating the clinical effectiveness of 18 unique AI tools for a variety of conditions (tables 1 and 2, online supplemental file 1). Three of the completed AI-RCTs[36 39 45] had previously published protocols.[35 38 44] The identified reports were published over a 10-year period (2009–2020). Half of the AI-RCTs were multicentre (52%) and the majority compared the AI-based intervention to a single control intervention (87%). The median target sample size reported in the protocols of AI-RCTs was 298 (IQR 219–850), whereas for the published AI-RCTs was 214 (IQR 100–437) (table 2, online supplemental table 1). Power calculations were available in 18 out of 23 AI-RCTs. The control arms consisted of standard-of-care interventions in all but one study in which a sham intervention was used as control. In one trial, the investigators also considered a historical control group in addition to the two randomised groups in the trial.[37] Ten AI-RCTs were funded by non-industry sponsors and seven trials did not specify the financial source. The investigators did not specify any strategies for handling missing data in most AI-RCTs (19 out of 23, 83%). Outcome ascertainment was based on electronic health records in the minority of the AI-RCTs (4 out of 23, 17%), while in the remaining studies either was unclear or conventional adjudication methods were applied. A binary or continuous primary outcome was considered in 7 (30%) and 14 (61%) of the trials. Among the 18 unique AI tools (table 1), 10 tools were examined for therapeutic outcomes, 6 for diagnostic and 2 for feasibility. The results according to the primary outcome favoured the AI intervention in 82% of the completed AI-RCTs (14 out of 17), with 1 trial claiming lower in-hospital mortality rates with the AI intervention[25] (table 2, online supplemental table 2). None of the AI-RCTs reported their intention to provide access to the coding of the AI tool. Online supplemental table 3 summarises the detailed risk-of-bias judgement for each domain and the overall judgement for each AI-RCT. Three trials were at low RoB, five trials were judged to raise 'some concerns' and nine to be at 'high RoB', mainly due to the lack of appropriate/complete reporting related to adherence of intended interventions and in measurement of the outcome of interest.

Development, external validation and clinical evaluation pathways of AI tools

We identified considerable dissimilarities in the patterns of development, external validation and clinical evaluation steps among AI tools (figures 1 and 2, online supplemental table 4). A peer-reviewed publication describing the development process was not found for 8 out of the 18 unique AI tools. In 12 AI-RCTs, the study population originated from the same geographic area and population as the one where the AI tool was developed in. We were able to identify at least one external validation study linked to a trial only in 11 out of the 23 ongoing/completed AI-RCTs. All of the external validation studies considered a different recruitment period compared with that in the development study, but from the same geographical area in all 11 cases. The number of external validation studies ranged from 1 to 4 per AI tool (figure 1). Three AI tools were evaluated in two different AI-RCTs, and one AI tool was evaluated in three different AI-RCTs with differences in patient populations and examined outcomes (table 1 and figure 1). Among the AI tools with external validation studies, in 6 cases the external validation studies were published at the same time or clearly after the corresponding AI-RCT (figure 2). In those six cases, the external validation studies applied the AI tool in different populations and/or clinical settings, compared with those where it was developed and those studied in the AI-RCT.

Among the 17 completed AI-RCTs, the distribution of the sample sizes and timelines of publications for development, external validation and AI-RCT reports is shown in figures 2 and 3. The sample sizes of the development studies were larger than the respective external validation studies and AI-RCTs, whereas external validation studies and AI-RCTs did not differ in sample sizes. Median time from publication of a development study to publication of the respective AI-RCT was 1.4 years (IQR 0.2–2.2). The time lag between publication of the development studies to the publication of AI-RCTs varied for different AI tools, but there was considerable overlap of the timelines of external validation and AI-RCT publications (table 1, figure 2, online supplemental tables 1 and 4).

DISCUSSION

Large scale real-world data collected from electronic-health records have allowed the development of diagnostic and prognostic tools based on machine learning approaches.[46–52] Evaluations of the clinical impact of such tools in dedicated RCTs are now starting to emerge in the literature. Our empirical assessment of the literature identified significant variation in the patterns of AI tool development (training, validation, testing) and external (independent) validation leading up to their evaluation in dedicated AI-RCTs. In this early phase of novel AI-RCTs, trials are characterised by heterogeneous design and reporting. Data that would allow independent

Table 1 Descriptive summary of 18 artificial intelligence tools evaluated in AI-RCTs

AI-RCT/AI tool	Medical field	Aim	Description	Softwares/ packages used	Primary outcome	Outcome classification	Main finding
El-Soll *et al*[20]/na	Pulmonary diseases	Prediction of optimal CPAP titration	A general regression neural network with tree-layer structure (input layer, hidden layer and output layer) was trained to predict optimal CPAP pressure based on five input variables.	Neuroshell 2, Ward Systems, Frederick, MD	Time of achieving optimal continuous positive airway pressure titration	Therapeutic	AI guided CPAP titration resulted in lower time to optimal CPAP and lower titration failure rate.
Martin *et al*[21]/Patient Journey Record system (PaJR)	Chronic diseases	Early detection of adverse trajectories and reduction of readmissions	Summaries of semistructured phone calls about well-being and health-concerns analysed by machine learning-based and rule-based algorithms. By detection of signs of health deterioration, an alarm was triggered. Alarms were reviewed by a clinical case manager who decided subsequent interventions.	Not specified	Unplanned emergency ambulatory care sensitive admissions	Therapeutic	AI tool allowed early identification of health concerns and resulted in reduction of emergency ambulatory care sensitive admissions.
Zeevi *et al*[22]; Popp *et al*[27]/na	Nutrition/ endocrinology	Prediction of postprandial glycaemic response	A machine learning algorithm employing stochastic gradient boosting regression was developed to predict personalised postprandial glycaemic responses to real-life meals. Inputs included blood parameters, dietary habits, anthropometrics, physical activity and gut microbiota.	Code adapted from the sklearn 0.15.2 Gradient Boosting Regressor class	Postprandial glycaemic responses	Therapeutic	AI tool accurately predicted postprandial glycaemic responses. Individualised dietary interventions resulted in lower postprandial glycaemic responses and alterations to gut microbiota.
Piette *et al*[23]/na	Behavioural	Improvement of chronic low back pain by personalised cognitive behavioural therapy	A reinforcement learning algorithm is employed to customise cognitive behavioural therapy in patients with chronic low back pain. The algorithm learns from patient feedback and pedometer step counts to provide personalised therapy recommendations.	Not specified	24-item Roland Morris Disability Questionnaire	Therapeutic	Not applicable (protocol of AI-RCT)
Sadasivam *et al*[24]/ PERSPeCT	Behavioural	Smoking cessation	A hybrid recommender system employing content-based and collaborative filtering methods was developed to provide personalised messages supporting smoking cessation. Data sources included message-metadata together with implicit (ie, website view patterns) and explicit (item ratings) user feedbacks. Each participant received AI-selected messages from a message database that matched their readiness to quit status.	Not specified	Smoking cessation	Therapeutic	After 30 days, there was no difference in smoking cessation rates, although those receiving AI-tailored computer messages rated them as being more influential.

Continued

Table 1 Continued

AI-RCT/AI tool	Medical field	Aim	Description	Softwares/ packages used	Primary outcome	Outcome classification	Main finding
Shimabukuro *et al*[25]/ InSight	Infectious diseases	Sepsis prediction	A machine learning based classifier with gradient tree boosting was developed to generate risk scores predictive of sepsis, severe sepsis or septic shock based on electronic health record data. Depending on the predicted risk, an alarm was triggered. Further evaluation and treatment was according to standard guidelines.	Matlab	Average hospital length of stay	Therapeutic	AI-guided monitoring decreased length of hospital stay and in-hospital mortality.
Fulmer*et al*[26]/Tess	Behavioural	Reduction of depression and anxiety	An AI-based chatbot was designed to deliver personalised conversations in the form of integrative mental health support, psychoeducation and reminders. Users could enter both free-text and/or select predefined responses.	Not specified	Self-report tools (PHQ-9, GAD-7, PANAS) for symptoms of depression and anxiety	Therapeutic	AI-based intervention resulted in reduction of symptoms of depression and anxiety.
Wang *et al*[28 32]/ EndoScreener	Gastroenterology	Automatic polyp and adenom detection	A deep CNN based on the SegNet architecture was trained to automatically identify polyps in real time during colonoscopy.	Not specified	Adenoma detection rate	Diagnostic	Automatic polyp detection system resulted in a significant increased detection rate of adenomas and polyps.
Wu *et al*[29]; Chen *et al*[33]/Wisense/ Endoangel	Gastroenterology	Quality improvement of endoscopy by automatic identification of blind spots	A deep CNN combined with deep reinforcement learning was designed to automatically detect blind spots during EGD.	TensorFlow	Blind spot rate	Feasibility	AI reduced blind spot rate during esophagogastroduodenoscopy
Gong *et al*[34]/ Wisense/ Endoangel	Gastroenterology	Quality improvement of endoscopy by automatic identification of adenomas	A deep CNN combined with deep reinforcement learning was designed to automatically detect adenomas during colonoscopy.	TensorFlow	Adenoma detection rate	Diagnostic	AI increased adenoma detection rate during colonoscopy
Oka *et al*[30]/Asken	Nutrition/ endocrinology	Automated nutritional intervention to improve glycaemic control in patients with diabetes mellitus	Participants use a mobile app to select foods from a large database (>100 000) of menus, which are analysed with regards to their energy and nutrition content by an AI-powered photo analysis system. The trial will compare dietary interventions based on AI-supported vs standard nutritional therapy.	Not specified	Change in glycated haemoglobin levels	Therapeutic	Not applicable (protocol of AI-RCT)

Continued

Table 1 Continued

AI-RCT/AI tool	Medical field	Aim	Description	Softwares/ packages used	Primary outcome	Outcome classification	Main finding
Lin *et al*[31]/CC-Cruiser	Ophtalmology	Diagnosis and risk stratification of childhood cataracts	A collaborative cloud platform encompassing automatic analysis of uploaded split-lamp photographs of the ocular anterior segment by an AI engine was established. Output includes diagnosis, risk stratification and treatment recommendations.	Not specified	Diagnostic performance for childhood cataract	Diagnostic	AI tool was less accurate than senior consultants in diagnosing childhood cataracts, but was less time-consuming.
Wijnberge *et al*[35 36]; Schneck *et al*[37]; Maheshwari *et al*[38 39]/EV1000 HPI monitoring device	Surgery/anaesthesia	Prediction of intraoperative hypotension	A machine learning algorithm to predict hypotensive episodes from arterial pressure waveforms was designed. The model output was implemented as an early warning system based on the estimated 'hypotension prediction index'(0–100, with higher numbers reflecting higher likelihood of incipient hypotension) and included information about the underlying cause for the predicted hypotension (vasoplegia, hypovolaemia, low contractility).	Matlab	Time-weighted average of hypotension during surgery/frequency and absolute and relative duration of intraoperative hypotension	Therapeutic	The AI-based early warning system performed different under different clinical settings (ie, elective non-cardiac surgery, primary total hip arthroplasty, moderate to high risk non-cardiac surgery patients).
Auloge *et al*[40]/na	Orthopaedics	Facilitation of percutaneous vertebroplasty by augmented reality/ artificial intelligence-based navigation	A navigation system integrating four video cameras within the flat-panel detector of a standard C-arm fluoroscopy machine was developed, including an AI software that automatically recognised osseous landmarks, identified each vertebral level and displayed 2D/3D planning images on the user interface. After manual selection of the target vertebra, the software suggests an optimal trans-pedicular approach. Once trajectory is validated, the C-arm automatically rotates and the virtual trajectory is superimposed over the real-world camera input with overlaid, motion-compensated needle trajectories.	Not specified	Technical feasibility of trocar placement using augmented reality/artificial intelligence guidance	Feasibility	AI-guided percutaneous vertebroplasty was feasible and resulted in lower radiation exposure compared with standard fluoroscopic guidance.

Continued

Table 1 Continued

AI-RCT/AI tool	Medical field	Aim	Description	Softwares/ packages used	Primary outcome	Outcome classification	Main finding
Wong *et al*[41]/Everion/ Biovitals	Infectious diseases	Early detection of COVID-19 in quarantine subjects	Data from a wearable biosensor worn on the upper arm are automatically transferred in real time through a smartphone app to a cloud storage platform and subsequently analysed by the AI software. The results (including risk prediction of critical events) are displayed on a web-based dashboard for clinical review.	Not specified	Time to diagnosis of coronavirus disease 19	Diagnostic	Not applicable (protocol of AI-RCT)
Aguilera *et al*[42]/na	Behavioural	Increase physical activity in patients with diabetes and depression by tailored messages via AI mobile health application	Participants receive daily messages from a messaging bank, with message category, timing and frequency being selected by a reinforcement learning algorithm. The algorithm employs Thompson Sampling to continuously learn from contextual features like previous physical activity, demographic and clinical characteristics.	Not specified	Improvement in physical activity defined by daily step counts	Therapeutic	Not applicable (protocol of AI-RCT)
Hill *et al*[43]/na	Cardiology	Atrial fibrillation detection	An atrial fibrillation risk prediction algorithm was developed using machine learning techniques on retrospective data from nearly 3 000 000 adult patients without history of atrial fibrillation. The output is provided as a risk score for the likelihood of atrial fibrillation.	R	Prevalence of diagnosed atrial fibrillation	Diagnostic	Not applicable (protocol of AI-RCT)
Yao *et al*[44 45]/na	Cardiology	ECG AI-guided screening for low left ventricular ejection fraction	A CNN model has been trained to predict low LVEF from 10 s 12-lead ECGs strips from nearly 98'000 patients with paired ECG-TTE data. The final model consisted of 6 convolutional layers, each followed by a nonlinear 'Relu' activation function, a batch-normalisation layer and a max-pooling layer. The binary output will be incorporated into the electronic health record and triggered a recommendation for TTE in case of a positive screening result (predicted LVEF ≤35%).	Keras, TensorFlow, Python	Newly discovered left ventricular ejection fraction <50%	Diagnostic	An AI algorithm applied on existing ECGs enabled the early diagnosis of low left ventricular ejection fraction in patients managed in primary care practices.

AI-RCT, artificial intelligence randomised controlled trial; CNN, convolutional neural network; CPAP, continuous positive airway pressure; GAD-7, General Anxiety Disorder-7; LVEF, left ventricular ejection fraction; na, not available; PANAS, Positive and Negative Affect Schedule; PHQ-9, Patient Health Questionnaire-9; TTE, transthoracic echocardiography.

Table 2 Characteristics of peer-reviewed protocols and completed RCTs evaluating artificial intelligence tools

Characteristics	AI-RCTs (n=23)	Protocols of AI-RCTs (n=6)	Completed AI-RCTs (n=17)
No of centres, n (%)			
Single	11 (48)	1 (17)	10 (59)
Multicentre	12 (52)	5 (83)	7 (41)
Geographic area, n (%)			
Asia	8 (35)	2 (33)	6 (35)
Europe	5 (22)	1 (17)	4 (24)
North America	9 (39)	3 (50)	6 (35)
Other	1 (4)	0 (0)	1 (6)
Arms of randomisation, n (%)			
Two	20 (87)	5 (83)	15 (88)
Three	3 (13)	1 (17)	2 (12)
Level of randomisation, n (%)			
Patients	22 (96)	6 (100)	16 (94)
Clinicians	1 (4)	0 (0)	1 (6)
Sample size			
Median (IQR)	214 (108–571)	298 (219–830)	214 (100–437)
Min	20	100	20
Max	22 641	18 000	22 641
Power calculations, n (%)			
Yes	18 (78)	6 (100)	12 (71)
No	5 (22)	0 (0)	5 (29)
Type of control intervention, n (%)			
Standard of care	22 (96)	6 (100)	16 (94)
Sham procedure	1 (4)	0 (0)	1 (6)
Funding source, n (%)			
Industry related	4 (17)	1 (17)	3 (18)
Non-industry related	10 (43)	4 (66)	6 (35)
None reported	7 (30)	1 (17)	6 (35)
None	2 (9)	0 (0)	2 (12)
Data sources, n (%)			
Dedicated personnel	5 (22)	2 (33)	3 (18)
Dedicated personnel and EHR	4 (17)	2 (33)	2 (12)
EHR	4 (17)	2 (33)	2 (12)
Not applicable	4 (17)	0 (0)	4 (23)
Not specified	6 (27)	0 (0)	6 (35)
Strategies for missing data, n (%)			
Specified	4 (17)	4 (67)	0 (0)
Not specified	19 (83)	2 (33)	17 (100)
Primary outcome(s), n (%)			
Binary	7 (30)	0 (0)	7 (41)
Binary and continuous	1 (4)	0 (0)	1 (6)
Categorical	1 (4)	1 (17)	0 (0)
Continuous	14 (61)	5 (83)	9 (53)
Primary outcome favours AI tool, n (%)			

Continued

Table 2 Continued

Characteristics	AI-RCTs (n=23)	Protocols of AI-RCTs (n=6)	Completed AI-RCTs (n=17)
Yes	13 (57)	0 (0)	13 (76)
No	2 (9)	0 (0)	2 (12)
Not applicable	8 (34)	6 (100)	2 (12)
Different geographic area of study population in development study and AI-RCT, n (%)			
Yes	3 (14)	1 (17)	2 (12)
No	12 (52)	1 (17)	11 (65)
Not applicable*	8 (34)	4 (66)	4 (23)
External validation of AI tool, n (%)			
Yes	11 (48)	2 (33)	9 (53)
No	12 (52)	4 (67)	8 (47)
Different geographic area†	0 (0)	0 (0)	0 (0)
Different time period†	11 (48)	2 (33)	9 (53)

*The respective development study was not identified.
†Compared with the development study.
AI-RCTs, artificial intelligence randomised controlled trials; EHR, electronic health records.

replication and implementation of AI tools were not available in any of the AI-RCTs.

There is growing recognition that AI tools need to be held to the same rigorous standard of evidence as other diagnostic and therapeutic tools in medicine with standardised reporting.[53–55] The recently published extensions of the COSNORT and Standard Protocol Items: Recommendations for Interventional Trials (SPIRIT) statements for RCTs of AI-based interventions (namely Consolidated Standards of Reporting Trials (CONSORT)-AI[56] and SPIRIT-AI)[57] are beginning to provide such a framework.

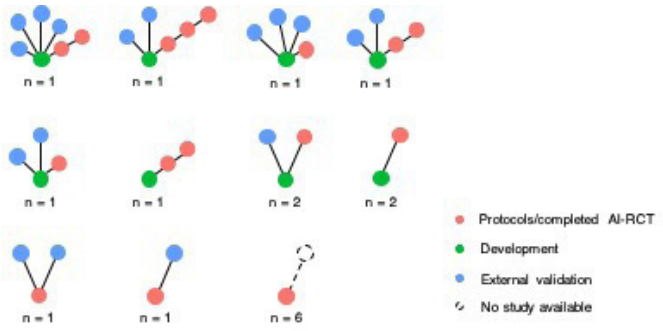


Figure 1 Patterns of pathways of development (training, validation and/or testing), external validation and clinical evaluation of artificial intelligence tools in ongoing and completed clinical trials (n=23). In network level, each circle corresponds to an individual study (green, blue, and red for development, external validation and AI-RCTs, respectively). The number below each network represents the number of unique AI tools having identified with the respective pattern (network) of studies. For example, the first network of the top row corresponds to a unique AI tool for which a development study (green circle), four external validation studies (blue circles), and two AI-RCTs (red circles) were found. AI-RCTs, artificial intelligence randomised controlled trials.

Among the items mandated by these documents, investigators in AI-RCT have to provide better clarity around the intended use of the AI intervention, descriptions how the AI intervention can be integrated into the trial setting, and the setting expectations that investigators make the AI intervention and/or its code assessable. Although most of the studies included in the current review were published before these guidelines, the marked heterogeneity in current reporting underscore the urgency of this call and provide a standard for the ongoing evaluation of these kinds of studies.

RCTs remain the cornerstone of evaluation of diagnostic or therapeutic interventions proposed for clinical use, and this should be no less true for AI interventions. While the experience with the clinical application of AI tools is still early, the evaluation standards of these tools should follow well established norms. AI has demonstrated great promise in transforming many aspects of patient care and healthcare delivery, but the rigorous evaluation standards has lagged for AI tools. Despite numerous published AI applications in medicine,[1–3] in this empirical assessment we have found that a very small fraction has so far undergone evaluation in dedicated clinical trials. We identified significant variation of model development processes leading up to the AI-RCTs. After initial development of an AI tool, at least one external validation study for that particular tool was found for only 11 out of the 23 AI-RCTs. Furthermore, the AI-RCTs were almost always conducted in the same geographic areas as their respective development studies. Thus, the AI-RCTs in this empirical assessment often failed to provide sufficient information regarding the generalisability and external validity of the AI tools. When considering the application of AI tools in the real world, a 'table of

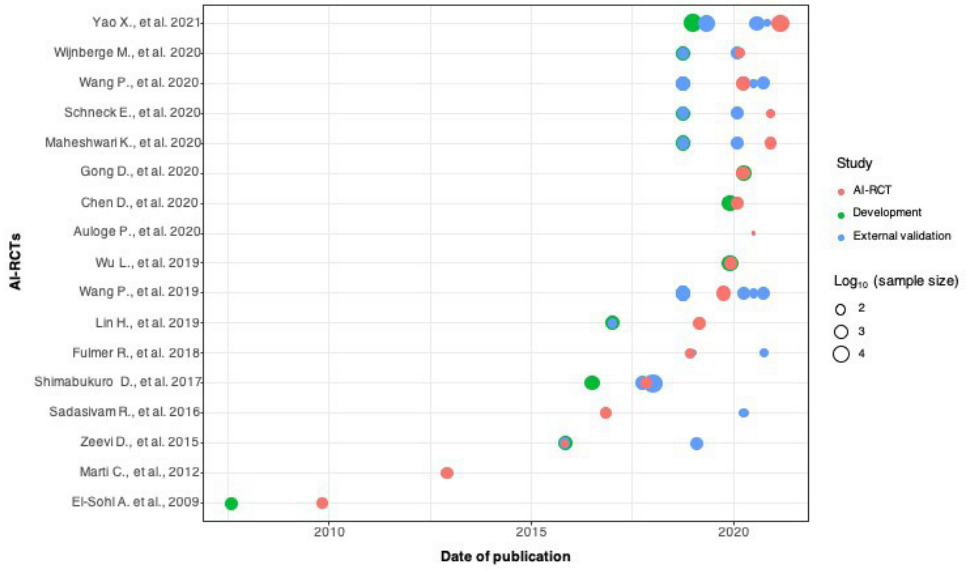


Figure 2 Timelines of publications and sample sizes of development (training, validation and/or testing), external validation studies and completed AI-RCTs (n=17). Each circle corresponds to a unique study (development (training, validation, testing) studies in green, external validation studies in blue, and AI-RCTs in red). Due to the wide range of studies' sample sizes, the values are displaying in logarithmic (log_{10}) scale. AI-RCTs, artificial intelligence randomised controlled trials.

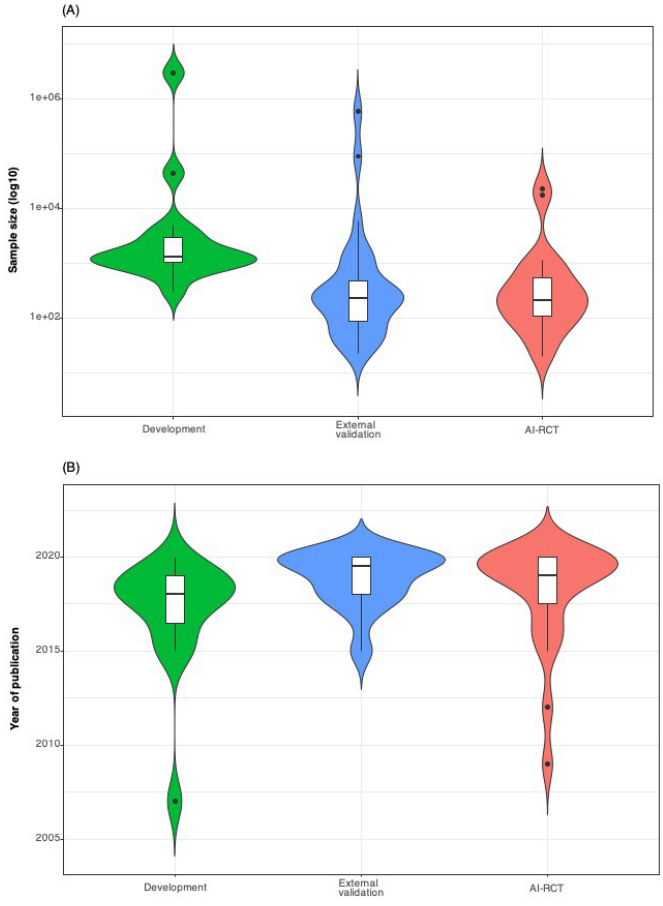


Figure 3 Violin plots showing in comparison the distributions of sample sizes (A) and years of publication (B) of development (training, validation and/or testing), external validation studies and completed AI-RCTs (n=17). AI-RCT, artificial intelligence randomised controlled trials.

ingredients' accompanying the AI tool could be of value. Such a label would include information on how the tool was developed and whether it has been externally validated, including the specific populations, demographic profiles, racial mix, inpatient versus outpatient settings, and other key details. This would allow a potential user to determine whether the AI tool is applicable to their patient or population of interest and whether any deviations in diagnostic or prognostic performance are to be expected.

Along these lines, as with any type of RCT, the choice of primary outcomes in AI-RCTs is also important to consider. Improvement in therapeutic efficacy outcomes with direct patient relevance may be the ultimate criterion of value of an AI tool, but these may also be the most difficult to demonstrate improvements for. The number of studies in each of the three outcome classes in our study (therapeutic, diagnostic, feasibility) was too small to reach conclusions about differences in the probability of statistically significant results between classes. It should also be noted that for diagnostic AI tools, diagnostic performance outcomes that align with the scope of the intervention would be appropriate. However, interpretation of such findings should account for likely dilution of any effect when translating differences in diagnostic outcomes to downstream clinical outcomes.[58] Ultimately, investigation of patient-centric outcomes, should remain a priority whenever possible.

The optimal process for the clinical evaluation of AI tools, ranging from model development to AI-RCTs to real-world implementation, is not yet well defined. Dedicated guidelines on the development, reporting and bridging the development-to-implementation gap of AI tools for prognosis or diagnosis, namely Transparent Reporting of

a multivariable prediction model for Individual Prognosis or Diagnosis-AI (TRIPOD-AI),[59] Prediction model Risk Of Bias ASsessment Tool-AI (PROBAST-AI),[59] Developmental and Exploratory Clinical Investigation of Decision-AI (DECIDE-AI),[60] Standards for Reporting of Diagnostic Accuracy Studies-AI (STARD-AI),[61] Quality Assessment of Diagnostic Accuracy Studies-AI (QUADAS-AI),[62] will be available soon. The heterogeneity in development, validation and reporting in the existing AI literature that we found in this study might be largely attributable to the lack of consensus on research practices and reporting standards in this space. The translational process from development to clinical evaluation of AI tools is in the early phase of a broader scrutiny of AI in various medical disciplines. The upcoming guideline documents are likely to enhance the reliability, replicability, validity and generalisability of this literature.

Furthermore, it is unknown whether all AI tools necessitate testing in traditional, large-scale AI-RCTs.[63] Well-powered, large RCTs that are likely to provide conclusive results are costly, resource intensive and take a long time to complete. Therefore, a clinical evaluation model that routinely requires RCTs may not represent a realistic expectation for the majority of AI tools. However, the ongoing digital transformation in healthcare allows researchers to simplify time-consuming and costly steps of traditional RCTs and to improve efficiency. For example, patient recruitment, follow-up and outcome ascertainment may be performed via nationwide linkage to centralised electronic health records. Natural language processing tools may allow automated screening for patient eligibility and collection of information of patient characteristics and outcomes. Existing web-based, patient-facing portals that are the norm for most healthcare institutions may allow a fully virtual consent process for recruitment. for outcomes' ascertainment. The extensions of the COSNORT and SPIRIT statements for RCTs of AI-based interventions (namely CONSORT-AI[56] and SPIRIT-AI)[57] underscore these concepts for facilitating a novel model of AI-RCT.

Limitations

Our empirical evaluation has limitations. First, a number of potentially eligible ongoing trials have not been included, since we summarised peer-reviewed protocols and final reports of AI-RCTs published in PubMed, whereas trials registered in online registries were not considered. However, as has been previously shown,[14–18 64] registered protocols often suffer from incomplete reporting, lack of compliance with the conditions for registration and out-of-date information, which would not have allowed us to appropriately characterise the AI tools and their respective development pathways. Second, as part of this evaluation we did not consider a control group of trials (ie, trials evaluating the clinical impact of traditional diagnostic or prognostic tools). However, such trials could not be directly comparable to the AI-RCTs due to fundamental differences in studied interventions and populations. Third, we were not able to comparatively assess the discriminatory performance of the AI tools across the distinct steps of training/validation/testing and external validation, since such performance metrics were neither systematically nor uniformly reported.

Conclusion

In conclusion, we have found that evaluation of AI tools in dedicated RCTs is still infrequent. There is significant variation in patterns of development and validation for AI tools before their evaluation in RCTs. Published peer-reviewed protocols and completed AI-RCTs also varied in design and reporting. Most AI-RCTs do not test the AI tools in geographical areas outside of those where the tools were developed, therefore generalisability remains largely unaddressed. As AI applications are increasingly reported throughout medicine, there is a clear need for structured evaluation of their impact on patients with a focus on effectiveness and safety outcomes, but also costs and patient-centred care, before their large-scale deployment.[65] The upcoming guidelines for AI tools aim to guide researchers and fill the translational gaps in the conduct and reporting of development and translation steps. All steps in the translation pathway of these tools should serve the development of meaningful and impactful AI tools without compromise under the pressure of innovation.

Acknowledgements None.

Contributors All authors reviewed the final manuscript for submission and contributed to the study as follows: GCMS contributed to the study design, data extraction and management, interpretation and writing of manuscript. GCMS is the guarantor and responsible for the overall content. GCMS accepts full responsibility for the work and the conduct of the study, had access to the data, and controlled the decision to publish. RS contributed to data extraction and management. PAN contributed to interpretation and writing of manuscript. PAF contributed to interpretation and writing of manuscript. KCS contributed to study design, interpretation and writing of manuscript. CJP contributed to interpretation and writing of manuscript.

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

Competing interests None declared.

Patient and public involvement statement Not applicable.

Patient consent for publication Not applicable.

Ethics approval Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement All data relevant to the study are included in the article or uploaded as online supplemental information. Not applicable.

ORCID iD

George C M Siontis http://orcid.org/0000-0003-2128-9205

REFERENCES

1 Gulshan V, Peng L, Coram M, *et al*. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus Photographs. *JAMA* 2016;316:2402–10.
2 Esteva A, Kuprel B, Novoa RA, *et al*. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017;542:115–8.
3 Ehteshami Bejnordi B, Veta M, Johannes van Diest P, *et al*. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA* 2017;318:2199–210.
4 Royston P, Moons KGM, Altman DG, *et al*. Prognosis and prognostic research: developing a prognostic model. *BMJ* 2009;338:b604–7.
5 Siontis GCM, Tzoulaki I, Castaldi PJ, *et al*. External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. *J Clin Epidemiol* 2015;68:25–34.
6 Moons KGM, Altman DG, Vergouwe Y, *et al*. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. *BMJ* 2009;338:b606–90.
7 Collins GS, Reitsma JB, Altman DG, *et al*. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ* 2015;350:g7594.
8 Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. *Lancet* 2019;393:1577–9.
9 Liu X, Faes L, Calvert MJ, *et al*. Extension of the CONSORT and spirit statements. *Lancet* 2019;394:1225.
10 Park SH, Han K. Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. *Radiology* 2018;286:800–9.
11 Kohli M, Prevedello LM, Filice RW, *et al*. Implementing machine learning in radiology practice and research. *AJR Am J Roentgenol* 2017;208:754–60.
12 Aerts HJWL, . The potential of Radiomic-Based phenotyping in precision medicine: a review. *JAMA Oncol* 2016;2:1636–42.
13 Vollmer S, Mateen BA, Bohner G, *et al*. Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. *BMJ* 2020;368:l6927.
14 DeVito NJ, Bacon S, Goldacre B. Compliance with legal requirement to report clinical trial results on ClinicalTrials.gov: a cohort study. *Lancet* 2020;395:361–9.
15 Zarin DA, Tse T, Williams RJ, *et al*. Update on trial registration 11 years after the ICMJE policy was established. *N Engl J Med* 2017;376:383–91.
16 Anderson ML, Chiswell K, Peterson ED, *et al*. Compliance with results reporting at ClinicalTrials.gov. *N Engl J Med* 2015;372:1031–9.
17 Adam GP, Springs S, Trikalinos T, *et al*. Does information from ClinicalTrials.gov increase transparency and reduce bias? results from a five-report case series. *Syst Rev* 2018;7:59.
18 Zarin DA, Tse T, Williams RJ, *et al*. The ClinicalTrials.gov results database--update and key issues. *N Engl J Med* 2011;364:852–60.
19 Sterne JAC, Savović J, Page MJ, *et al*. Rob 2: a revised tool for assessing risk of bias in randomised trials. *BMJ* 2019;366:l4898.
20 El Solh A, Akinnusi M, Patel A, *et al*. Predicting optimal CPAP by neural network reduces titration failure: a randomized study. *Sleep Breath* 2009;13:325–30.
21 Martin CM, Vogel C, Grady D, *et al*. Implementation of complex adaptive chronic care: the patient journey record system (PaJR). *J Eval Clin Pract* 2012;18:1226–34.
22 Zeevi D, Korem T, Zmora N, *et al*. Personalized nutrition by prediction of glycemic responses. *Cell* 2015;163:1079–94.
23 Piette JD, Krein SL, Striplin D, *et al*. Patient-Centered pain care using artificial intelligence and mobile health tools: protocol for a randomized study funded by the US department of Veterans Affairs health services research and development program. *JMIR Res Protoc* 2016;5:e53.
24 Sadasivam RS, Borglund EM, Adams R, *et al*. Impact of a collective intelligence tailored messaging system on smoking cessation: the perspect randomized experiment. *J Med Internet Res* 2016;18:e285.
25 Shimabukuro DW, Barton CW, Feldman MD, *et al*. Effect of a machine learning-based severe sepsis prediction algorithm on patient survival and hospital length of stay: a randomised clinical trial. *BMJ Open Respir Res* 2017;4:e000234.
26 Fulmer R, Joerin A, Gentile B, *et al*. Using psychological artificial intelligence (Tess) to relieve symptoms of depression and anxiety: randomized controlled trial. *JMIR Ment Health* 2018;5:e64.
27 Popp CJ, St-Jules DE, Hu L, *et al*. The rationale and design of the personal diet study, a randomized clinical trial evaluating a personalized approach to weight loss in individuals with pre-diabetes and early-stage type 2 diabetes. *Contemp Clin Trials* 2019;79:80–8.
28 Wang P, Berzin TM, Glissen Brown JR, *et al*. Real-Time automatic detection system increases colonoscopic polyp and adenoma detection rates: a prospective randomised controlled study. *Gut* 2019;68:1813–9.
29 Wu L, Zhang J, Zhou W, *et al*. Randomised controlled trial of WISENSE, a real-time quality improving system for monitoring blind spots during esophagogastroduodenoscopy. *Gut* 2019;68:2161–9.
30 Oka R, Nomura A, Yasugi A, *et al*. Study Protocol for the Effects of Artificial Intelligence (AI)-Supported Automated Nutritional Intervention on Glycemic Control in Patients with Type 2 Diabetes Mellitus. *Diabetes Ther* 2019;10:1151–61.
31 Lin H, Li R, Liu Z, *et al*. Diagnostic efficacy and therapeutic decision-making capacity of an artificial intelligence platform for childhood cataracts in eye clinics: a multicentre randomized controlled trial. *EClinicalMedicine* 2019;9:52–9.
32 Wang P, Liu X, Berzin TM, *et al*. Effect of a deep-learning computer-aided detection system on adenoma detection during colonoscopy (CADe-DB trial): a double-blind randomised study. *Lancet Gastroenterol Hepatol* 2020;5:343–51.
33 Chen D, Wu L, Li Y, *et al*. Comparing blind spots of unsedated ultrafine, sedated, and unsedated conventional gastroscopy with and without artificial intelligence: a prospective, single-blind, 3-parallel-group, randomized, single-center trial. *Gastrointest Endosc* 2020;91:332–9.
34 Gong D, Wu L, Zhang J, *et al*. Detection of colorectal adenomas with a real-time computer-aided system (ENDOANGEL): a randomised controlled study. *Lancet Gastroenterol Hepatol* 2020;5:352–61.
35 Wijnberge M, Schenk J, Terwindt LE, *et al*. The use of a machine-learning algorithm that predicts hypotension during surgery in combination with personalized treatment guidance: study protocol for a randomized clinical trial. *Trials* 2019;20:582.
36 Wijnberge M, Geerts BF, Hol L, *et al*. Effect of a machine Learning-Derived early warning system for intraoperative hypotension vs standard care on depth and duration of intraoperative hypotension during elective noncardiac surgery: the hype randomized clinical trial. *JAMA* 2020;323:1052–60.
37 Schneck E, Schulte D, Habig L, *et al*. Hypotension prediction index based protocolized haemodynamic management reduces the incidence and duration of intraoperative hypotension in primary total hip arthroplasty: a single centre feasibility randomised blinded prospective interventional trial. *J Clin Monit Comput* 2020;34:1149–58.
38 Maheshwari K, Shimada T, Fang J, *et al*. Hypotension prediction index software for management of hypotension during moderate- to high-risk noncardiac surgery: protocol for a randomized trial. *Trials* 2019;20:255.
39 Maheshwari K, Shimada T, Yang D, *et al*. Hypotension prediction index for prevention of hypotension during moderate- to high-risk noncardiac surgery. *Anesthesiology* 2020;133:1214–22.
40 Auloge P, Cazzato RL, Ramamurthy N, *et al*. Augmented reality and artificial intelligence-based navigation during percutaneous vertebroplasty: a pilot randomised clinical trial. *Eur Spine J* 2020;29:1580–9.
41 Wong CK, Ho DTY, Tam AR, *et al*. Artificial intelligence mobile health platform for early detection of COVID-19 in quarantine subjects using a wearable biosensor: protocol for a randomised controlled trial. *BMJ Open* 2020;10:e038555.
42 Aguilera A, Figueroa CA, Hernandez-Ramos R, *et al*. mHealth APP using machine learning to increase physical activity in diabetes and depression: clinical trial protocol for the DIAMANTE study. *BMJ Open* 2020;10:e034723.
43 Hill NR, Arden C, Beresford-Hulme L, *et al*. Identification of undiagnosed atrial fibrillation patients using a machine learning risk prediction algorithm and diagnostic testing (PULsE-AI): study protocol for a randomised controlled trial. *Contemp Clin Trials* 2020;99:106191.
44 Yao X, McCoy RG, Friedman PA, *et al*. Ecg AI-Guided screening for low ejection fraction (Eagle): rationale and design of a pragmatic cluster randomized trial. *Am Heart J* 2020;219:31–6.
45 Yao X, Rushlow DR, Inselman JW. Artificial intelligence-enabled electrocardiograms for identification of patients with low ejection fraction: a pragmatic, cluster-randomized clinical trial. *Nat Med* 2021.

46 Nevin L, PLOS Medicine Editors. Advancing the beneficial use of machine learning in health care and medicine: toward a community understanding. *PLoS Med* 2018;15:e1002708.
47 Ravi D, Wong C, Deligianni F, *et al*. Deep learning for health informatics. *IEEE J Biomed Health Inform* 2017;21:4–21.
48 Hae H, Kang S-J, Kim W-J, *et al*. Machine learning assessment of myocardial ischemia using angiography: development and retrospective validation. *PLoS Med* 2018;15:e1002693.
49 Rodriguez-Ruiz A, Lång K, Gubern-Merida A, *et al*. Stand-Alone artificial intelligence for breast cancer detection in mammography: comparison with 101 radiologists. *J Natl Cancer Inst* 2019;111:916–22.
50 McKinney SM, Sieniek M, Godbole V, *et al*. International evaluation of an AI system for breast cancer screening. *Nature* 2020;577:89–94.
51 Siontis KC, Noseworthy PA, Attia ZI, *et al*. Artificial intelligence-enhanced electrocardiography in cardiovascular disease management. *Nat Rev Cardiol* 2021;18:465–78.
52 Siontis KC, Yao X, Pirruccello JP, *et al*. How will machine learning inform the clinical care of atrial fibrillation? *Circ Res* 2020;127:155–69.
53 Setting guidelines to report the use of AI in clinical trials. *Nat Med* 2020;26:1311.
54 Wu E, Wu K, Daneshjou R, *et al*. How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals. *Nat Med* 2021;27:582–4.
55 Hernandez-Boussard T, Bozkurt S, Ioannidis JPA, *et al*. MINIMAR (minimum information for medical AI reporting): developing reporting standards for artificial intelligence in health care. *J Am Med Inform Assoc* 2020;27:2011–5.
56 Liu X, Rivera SC, Moher D, *et al*. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *BMJ* 2020;370:m3164.
57 Rivera SC, Liu X, Chan A-W, *et al*. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *BMJ* 2020;370:m3210.
58 Siontis KC, Siontis GCM, Contopoulos-Ioannidis DG, *et al*. Diagnostic tests often fail to lead to changes in patient outcomes. *J Clin Epidemiol* 2014;67:612–21.
59 Collins GS, Dhiman P, Andaur Navarro CL, *et al*. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. *BMJ Open* 2021;11:e048008.
60 DECIDE-AI Steering Group. DECIDE-AI: new reporting guidelines to bridge the development-to-implementation gap in clinical artificial intelligence. *Nat Med* 2021;27:186–7.
61 Sounderajah V, Ashrafian H, Golub RM, *et al*. Developing a reporting guideline for artificial intelligence-centred diagnostic test accuracy studies: the STARD-AI protocol. *BMJ Open* 2021;11:e047709.
62 Sounderajah V, Ashrafian H, Rose S, *et al*. A quality assessment tool for artificial intelligence-centered diagnostic test accuracy studies: QUADAS-AI. *Nat Med* 2021;27:1663–5.
63 Collins R, Bowman L, Landray M, *et al*. The magic of randomization versus the myth of real-world evidence. *N Engl J Med* 2020;382:674–8.
64 Califf RM, Zarin DA, Kramer JM, *et al*. Characteristics of clinical trials registered in ClinicalTrials.gov, 2007-2010. *JAMA* 2012;307:1838–47.
65 Perakslis E, Ginsburg GS. Digital Health-The need to assess benefits, risks, and value. *JAMA* 2021;325:127–8.

BMJ Health & Care Informatics

Association of persistent acute kidney injury and renal recovery with mortality in hospitalised patients

Tezcan Ozrazgat-Baslanti,[1,2] Tyler J Loftus,[2,3] Yuanfang Ren,[1,2] Esra Adiyeke,[1,2] Shunshun Miao,[1,2] Haleh Hashemighouchani,[1,2] Rubab Islam,[1] Rajesh Mohandas,[1] Saraswathi Gopal,[1] Elizabeth A Shenkman,[4] Panos Pardalos,[5] Babette Brumback,[6] Mark S Segal,[1] Azra Bihorac [1,2]

To cite: Ozrazgat-Baslanti T, Loftus TJ, Ren Y, *et al*. Association of persistent acute kidney injury and renal recovery with mortality in hospitalised patients. *BMJ Health Care Inform* 2021;**28**:e100458. doi:10.1136/bmjhci-2021-100458

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/bmjhci-2021-100458).

Received 05 August 2021
Accepted 08 November 2021

For numbered affiliations see end of article.

Correspondence to
Dr Azra Bihorac;
abihorac@ufl.edu

ABSTRACT

Objectives Acute kidney injury (AKI) affects up to one-quarter of hospitalised patients and 60% of patients in the intensive care unit (ICU). We aim to understand the baseline characteristics of patients who will develop distinct AKI trajectories, determine the impact of persistent AKI and renal non-recovery on clinical outcomes, resource use, and assess the relative importance of AKI severity, duration and recovery on survival.

Methods In this retrospective, longitudinal cohort study, 156 699 patients admitted to a quaternary care hospital between January 2012 and August 2019 were staged and classified (no AKI, rapidly reversed AKI, persistent AKI with and without renal recovery). Clinical outcomes, resource use and short-term and long-term survival adjusting for AKI severity were compared among AKI trajectories in all cohort and subcohorts with and without ICU admission.

Results Fifty-eight per cent (31 500/54 212) had AKI that rapidly reversed within 48 hours; among patients with persistent AKI, two-thirds (14 122/22 712) did not have renal recovery by discharge. One-year mortality was significantly higher among patients with persistent AKI (35%, 7856/22 712) than patients with rapidly reversed AKI (15%, 4714/31 500) and no AKI (7%, 22 117/301 466). Persistent AKI without renal recovery was associated with approximately fivefold increased hazard rates compared with no AKI in all cohort and ICU and non-ICU subcohorts, independent of AKI severity.

Discussion Among hospitalised, ICU and non-ICU patients, persistent AKI and the absence of renal recovery are associated with reduced long-term survival, independent of AKI severity.

Conclusions It is essential to identify patients at risk of developing persistent AKI and no renal recovery to guide treatment-related decisions.

Summary

What is already known?

- In surgical sepsis, acute kidney injury (AKI) trajectory subgroups have unique physiologic signatures, suggesting utility for targeted, therapeutic interventions; it is unknown whether similar subgroups exist among all hospitalised patients.
- Early recovery after AKI is associated with favourable long-term outcomes; it is unclear whether this association is affected by critical illness and AKI severity.

What does this paper add?

- To our knowledge, this study is the first large scale, granular description of associations among patient baseline characteristics, illness severity, AKI trajectory and severity and other clinical outcomes.
- Among large and diverse cohort of hospitalised patients and in subset of critically ill patients, persistent AKI and the absence of renal recovery were associated with fourfold to fivefold increased risk to die within a period of 3 years compared with patients who did not develop AKI, independent of AKI severity.
- Our study is strengthened by the use of validated computable phenotype for kidney health encompassing both chronic kidney disease and AKI while maintaining consistency with Kidney Disease: Improving Global Outcomes and Acute Disease Quality Initiative guidelines and addressing the potential racial biases introduced by race adjustments in glomerular filtration rate and creatinine using comprehensive reference creatinine calculations.
- The identification of AKI trajectory subgroups facilitates prognostication and identifies patients who may benefit from nephrology consultation and preventive measures.

INTRODUCTION

Acute kidney injury (AKI) affects nearly one-quarter of hospitalised patients worldwide and up to 60% of patients in the intensive care unit (ICU).[1–3] The delayed or incomplete recovery of renal function confers increased risk for chronic critical illness with poor long-term survival and quality of life.[4] Prevention, early diagnosis, and appropriate treatment with euvolaemia, avoidance of nephrotoxic substances, and relief of obstructive uropathy have variable efficacy in improving patient outcomes. To optimise these management strategies and their early

delivery, it is necessary to understand the trajectories of AKI and recovery among hospitalised patients.

AKI trajectories can be classified as rapidly reversed, persistent with renal recovery or persistent without renal recovery. These trajectory subgroups are important for risk-stratification in surgical sepsis patients, for whom AKI trajectory subgroups have unique physiological signatures of immunological and endothelial dysfunction, suggesting potential utility for targeted, therapeutic interventions.[5–7] Yet, it remains unknown whether these clinical trajectories apply to broader, heterogeneous cohorts of hospitalised patients and associated long-term outcomes remain unclear.[8–11]

We performed a retrospective, longitudinal study of 355 678 adult hospitalisations, 78 769 of which included ICU admission. Our objectives were to understand the baseline characteristics of patients who will develop distinct AKI trajectories, determine the impact of persistent AKI and renal non-recovery on clinical outcomes, resource use and assess the relative importance of AKI severity, duration and recovery on survival.

METHODS

Study design

Using the University of Florida Health (UFH) Integrated Data Repository as Honest Broker, we created a single-centre, longitudinal dataset extracted directly from the electronic health records of 156 699 patients ≥18 years admitted to UFH between 1 January 2012 and 22 August 2019. After exclusion of encounters with no serum creatinine measurement to determine AKI status during hospitalisation and within 48 hours of hospital admission, our final cohort included 355 678 hospital encounters from 138 140 patients (online supplemental figure 1, supplemental methods).

Assessment of kidney function

We developed and validated computable phenotype algorithms for comprehensive kidney health assessments during hospital admission to determine AKI status and classification.[12] Kidney Disease: Improving Global Outcomes (KDIGO) serum creatinine criteria[13] and consensus report of the Acute Disease Quality Initiative (ADQI) 16 Workgroup on renal recovery were used as conceptual frameworks for our computable phenotype algorithms.[5 7 14 15] Stage 1 AKI was termed 'mild AKI'; stages 2 and 3 were termed 'severe AKI'. Duration and evidence of renal recovery[5] were used to define rapidly reversed and persistent AKI with and without renal recovery at discharge. We defined an episode of AKI as beginning with AKI onset and ending if there are two consecutive days without AKI identified, thus allowing us to identify a new episode of AKI in a patient who has recovered from a previous episode of AKI. Persistent AKI was defined as an AKI episode lasting beyond 48 hours. Rapid reversal of AKI was defined as complete reversal of AKI by KDIGO criteria within 48 hours of AKI onset, and remaining as such. Frequency of creatinine testing within the first 2 days of AKI onset is reported in online supplemental table 1. Renal recovery was adjudicated for each episode of AKI based on normalisation criteria at the time of hospital discharge. We grouped each encounter based on the worst trajectory group during hospitalisation as persistent AKI without renal recovery, persistent AKI with renal recovery, rapidly reversed AKI or no AKI. Reference creatinine was determined using preadmission measurements (n=302 349, 85%)[7 16] or the estimated creatinine using the Modification of Diet in Renal Disease (MDRD) Study equation assuming that baseline estimated glomerular filtration rate (eGFR) is 75 mL/min/per 1.73 m^2 (n=52 544, 15%) (online supplemental methods).[13 17 18] Reference creatinine was used to estimate preadmission reference GFR using the Chronic Kidney Disease Epidemiology Collaboration (CKD-EPI) equation.[19] The race multiplier was removed to avoid the undesirable effects of racial corrections in MDRD and CKD-EPI formulae.[20–22] Online supplemental table 2 shows results of sensitivity analyses that shows reclassification in AKI trajectory group when race correction was included. For each patient, we calculated daily kinetic GFR using the estimate of creatinine production rate and per cent change in creatinine.[23]

Outcomes

Primary clinical outcomes were hospital, 1-year and 3-year mortality. Primary renal outcomes were new renal replacement therapy (RRT) and new CKD within 90 days or 1 year of hospital discharge as well as CKD progression within 1 year of hospital discharge. Other exploratory outcomes included hospital and 30-day outcomes (online supplemental methods).

Statistical analysis

Overall survival of each trajectory group was evaluated in 138 140 patients using log-rank and Kaplan-Meier methods. Propensity score-based inverse weighting was used to plot adjusted Kaplan-Meier curves in which the probability of being in a trajectory group was calculated using multinomial logistic model that included patient demographics (age, gender, African-American race) and Charlson Comorbidity Index score. Cox proportional-hazards regression used to assess associations between groups of interest (AKI, AKI severity, AKI trajectories and combination of AKI trajectory and severity) and time to death while controlling for demographics, Charlson Comorbidity Index score and provision of mechanical ventilation and ICU admission for ≥2 days, with the exception of exclusion of variable for prolonged ICU admission and mechanical ventilation for subgroup analysis of non-ICU cohort. Multivariate logistic regression was used to model hospital mortality with similar baseline characteristics variables. Models were also run with and without AKI severity to examine change in association after further adjustment of AKI severity included as indicators of severe AKI or stage 3 AKI. Survival models

were started at hospital discharge and followed up to 3 years. Model discrimination was assessed using Harrell's concordance index. Kinetic GFR values were visualised using line plots illustrating average values with 95% CIs over time. All p values were adjusted for multiple comparisons using Bonferroni methods.[24] Statistical analyses were performed with R V.3.5.3 and Python V.3.8 software (online supplemental methods).

RESULTS

Clinical characteristics of patients

Average age was 54 years with female and male sex approximately equally distributed (table 1, online supplemental table 3). The most common comorbidity was hypertension (63%) and the most common admission diagnosis was disease of the circulatory system (18%). Eighty-three per cent of all patients had urgent admission and 15% were transferred from another hospital. While 20% of all admissions included inpatient surgery, about 22% (78 769/355 678) of hospitalisations required ICU admission. Average age was higher for the ICU cohort (59 vs 54 years) with a higher proportion of male sex (54% vs 42%) and lower proportion of African-American race (18% vs 26%) (online supplemental tables 4 and 5).

Clinical trajectories of patients with AKI during hospitalisation

Overall, 54 212 patients (15%) developed AKI; 37 973 (11%) had AKI within 48 hours of admission (table 1, figure 1A). While 58% (31 500/54 212) had AKI that rapidly reversed within 48 hours, the remaining 42% (22 712/54 212) had persistent AKI. By the time of discharge or death, 62% (14 122/22 712) of all subjects with persistent AKI did not recover renal function.

We examined clinical trajectories of AKI in encounters stratified by requirement of ICU admission. Prevalence of AKI was higher in ICU cohort (35%, 27 711/78 769) than the non-ICU cohort (10%, 26 501/276 909) (figure 2A,D). In the non-ICU cohort, 69% (18 222/26 501) had rapidly reversed AKI; the remaining 31% (8279/26 501) had persistent AKI with 67% (5549/8279) of them not recovering renal function at discharge or death. Meanwhile, among ICU cohort, 48% (13 278/27 711) had rapidly reversed AKI; the remaining 52% (14 433/27 711) had persistent AKI with 59% (8573/14 433) of them not recovering renal function at discharge or death.

Regardless of trajectory and ICU admission, AKI patients had a greater burden of comorbid disease and had lower reference eGFR, especially for patients with persistent AKI (table 1, online supplemental tables 3–5). Forty per cent of all AKI patients had CKD with moderate/severe stage (55%). A greater proportion of AKI patients were transferred from another hospital (25% vs 13%). Sepsis, acute renal failure, congestive heart failure and respiratory disease were the most common admission diagnosis for persistent AKI patients (online supplemental figure 2). Patients without AKI had greater incidence of abdominal and chest pain as the admission diagnosis. Within 48 hours of admission, patients with persistent AKI had significantly higher blood urea nitrogen (mean range 35–36 mg/dL, SD range 25–26 mg/dL), serum creatinine (median range 1.5–1.6 mg/dL, IQR range 0.9–2.4 mg/dL), serum creatinine-reference creatinine ratio (mean range 1.9–2.1, SD range 1.4–1.8) and cystatin C (median 1.4 mg/L, IQR range 0.9–2.1 mg/L) compared with others (table 1, online supplemental table 6). Similar trends have been observed in ICU and non-ICU cohorts (online supplemental tables 7 and 8). We have observed that nephrotoxic exposure within first 2 and 3 days of hospital admission and between hospital admission and first AKI onset was significantly higher in persistent AKI patients compared with patients rapidly reversed AKI (table 1, online supplemental tables 6–8).

Compared with patients with rapidly reversed AKI, patients with persistent AKI were more likely to present with more severe stage (stage 3) (18%, 4143/22 712 vs 4%, 1221/31 500) and had greater incidence of RRT within 48 hours of admission (3%, 731/22 712 vs 0.03%, 10/31 500), with similar trends for the entire hospitalisation (table 2). Persistent AKI patients received significantly more blood products than others (14%, n=3259) and exhibited significantly greater fluid retention with an average fluid overload of approximately 1.2% of admission volume within 48 hours of admission. Volumes of intravenous saline infusions were higher in patients that developed AKI.

Early and sustained decline in kinetic GFR below 60 mL/min/1.73 m^2 was demonstrated among persistent AKI patients (online supplemental figure 3). Among patients with persistent AKI, those who failed to recover renal function at discharge had sustained kinetic GFR approximately 60 mL/min/1.73 m^2; those with renal recovery exhibited gradually increasing kinetic GFR.

Correlation of trajectories with biomarker profile

There were significant differences in biomarker distributions across trajectory groups within 24 hours of admission (online supplemental tables 6–8). Regardless of AKI trajectories and ICU admission requirement, AKI patients had lower systolic, diastolic and mean blood pressure, higher average glucose, lower average platelet counts and average albumin compared with patients without AKI. These differences were greatest within the ICU cohort.

Persistent AKI patients sustained longer duration of mean arterial blood pressure below 60 mm Hg (median range 111–120 min, IQR range 40–300 min) and received more vasopressors compared with patients without persistent AKI. Almost half of all patients with persistent AKI were admitted to the ICU and had greater incidence of mechanical ventilation (19%,

Table 1 Baseline and early admission characteristics by trajectory groups in all cohort

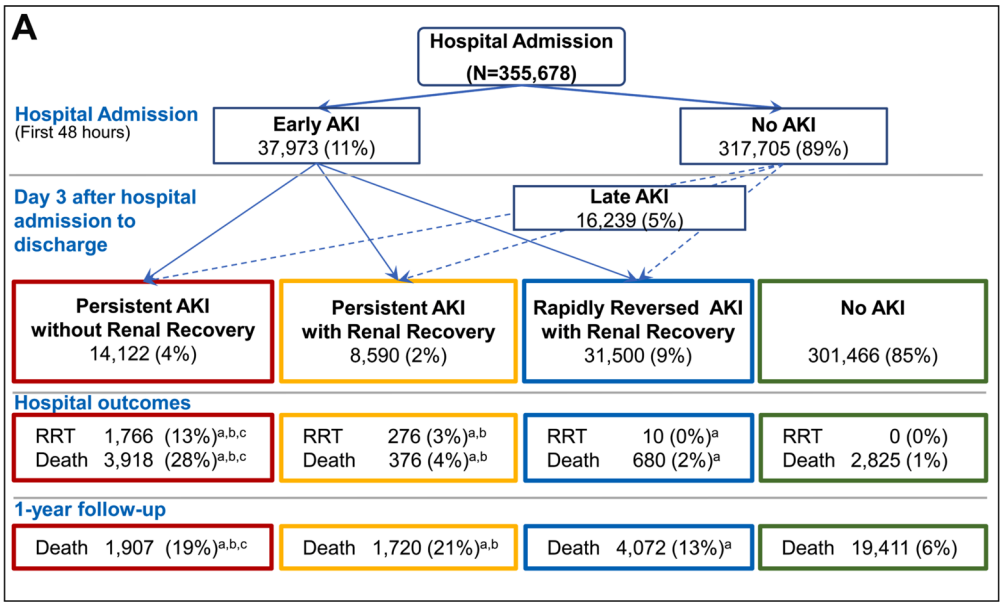
Variables	All subjects (N=355 678)	AKI (N=54 212, 15%)	Persistent AKI without renal recovery (N=14 122, 4%)	Persistent AKI with renal recovery (N=8590, 2%)	Rapidly reversed AKI (N=31 500, 9%)	No AKI (N=301 466, 85%)
Preadmission clinical characteristics						
Age, years, mean (SD)	54 (19)	60 (17)*	61 (17)*†	60 (17)*†	59 (17)*	53 (19)
Female sex, n (%)	196 023 (55)	27 146 (50)*	7137 (51)*‡	4116 (48)*†	15 893 (50)*	168 877 (56)
African American ethnicity, n (%)	85 825 (24)	12 411 (23)*	3057 (22)*†	1928 (22)*	7426 (24)*	73 414 (24)
Distance from hospital (mile), median (IQR)	14 (3–30)	21 (3–37)*	23 (9–41)*†‡	24 (10–42)†	18 (3–36)*	14 (3–27)
Comorbidities, n (%)						
Hypertension	225 192 (63)	36 556 (67)*	9100 (64)*†‡	5929 (69)*	21 527 (68)*	188 636 (63)
Chronic pulmonary disease	149 551 (42)	23 044 (43)*	5517 (39)*†‡	3771 (44)*	13 756 (44)*	126 507 (42)
Cardiovascular disease	129 930 (37)	23 451 (43)*	5619 (40)*†‡	3960 (46)*†	13 872 (44)*	106 479 (35)
Diabetes mellitus	104 546 (29)	18 331 (34)*	4359 (31)*†‡	2928 (34)*	11 044 (35)*	86 215 (29)
Chronic kidney disease	85 942 (24)	21 421 (40)*	5442 (39)*‡	3864 (45)*†	12 115 (38)*	64 521 (21)
Moderate/severe (≥G-stage 3)	34 956 (41)	11 739 (55)*	3025 (56)*	2211 (57)*†	6503 (54)*	26 236 (41)
Preadmission estimated glomerular filtration rate (mL/min per 1.73 m^2), median (IQR)	63.6 (46.4–85.1)	55.9 (37.9–78.7)*	53.9 (32.6–81.1)*†	54.1 (36.6–76.2)*†	57.2 (40.1–78.6)*	65.6 (49.5–86.6)
Admission characteristics, n (%)						
Emergent admission	295 286 (83)	45 927 (85)*	12 091 (86)*†	7355 (86)*†	26 481 (84)*	249 359 (83)
Transfer from another hospital	53 265 (15)	13 706 (25)*	4626 (33)*†	2721 (32)*†	6359 (20)*	39 559 (13)
Surgery on admission day	51 310 (14)	8327 (15)*	2179 (15)*‡	1455 (17)*†	4693 (15)*	42 983 (14)
Surgery at any time	72 541 (20)	15 216 (28)*	4068 (29)*†‡	3272 (38)*†	7876 (25)*	57 325 (19)
Kidney function within 48 hours of the admission						
AKI, n (%)	37 973 (11)	37 973 (70)*	9706 (69)*†	5908 (69)*†	22 359 (71)*	0 (0)
Stage 1	25 963 (68)	25 963 (68)*	4538 (47)*†‡	3132 (53)*†	18 293 (82)*	0 (0)
Stage 2	6646 (18)	6646 (18)*	2328 (24)*†	1473 (25)*†	2845 (13)*	0 (0)
Stage 3	5364 (14)	5364 (14)*	2840 (29 *†‡	1303 (22)*†	1221 (5)*	0 (0)
Stage three without RRT	4623 (12)	4623 (12)*	2217 (23)*†‡	1195 (20)*†	1211 (5)*	0 (0)
Stage three with RRT	741 (2)	741 (2)*	623 (6)*†‡	108 (2)*†	10 (0)*	0 (0)
Highest blood urea nitrogen (mg/dL), mean (SD)	18 (13)	30 (22)*	35 (26)*†‡	36 (25)*†	27 (17)*	15 (9)

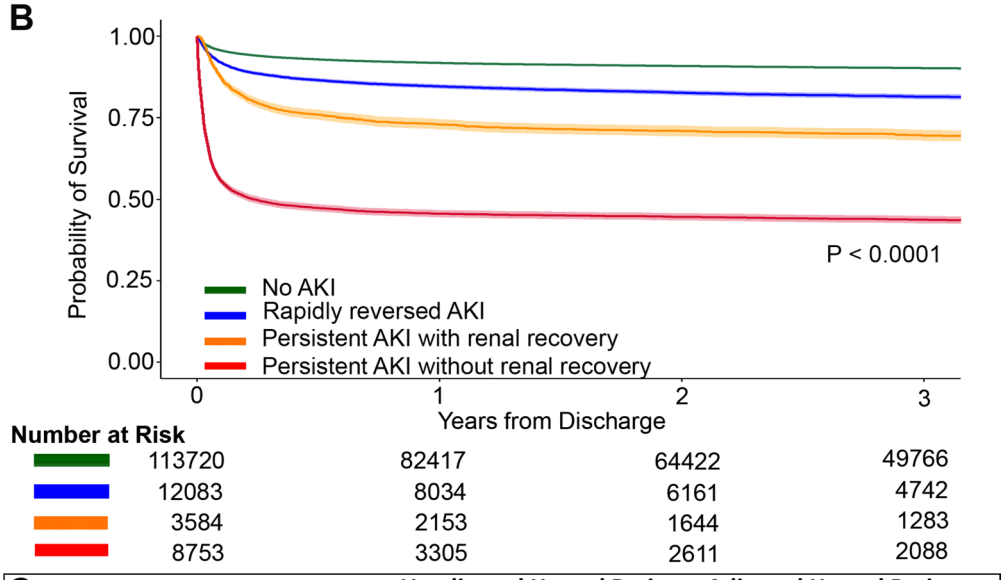
Continued

Table 1 Continued

Variables	All subjects (N=355 678)	AKI (N=54 212, 15%)	Persistent AKI without renal recovery (N=14 122, 4%)	Persistent AKI with renal recovery (N=8590, 2%)	Rapidly reversed AKI (N=31 500, 9%)	No AKI (N=301 466, 85%)
Highest serum creatinine (mg/dL), median (IQR)	0.9 (0.7–1.1)	1.4 (1.0–1.9)*	1.5 (0.9–2.4)*†‡	1.6 (1.0–2.4)*†	1.3 (1.0–1.7)*	0.8 (0.7–1.0)
Reference creatinine (mg/dL), median (IQR)	0.8 (0.7–1.0)	0.8 (0.7–1.1)*	0.8 (0.7–1.1)*†‡	0.9 (0.7–1.2)*†	0.8 (0.7–1.1)*	0.8 (0.7–0.9)
Highest/reference creatinine, mean (SD)	1.2 (0.6)	1.8 (1.2*	2.1 (1.8)*†	1.9 (1.4)*†	1.6 (0.7)*	1.1 (0.2)
Count of nephrotoxic drug, mean (SD)						
Within 2 days after hospital admission	0.86 (0.97)	1.26 (1.03)*	1.41 (1.03)*†	1.39 (1.03)*†	1.17 (1.02)*	0.79 (0.94)
Within 3 days after hospital admission	0.92 (1.00)	1.39 (1.07)*	1.54 (1.07)*†	1.55 (1.07)*†	1.27 (1.06)*	0.83 (0.97)
Between hospital admission and first AKI onset	1.33 (1.18)	1.33 (1.18)	1.49 (1.20)†	1.47 (1.20)†	1.23 (1.15)	NA

*P<0.05 compared with no AKI.
†P<0.05 compared with rapidly reversed AKI.
‡P<0.05 compared with persistent AKI with renal recovery.
§Cardiovascular disease was considered if there was a history of congestive heart failure, coronary artery disease of peripheral vascular disease.
AKI, acute kidney injury; NA, not applicable; RRT, renal replacement therapy.



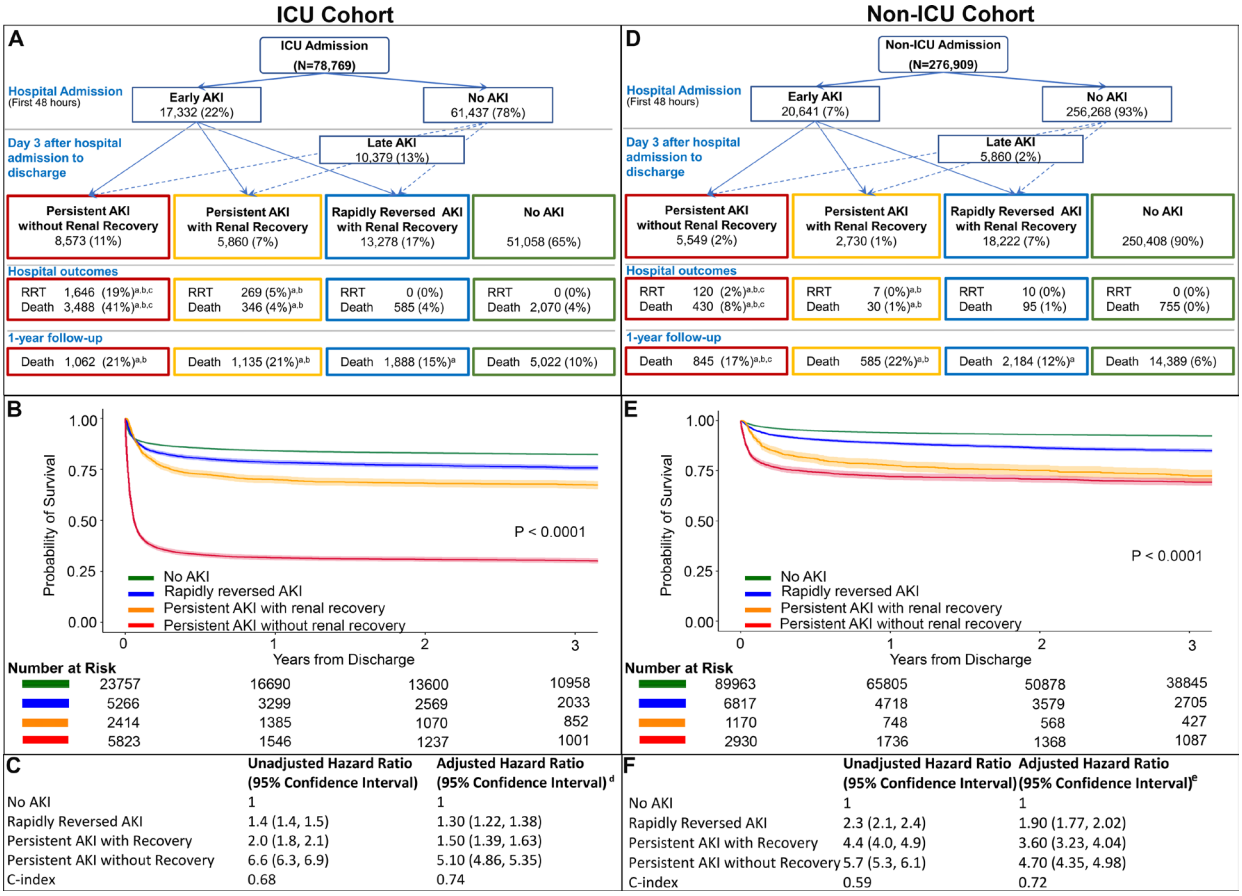


C

	Unadjusted Hazard Ratio (95% Confidence Interval)	Adjusted Hazard Ratio (95% Confidence Interval)[d]
No AKI	1	1
Rapidly Reversed AKI	2.2 (2.1, 2.3)	1.60 (1.50, 1.67)
Persistent AKI with Recovery	3.8 (3.5, 4.0)	1.90 (1.80, 2.03)
Persistent AKI without Recovery	9.7 (9.4, 10.0)	5.63 (5.40, 5.86)
C-index	0.67	0.77

Figure 1 Hospital and long-term outcomes by trajectories of acute kidney injury (AKI) in hospitalised adult patients. (A) Trajectories of AKI in hospitalised adult patients. 1-year follow-up outcome was reported among hospital survivors. (B) Adjusted Kaplan-Meier survival curves and number at risk by AKI trajectories. Propensity score based inverse weighting was used to plot adjusted Kaplan-Meier curves where propensity of being in a trajectory group was calculated using multinomial logistic model that included patient demographics (age, gender, ethnicity), and Charlson Comorbidity Index score. (C) Hazard ratios for all-cause mortality by AKI trajectories. [a]Significantly different from no AKI group (Bonferroni-adjusted p<0.05). [b]Significantly different from rapidly reversed AKI group (Bonferroni-adjusted p<0.05). [c]Significantly different from persistent AKI with renal recovery (Bonferroni-adjusted p<0.05). [d]Adjusted for age, gender, ethnicity, Charlson Comorbidity Index score, and need for mechanical ventilation for more than 2 days and need for intensive care unit admission for more than 2 days. RRT, renal replacement therapy.

4363/22 712) within 24 hours of hospital admission. Other biomarkers that were significantly different in patients with persistent AKI included lower average arterial oxygen tension/fractional inspired oxygen ratio (mean range 295–314, SD range 199–205), higher average lactate (mean range 2.9–3.6 mmol/L,



C

	Unadjusted Hazard Ratio (95% Confidence Interval)	Adjusted Hazard Ratio (95% Confidence Interval)[d]
No AKI	1	1
Rapidly Reversed AKI	1.4 (1.4, 1.5)	1.30 (1.22, 1.38)
Persistent AKI with Recovery	2.0 (1.8, 2.1)	1.50 (1.39, 1.63)
Persistent AKI without Recovery	6.6 (6.3, 6.9)	5.10 (4.86, 5.35)
C-index	0.68	0.74

F

	Unadjusted Hazard Ratio (95% Confidence Interval)	Adjusted Hazard Ratio (95% Confidence Interval)[e]
No AKI	1	1
Rapidly Reversed AKI	2.3 (2.1, 2.4)	1.90 (1.77, 2.02)
Persistent AKI with Recovery	4.4 (4.0, 4.9)	3.60 (3.23, 4.04)
Persistent AKI without Recovery	5.7 (5.3, 6.1)	4.70 (4.35, 4.98)
C-index	0.59	0.72

Figure 2 Hospital and long-term outcomes by trajectories of acute kidney injury (AKI) in hospitalised adult patients stratified by ICU admission. (A) Trajectories of AKI in hospitalised adult patients who have been admitted to ICU during hospitalisation. 1-year follow-up outcome was reported among hospital survivors. (B) Adjusted Kaplan-Meier survival curves and number at risk by AKI trajectories injury in hospitalised adult patients who have been admitted to ICU during hospitalisation. Propensity score based inverse weighting was used to plot adjusted Kaplan-Meier curves where propensity of being in a trajectory group was calculated using multinomial logistic model that included patient demographics (age, gender, ethnicity) and Charlson Comorbidity Index score. (C) HRs for all-cause mortality by AKI trajectories in hospitalised adult patients who have been admitted to ICU during hospitalisation. (D) Trajectories of AKI in hospitalised adult patients who have not been admitted to ICU during hospitalisation. 1-year follow-up outcome was reported among hospital survivors. (E) Adjusted Kaplan-Meier survival curves and number at risk by AKI trajectories in hospitalised adult patients who have not been admitted to ICU at any time during hospitalisation. Propensity score based inverse weighting was used to plot adjusted Kaplan-Meier curves where propensity of being in a trajectory group was calculated using multinomial logistic model that included patient demographics (age, gender, ethnicity) and Charlson Comorbidity Index score. (F) Hazard ratios for all-cause mortality by AKI trajectories in hospitalised adult patients who have not been admitted to ICU at any time during hospitalisation. [a]Significantly different from no AKI group (Bonferroni-adjusted p<0.05). [b]Significantly different from rapidly reversed AKI group (Bonferroni-adjusted p<0.05). [c]Significantly different from persistent AKI with renal recovery (Bonferroni-adjusted p<0.05). [d]Adjusted for age, gender, ethnicity, Charlson Comorbidity Index score, and need for mechanical ventilation for more than 2 days and need for ICU admission for more than 2 days. [e]Adjusted for age, gender, ethnicity, and Charlson Comorbidity Index score. ICU, intensive care unit; RRT, renal replacement therapy.

SD range 2.7–4.0 mmol/L) and lower average haematocrit (31%, SD 7%).

Persistence of kidney dysfunction and absence of recovery affect short-term and long-term outcomes

Median duration of AKI was 5 (IQR 3–8) days among patients with persistent AKI (table 2, online supplemental table 9). Persistent AKI patients required significantly more hospital resources, with longer mechanical ventilation (5 days), ICU admission (7 days) and hospital admission (10 days) compared with patients without AKI. Patients who required ICU admission had worse AKI stage, more AKI days and higher percentage of recurrent AKI (online supplemental tables 9–11), especially in the subset of persistent AKI patients.

Patients with persistent AKI without recovery of renal function had significantly higher in-hospital mortality (28%), followed by the next highest mortality rate in patients with persistent AKI with renal recovery (4%). Even after for adjustment for AKI severity and baseline

Table 2 Renal characteristics, resource utilisation and hospital outcomes during entire hospitalisation by trajectories of AKI in all cohort

Variables	All subjects (N=355 678)	AKI (N=54 212, 15%)	Persistent AKI without renal recovery (N=14 122, 4%)	Persistent AKI with renal recovery (N=8590, 2%)	Rapidly reversed AKI (N=31 500, 9%)	No AKI (N=301 466, 85%)
Renal characteristics during entire hospitalisation						
Worst AKI staging, n (%)						
Stage 1	36 258 (10)	36 258 (67)*	5210 (37)*†‡	4176 (49)*†	26 872 (85)*	0 (0)
Stage 2	9551 (3)	9551 (18)*	3762 (27)*†‡	2492 (29)*†	3297 (10)*	0 (0)
Stage 3	8403 (2)	8403 (16)*	5150 (36)*†‡	1922 (22)*†	1331 (4)*	0 (0)
Stage three without RRT	6351 (2)	6351 (12)*	3384 (24)*†‡	1646 (19)*†	1321 (4)*	0 (0)
Stage three with RRT	2052 (1)	2052 (4)*	1766 (13)*†‡	276 (3)†	10 (0)*	0 (0)
AKI duration, days, median (IQR)	2 (1–4)	2 (1–4)	5 (3–9) b‡	4 (3–7)	1 (1–2)	NA
Recurrent AKI, n (%)	6466 (2)	6466 (12)*	2173 (15)*†‡	1957 (23)*†	2336 (7)*	0 (0)
No renal recovery at discharge/death, n (%)	22 240 (6)	22 240 (41)*	14 122 (100)*†‡	0 (0)†	8118 (26)*	0 (0)
Resource utilisation during entire hospitalisation						
Hospital days, median (IQR)	3 (1–6)	7 (4–14)*	8 (4–15)*†‡	14 (8–24)*†	6 (3–10)*	2 (1–5)
Admission to ICU, n (%)	78 769 (22)	27 711 (51)*	8573 (61)*†‡	5860 (68)*†	13 278 (42)*	51 058 (17)
Days in ICU, median (IQR)	4 (2–7)	6 (3–12)*	6 (3–13)*†‡	9 (5–18)*†	5 (3–9)*	3 (–5)
Mechanical ventilation, n (%)	23 286 (7)	11 876 (22)*	4779 (34)*†	2876 (33)*†	4221 (13)*	11 410 (4)
Mechanical ventilation calendar days, median (IQR)	3 (2–6)	4 (2–9)*	4 (2–9)*†‡	5 (2–12)*†	3 (2–7)*	2 (1–4)
Vasopressor or inotropes used, n (%)	55 415 (16)	17 261 (32)*	6016 (43)*†	3781 (44)*†	7464 (24)*	38 154 (13)
Hospital disposition, n (%)						
Hospital mortality	7799 (2)	4974 (9)*	3918 (28)*†‡	376 (4)*†	680 (2)*	2825 (1)
Another hospital, LTAC, SNF, Hospice	34 092 (10)	10 028 (18)*	3011 (21)*†‡	2494 (29)*†	4523 (14)*	24 064 (8)
Home/rehab	313 787 (88)	39 210 (72)*	7193 (51)*†‡	5720 (67)*†	26 297 (83)*	274 577 (91)
30-day outcomes (among survivors), n (%)	347 879	49 238 (14)	10 204 (3)	8214 (2)	30 820 (9)	298 641 (86)
Death in 30 days of discharge	4934 (1)	1776 (4)*	570 (6)*†	418 (5)*†	788 (3)*	3158 (1)
Trajectory group for encounter with readmission within 30 days of discharge	83 592 (24)	12 748 (26)*	2528 (25)‡	2381 (29)*†	7839 (25)*	70 844 (24)
Persistent AKI with no renal recovery	2764 (3)	1297 (10)*	536 (21)*†‡	223 (9)*†	538 (7)*	1467 (2)

Continued

Table 2 Continued

Variables	All subjects (N=355 678)	AKI (N=54 212, 15%)	Persistent AKI without renal recovery (N=14 122, 4%)	Persistent AKI with renal recovery (N=8590, 2%)	Rapidly reversed AKI (N=31 500, 9%)	No AKI (N=301 466, 85%)
Persistent AKI with renal recovery	2118 (3)	933 (7)*	231 (9)*†	239 (10)*†	463 (6)*	1185 (2)
Rapidly reversed AKI	7505 (9)	2504 (20)*	502 (20)*	448 (19)*	1554 (20)*	5001 (7)
No AKI	59 164 (71)	7096 (56)*	1100 (44)*†‡	1337 (56)*†	4659 (59)*	52 068 (73)
Unknown	12 041 (14)	918 (7)	159 (6)*†	134 (6)*†	625 (8)*	11 123 (16)
Other complications during entire hospitalisation						
Venous thromboembolism, n (%)	15 755 (4)	5180 (10)*	1589 (11)*†‡	1290 (15)*†	2301 (7)*	10 575 (4)
Sepsis, n (%)	29 836 (8)	13 995 (26)*	5102 (36)*†‡	3275 (38)*†	5618 (18)*	15 841 (5)
Cardiovascular complication, n (%)	31 780 (9)	15 229 (28)*	5553 (39)*†	3469 (40)*†	6207 (20)*	16 551 (5)
Thirty-day mortality, n (%)	11 082 (3)	5655 (10)*	3962 (28)*†‡	506 (6)*†	1187 (4)*	5427 (2)
One-year mortality, n (%)	34 687 (10)	12 570 (23)*	5802 (41)*†‡	2054 (24)*†	4714 (15)*	22 117 (7)
Three-year mortality, n (%)	49 144 (14)	15 703 (29)*	6414 (45)*†‡	2669 (31)*†	6620 (21)*	33 441 (11)

*P<0.05 compared with rapidly no AKI.
†P<0.05 compared with rapidly reversed AKI.
‡P<0.05 compared with persistent AKI with renal recovery.
AKI, acute kidney injury; ICU, intensive care unit; LTAC, long-term acute care hospital; NA, not applicable; RRT, renal replacement therapy; SNF, skilled nursing facility.

characteristics (age, gender, Charlson Comorbidity Index score), odds of hospital mortality was significantly higher in all, ICU and non-ICU cohorts with OR range of 1.1–4.8 for persistent AKI without renal recovery and range of 12.0–40.6 for persistent AKI with renal recovery compared with no AKI group (online supplemental table 12). One-year mortality for patients with persistent AKI (35%, n=7856) was significantly higher compared with patients with rapidly reversed AKI (15%, n=4714) and those without AKI (7%, n=22 117).

One-year survival following persistent AKI without renal recovery was 46%, significantly lower than patients with persistent AKI with renal recovery (73%), rapidly reversed AKI (85%) and no AKI (92%) (figure 1B). Persistent AKI without renal recovery was associated with increased all-cause mortality with unadjusted and adjusted hazard ratios (HR) of 9.7 (95% CI 9.4 to 10.0) and 5.63 (95% CI 5.40 to 5.86), compared with no AKI group (figure 1C). One-year survival was substantially lower for persistent AKI without renal recovery group who required ICU admission. Adjusted hazard rate of all-cause mortality was approximately five times greater for persistent AKI without renal recovery group both in non-ICU and ICU cohorts (figure 2B, E, C and F). When further adjusted for AKI severity, HRs remained similar (online supplemental table 13).

A combination of AKI stage, duration and renal recovery at discharge classifications were used to perform analysis for seven subphenotypes. One-year survival after AKI was significantly lower relative to no AKI (figure 3A) while severe AKI was associated with lower survival compared with no AKI and mild AKI (figure 3B). One-year survival following severe-persistent AKI without renal recovery was 39%, significantly lower than for mild-persistent AKI without renal recovery (60%), severe-persistent AKI with renal recovery (73%), mild-persistent AKI with renal recovery (73%), severe-rapidly reversed AKI (84%), mild-rapidly reversed AKI (85%) and no AKI (92%) (figure 3D). While survival rates did not differ significantly between mild and severe AKI for rapidly reversed AKI and persistent AKI with renal recovery trajectories, they were significantly lower for the persistent AKI without renal recovery trajectory and mild-persistent AKI was associated with markedly worse outcomes than severe-rapidly reversed AKI (figure 3C,D). Similar trends were observed for ICU and non-ICU cohort (online supplemental figures 4 and 5). Sensitivity analysis excluding encounters whose reference creatinine was calculated using MDRD creatinine yielded similar HRs (online supplemental figure 6).

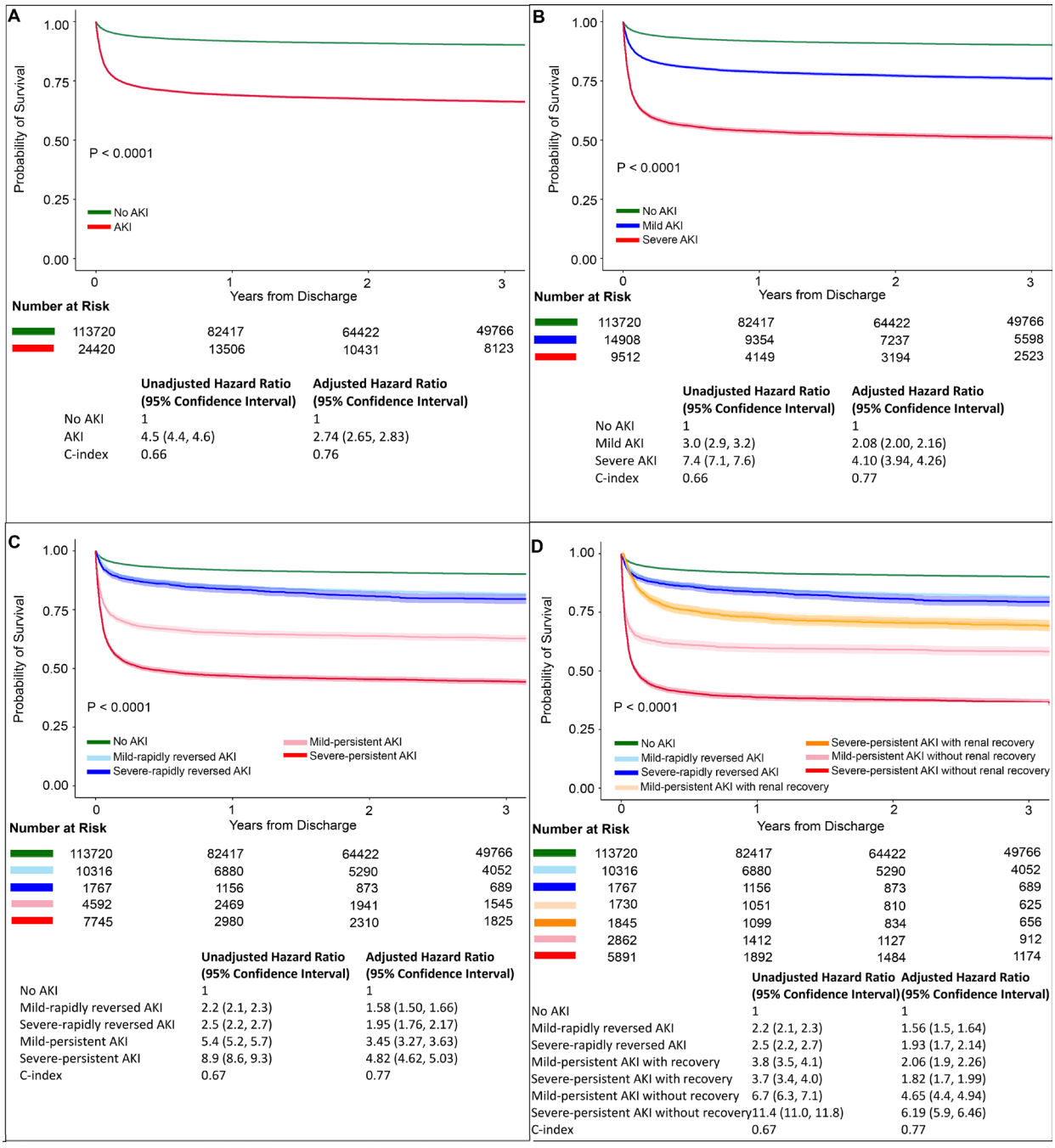


Figure 3 Adjusted Kaplan-Meier survival curves and number at risk by AKI subphenotypes obtained stratifying by (A) no AKI vs any AKI (B) AKI stratified by severity (C) AKI stratified by severity and duration (D) AKI stratified by severity and trajectories of AKI using duration and recovery of AKI. Propensity score based inverse weighting was used to plot adjusted Kaplan-Meier curves where propensity of being in a trajectory group was calculated using multinomial logistic model that included age, gender, ethnicity and Charlson Comorbidity Index score. Adjusted hazard ratios were obtained adjusting for the same variables as well as need for mechanical ventilation for more than 2 days and need for intensive care unit admission for more than 2 days. AKI, acute kidney injury.

Regardless trajectories, among 347 811 patients who survived 1 year after hospital discharge, AKI patients with CKD on admission had greater incidence of CKD progression (18%) within 1 year of admission compared with those without AKI (11%) (online supplemental tables 14–16). Incidence of new CKD and new RRT within 1-year follow-up were significantly higher among patients with AKI (16% and 3%, respectively) compared with patients without AKI (4% and 0.5%, respectively).

DISCUSSION

In a retrospective, longitudinal large cohort of hospitalised patients and in subset of patients who required and did not require ICU admission, we characterised distinct

AKI clinical trajectories and associated survival and resource use. A high proportion (15%) of all hospitalised patients developed AKI, and almost half (42%) developed persistent AKI. Compared with patients who did not develop critical illness during hospitalisation, the ICU cohort had higher proportions of AKI (35% vs 10%) and persistent AKI (52% vs 31%), consistent with evidence of burden of sepsis and organ dysfunction among ICU patients. Among patients with persistent AKI, most (62%) did not recover renal function prior to hospital discharge.

Hypotension, hyperglycaemic, thrombocytopaenia and hypoalbuminaemia were more frequent among patients with AKI; greater severity of these conditions was associated with worse AKI trajectory. There were significant, stepwise increases in 1-year mortality for patients with no AKI (7%), rapidly reversed AKI (15%), persistent AKI with renal recovery (24%) and persistent AKI without renal recovery (41%). Worse AKI trajectory was also associated with greater resource use, manifest as greater incidence of ICU admission, mechanical ventilation and vasopressor administration. Finally, the severity of AKI was combined with recovery trajectory to generate a more granular subphenotyping scheme that augmented discrimination for outcomes after persistent AKI with vs without renal recovery.

Our rationale for performing this work was that clinical trajectories of hospitalised patients with AKI had not been sufficiently described, although similar work has been performed using different patient populations. In a similar prospective, observational study performed at our centre, critically ill patients with surgical sepsis and AKI had a greater overall incidence of AKI compared with this study of hospitalised adults. Yet, associations between AKI trajectories and outcomes were similar between that study and this study, for which only 22% of all admissions involved surgery.[7] A previous retrospective analysis of critically ill patients classified AKI recovery phenotypes by AKI reversal within 7 days of onset and renal recovery at discharge, reporting that sepsis patients had greater incidence of AKI relapse without recovery and increased 1-year mortality, consistent with this study.[25] Similarly, a retrospective study of 5443 patients with septic shock found that subjects with rapid reversal of AKI within 24 hours of onset had lower in-hospital mortality.[26] A recent retrospective study of 350 patients admitted in ICU presented the value of accounting for time-dependent competing risk of discharge or death when assessing recovery pattern in determining AKI recovery trajectories.[27 28] A prospective study of ICU patients from two ICU populations (n=1914; 1867) identified higher mortality rates among non-resolving AKI.[6] Another prospective cohort study of 1538 hospitalised patients demonstrated graded associations among incident or progressive CKD, long-term dialysis, and all-cause death with worse outcomes after non-resolving AKI, intermediate outcomes after resolving AKI, and best outcomes among participants without AKI.[27] Collectively, previous studies have used similar methods to evaluate different patient populations, producing results that are consistent with ours. To our knowledge, this study is the first large-scale, granular description of associations among patient baseline characteristics, illness severity, AKI trajectory and severity, and other clinical outcomes. Using a large, diverse cohort of hospitalised patients as well as in ICU and non-ICU subcohorts, we have shown that significant decreases in long-term survival with persistent AKI and the absence of renal recovery, independent of AKI severity, suggesting importance of identification of AKI trajectories.

The clinical trajectories of AKI and recovery among hospitalised patients described herein could be applied to optimise prevention, early diagnosis and appropriate treatment of AKI. One strength of our study is the use of validated computable phenotype for kidney health encompassing both CKD and AKI while maintaining consistency with KDIGO and ADQI guidelines and addressing the potential racial biases introduced by race adjustments in GFR and creatinine using comprehensive reference creatinine calculations.[12] By defining a relevant classification system with strong associations among clinical trajectories, outcomes and resource use, we can develop standardised methods for predictive modelling and clinical decision support. Our findings suggest that patients from different AKI subgroups have distinct pathophysiological mechanisms related to hypotension, hyperglycaemic, thrombocytopaenia and hypoalbuminaemia. It remains plausible that these elements represent therapeutic targets for specific AKI subtypes. Systematic investigation of preventative and therapeutic strategies tailored to AKI trajectories may yield more consistent and generalisable results than diffuse, non-standardised investigations using variable classification systems.

Our single-institution design limits generalisability to other practice settings. As a retrospective study, our results may be influenced by selection bias. We sought to minimise selection bias by including all consecutive hospital admissions meeting relatively broad inclusion criteria. Due to lack of data on contrast agents and home medications, these were not reported and due to limitations on accurate data on urine output, only serum creatinine definition was used for defining AKI. Finally, biomarkers evaluated herein were limited to those collected for routine clinical use. Future investigations should seek development and validation of models that predict AKI trajectories at the time of hospital admission with subsequent dynamic predictions, and assess the efficacy of targeted preventative and therapeutic measures for patients at high risk for persistent AKI.

CONCLUSIONS

Among hospitalised patients and ICU cohorts, persistent AKI and the absence of renal recovery were associated with poor short-term and long-term survival, independent of AKI severity. Accurate and early identification of patients at increased risk for persistent AKI may facilitate the

provision of targeted treatments that prevent persistent AKI or promote renal recovery to improve survival and optimise resource use.

Author affiliations
[1]Department of Medicine, University of Florida, Gainesville, Florida, USA
[2]Precision and Intelligent Systems in Medicine (PrismaP), University of Florida, Gainesville, Florida, USA
[3]Department of Surgery, University of Florida, Gainesville, Florida, USA
[4]Department of Health Outcomes and Biomedical Informatics, University of Florida, Gainesville, Florida, USA
[5]Department of Industrial and Systems Engineering, University of Florida, Gainesville, Florida, USA
[6]Department of Biostatistics, University of Florida, Gainesville, Florida, USA

Contributors AB is the guarantor of this manuscript. TOB and AB had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. Concept and design: TOB and AB. Acquisition, analysis or interpretation of data: TOB, YR, EA, SM, TJL and AB. Drafting of the manuscript: TOB, TJL, YR, EA, SM, HH, RI and AB. Critical revision of the manuscript for important intellectual content: AB, TOB, TJL, YR, EA, SM, HH, RI, RM, SG, EAS, PP, BB and MSS. Obtained funding: TOB and AB. Supervision: AB, EAS, PP, BB and MSS.

Funding TOB was supported by K01 DK120784, R01 DK123078, and R01 DK121730 from the National Institute of Diabetes and Digestive and Kidney Diseases (NIH/NIDDK), R01 GM110240 from the National Institute of General Medical Sciences (NIH/NIGMS), R01 EB029699 from the National Institute of Biomedical Imaging and Bioengineering (NIH/NIBIB), and R01 NS120924 from the National Institute of Neurological Disorders and Stroke (NIH/NINDS). AB was supported R01 GM110240 from the National Institute of General Medical Sciences (NIH/NIGMS), R01 EB029699 and R21 EB027344 from the National Institute of Biomedical Imaging and Bioengineering (NIH/NIBIB), R01 NS120924 from the National Institute of Neurological Disorders and Stroke (NIH/NINDS), and by R01 DK121730 from the National Institute of Diabetes and Digestive and Kidney Diseases (NIH/NIDDK). TJL was supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number K23 GM140268. BB and MSS were supported by Sepsis and Critical Illness Research Center Award P50 GM-111152 from the National Institute of General Medical Sciences. Additionally, the Research reported in this publication was supported by the National Center for Advancing Translational Sciences of the National Institutes of Health under University of Florida Clinical and Translational Science Awards UL1TR000064 and UL1TR001427. National Institute of Diabetes and Digestive and Kidney Diseases (NIH/NIDDK), National Institute of General Medical Sciences (NIH/NIGMS), National Institute of Biomedical Imaging and Bioengineering (NIH/NIBIB), National Institute of Neurological Disorders and Stroke (NIH/NINDS) and National Center for Advancing Translational Sciences (NIH/NCATS).

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval The study was designed and approved by the Institutional Review Board of the University of Florida and the University of Florida Privacy Office (IRB 201901123).

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available on reasonable request.

ORCID iD
Azra Bihorac http://orcid.org/0000-0002-5745-2863

REFERENCES

1 Darmon M, Ostermann M, Cerda J, *et al*. Diagnostic work-up and specific causes of acute kidney injury. *Intensive Care Med* 2017;43:829–40.
2 James MT, Bhatt M, Pannu N, *et al*. Long-Term outcomes of acute kidney injury and strategies for improved care. *Nat Rev Nephrol* 2020;16:193–205.
3 Sawhney S, Fraser SD. Epidemiology of AKI: utilizing large databases to determine the burden of AKI. *Adv Chronic Kidney Dis* 2017;24:194–204.
4 Gardner AK, Ghita GL, Wang Z, *et al*. The development of chronic critical illness determines physical function, quality of life, and long-term survival among early survivors of sepsis in surgical ICUs. *Crit Care Med* 2019;47:566–73.
5 Chawla LS, Bellomo R, Bihorac A, *et al*. Acute kidney disease and renal recovery: consensus report of the acute disease quality initiative (ADQI) 16 Workgroup. *Nat Rev Nephrol* 2017;13:241–57.
6 Bhatraju PK, Mukherjee P, Robinson-Cohen C, *et al*. Acute kidney injury subphenotypes based on creatinine trajectory identifies patients at increased risk of death. *Crit Care* 2016;20:372.
7 Ozrazgat-Baslanti T, Loftus TJ, Mohandas R, *et al*. Clinical trajectories of acute kidney injury in surgical sepsis: a prospective observational study. *Ann Surg* 2020. doi:10.1097/SLA.0000000000004360. [Epub ahead of print: 13 Nov 2020].
8 Bellomo R, Kellum JA, Ronco C, *et al*. Acute kidney injury in sepsis. *Intensive Care Med* 2017;43:816–28.
9 Morrell ED, Kellum JA, Pastor-Soler NM, *et al*. Septic acute kidney injury: molecular mechanisms and the importance of stratification and targeting therapy. *Crit Care* 2014;18:501.
10 Poston JT, Koyner JL. Sepsis associated acute kidney injury. *BMJ* 2019;364:k4891.
11 Seymour CW, Kennedy JN, Wang S, *et al*. Derivation, validation, and potential treatment implications of novel clinical phenotypes for sepsis. *JAMA* 2019;321:2003–17.
12 Ozrazgat-Baslanti T, Hobson C, Motaei A. Development and validation of computable phenotype to identify and characterize kidney health in adult hospitalized patients. *arXiv preprint* 2019;2604673.
13 Kidney Disease: Improving Global Outcomes (KDIGO) Acute Kidney Injury Work Group. Clinical practice guideline for acute kidney injury. *Kidney inter, Suppl* 2012;2:1–138.
14 Khwaja A. KDIGO clinical practice guidelines for acute kidney injury. *Nephron Clin Pract* 2012;120:c179–84.
15 Kidney Disease: Improving Global Outcomes (KDIGO) Acute Kidney Injury Work Group. KDIGO 2012 clinical practice guideline for the evaluation and management of chronic kidney disease. *Kidney inter, Suppl* 2013;3:1–150.
16 Selby NM, Hill R, Fluck RJ, *et al*. Standardizing the early identification of acute kidney injury: the NHS England national patient safety alert. *Nephron* 2015;131:113–7.
17 Bellomo R, Ronco C, Kellum JA, *et al*. Acute renal failure - definition, outcome measures, animal models, fluid therapy and information technology needs: the Second International Consensus Conference of the Acute Dialysis Quality Initiative (ADQI) Group. *Crit Care* 2004;8:R204–12.
18 Závada J, Hoste E, Cartin-Ceba R, *et al*. A comparison of three methods to estimate baseline creatinine for RIFLE classification. *Nephrol Dial Transplant* 2010;25:3911–8.
19 Levey AS, Stevens LA, Schmid CH, *et al*. A new equation to estimate glomerular filtration rate. *Ann Intern Med* 2009;150:604–12.
20 Ahmed S, Nutt CT, Eneanya ND, *et al*. Examining the potential impact of race multiplier utilization in estimated glomerular filtration rate calculation on African-American care outcomes. *J Gen Intern Med* 2021;36:464–71.
21 Diao JA, Wu GJ, Taylor HA, *et al*. Clinical implications of removing race from estimates of kidney function. *JAMA* 2021;325:184–6.
22 Norris KC, Eneanya ND, Boulware LE. Removal of race from estimates of kidney function: first, do no harm. *JAMA* 2021;325:135–7.
23 Chen S. Retooling the creatinine clearance equation to estimate kinetic GFR when the plasma creatinine is changing acutely. *J Am Soc Nephrol* 2013;24:877–88.
24 Dolgun A, Demirhan H. Performance of nonparametric multiple comparison tests under heteroscedasticity, dependency, and skewed error distribution. *Commun Stat Simul Comput* 2017;46:5166–83.10.1080/03610918.2016.1146761

25 Kellum JA, Sileanu FE, Bihorac A, *et al*. Recovery after acute kidney injury. *Am J Respir Crit Care Med* 2017;195:784–91.
26 Sood MM, Shafer LA, Ho J, *et al*. Early reversible acute kidney injury is associated with improved survival in septic shock. *J Crit Care* 2014;29:711–7.
27 Bhatraju PK, Zelnick LR, Chinchilli VM, *et al*. Association between early recovery of kidney function after acute kidney injury and long-term clinical outcomes. *JAMA Netw Open* 2020;3:e202682.
28 Abdel-Nabey M, Ghrenassia E, Mariotte E, *et al*. Acute kidney injury recovery patterns in critically ill patients: results of a retrospective cohort study. *Crit Care Med* 2021;49:e683–92.

BMJ Health & Care Informatics

Association between the chronology of gestation and the morphometrical skin characteristics at childbirth: a development of predictive model

Ingrid Michelle Fonseca de Souza,[1] Gabriela Luiza Nogueira Vitral,[2] Marcelo Vidigal Caliari,[3] Zilma Silveira Nogueira Reis [4]

To cite: de Souza IMF, Vitral GLN, Caliari MV, *et al*. Association between the chronology of gestation and the morphometrical skin characteristics at childbirth: a development of predictive model. *BMJ Health Care Inform* 2021;**28**:e100476. doi:10.1136/bmjhci-2021-100476

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/bmjhci-2021-100476).

Received 25 August 2021
Accepted 10 November 2021

[1]Obstetrics and Gynecology, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil
[2]Pediatrics, Postgraduation Program, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil
[3]Department of General Pathology, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil
[4]Informatics Health Center, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil

Correspondence to
Dr Zilma Silveira Nogueira Reis; zilma.medicina@gmail.com

ABSTRACT

Objective The structural maturation of the skin is considered a potential marker of pregnancy dating. This study investigated the correlation between the morphometrical skin characteristics with the pregnancy chronology to propose models for predicting gestational age.

Methods A cross-sectional analysis selected 35 corpses of newborns. The biopsy was performed up to 48 hours after death in the periumbilical abdomen, palm and sole regions. Pregnancy chronology was based on the obstetric ultrasound before 14 weeks. The dimensions of the skin layers, area of glands and connective fibrous tissue were measured with imaging software support. Univariate and multivariate regression models on morphometric values were used to predict gestational age.

Results Gestational age at birth ranged from 20.3 to 41.2 weeks. Seventy-one skin specimens resulted in the analysis of 1183 digital histological images. The correlation between skin thickness and gestational age was positive and strong in both regions of the body. The highest univariate correlation between gestational age and skin thickness was using the epidermal layer dimensions, in palm ($r=0.867$, $p<0.001$). The multivariate modelling with the thickness of the abdominal epidermis, the dermis and the area of the sebaceous glands adjusted had the highest correlation with gestational age ($r=0.99$, $p<0.001$).

Conclusion The thickness of the protective epidermal barrier is, in itself, a potential marker of pregnancy dating. However, sets of values obtained from skin morphometry enhanced the estimation of the gestational age. Such findings may support non-invasive image approaches to estimate pregnancy dating with various clinical applications.

Summary

What is already known?

- Morphometric invasive analysis of fetal skin provides a visual examination of architectural patterns according to gestational age.
- Non-invasive ultrasound imaging indicates the epidermal thickness of the newborn's skin as one evolutionary indicator of the gestational chronology.

What does this paper add?

- Non-invasive analysis of newborn skin imaging can estimate the dating of pregnancy with various clinical applications.
- The protective epidermal barrier was, in itself, a potential marker of pregnancy dating through skin thickness imaging analysis.
- The multivariate model, including the thickness of the abdominal epidermis, the dermis, and the area of the sebaceous glands, had the highest correlation with gestational age.

INTRODUCTION

The anatomy of the human skin shows a clear relationship between its structure and function.[1] When well-differentiated, the skin provides a physical and immune barrier essential to newborn survival.[2] Skin's barrier function is mainly due to the stratum corneum which is a layer composed of flattened and differentiated corneocytes terminally separated by layers of densely compacted lipides.[1 3] Studies using skin biopsy are relevant to improve knowledge about the protective barrier during the perinatal period.[4 5] However, the specimen is difficult to obtain,[6] and the preparation of slides can result in artefacts and require multiple tissue samples.[6 7] Even so, microscopic methods with staining procedures allow to outline specific components and measure them in order to portray tissue modifications over time.[8 9]

It is not surprising that the chronology of pregnancy is considered the main indicator of newborn survival.[10] There are critical clinical relationships between epidermal barrier competence and neonatal survival, faced with the risk of hypothermia and infections.[4] Histological analysis suggests that epidermal development becomes complete in utero at approximately 34 gestational weeks but will only become functional in the first week of

life.[11] Preterm newborns with gestational age <37 weeks have the thinnest epidermis and a less developed functional barrier than full-term newborns,[12] being thus poorly prepared to face the extra-utero environment.[11] These have high rates of water loss and transcutaneous heat loss, in addition to the difficulty in maintaining homeostasis and having a deficient impermeable barrier.[13]

Visible changes in the clinical examination of the newborn's skin and also in a histological study of this tissue demonstrate that the functional and structural maturation of the skin is a potential marker of the chronology of pregnancy.[14 15] A non-invasive ultrasound imaging study indicates the thickness of the newborn's skin as one of the evolutionary indicators that can be objectively measured to estimate the gestational chronology.[7] In fact, the determining of gestational age with greater accuracy can positively affect perinatal results,[10 16] as it will direct the most appropriate interventions in neonatal care.[17] Furthermore, the chronology of gestation is the basis for the statistics of prematurity and nutritional status of the newborn, guiding public policies, which includes the analysis of perinatal mortality.[18] Nonetheless, the determination of gestational age at birth is not a trivial task since it is directly affected by access to high-cost technology, such as obstetric ultrasound, and by the imprecision of postnatal maturity clinical scores.[19] New approaches have been proposed, among them the analysis of skin maturity through its optical properties.[20]

This study investigated the correlation between the thickness of the skin layers, area of glands and fibrous connective tissue of the skin in corpses of newborns with the chronology of pregnancy to propose models for predicting gestational age based on morphometry values.

METHODS AND MATERIALS

Environment and subjects

Feasibility study evaluated 35 corpses of newborns, stillbirths or dead after birth, prospectively selected in accordance with the eligibility criteria, from January 2016 to September 2019. Based on the expectation of a linear correlation between epidermal thickening and gestational age,[7] a minimum sample of 17 bodies was calculated to detect a positive and moderate correlation, assuming an alpha error of 5% and one 20% beta error in a two-tailed hypothesis test. They met inclusion criteria as follows: childbirth with gestational age between 20 and 42 weeks of gestation, calculated using the crown-rump length measure ultrasonography-based reference, performed before 14 weeks of gestation.[10] In the case of stillbirths, the estimated interval between fetal death and childbirth was up to 3 days. For alive newborns selected after decease, the extra uterine life after birth did not exceed 48 hours of age, and biopsy was possible within 24 hours after neonatal death. Exclusion criteria were structural skin alterations or conditions that modify the skin, such as anhydramnios, hydrops, congenital skin diseases and clinical evidence of chorioamnionitis as maternal fever or foul-smelling amniotic fluid; tissue maceration assessed at the visual inspection of the corpses; oedema or autolysis verified during histological analysis.

The skin biopsy and tissue processing

Human skin specimens were withdrawn from three body regions: over the thenar eminence of palm (palm), over the periumbilical abdominal area and over the calcaneus area (sole of the foot). Punch biopsies cut a circle of 1 cm^2 of diameter with sufficient depth to reach the full skin thickness and partial hypodermis. The conventional

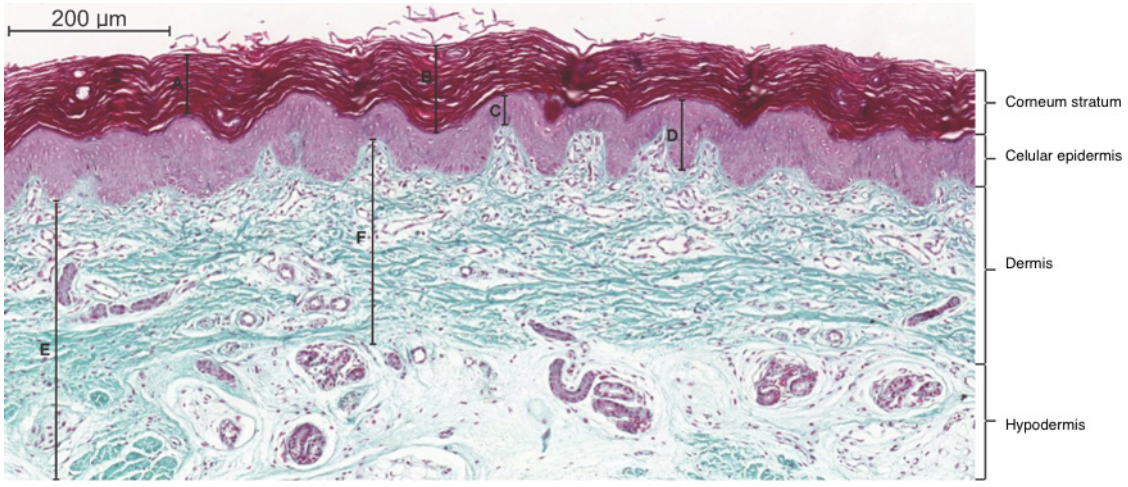


Figure 1 Photomicrograph of the skin on the bottom of the foot of stillbirth at 40 gestational weeks. A represents the measurement of the stratum corneum with a lower limit corresponding to the apex of the epidermal crest. B represents the measurement of the stratum corneum with a lower limit corresponding to the valley of the epidermal papillae. C represents the measurement of the cellular epidermis with a lower limit corresponding to the apex of the dermis. D represents the measurement of the epidermis with a lower limit corresponding to the valley of the epidermal papillae. E represents the measurement of the hypodermis with an upper limit corresponding to the valley of the epidermal papillae. F represents the measure of the upper limit dermis corresponding to the crest of the dermal papilla. Gomori trichrome. Bar=200 mm.

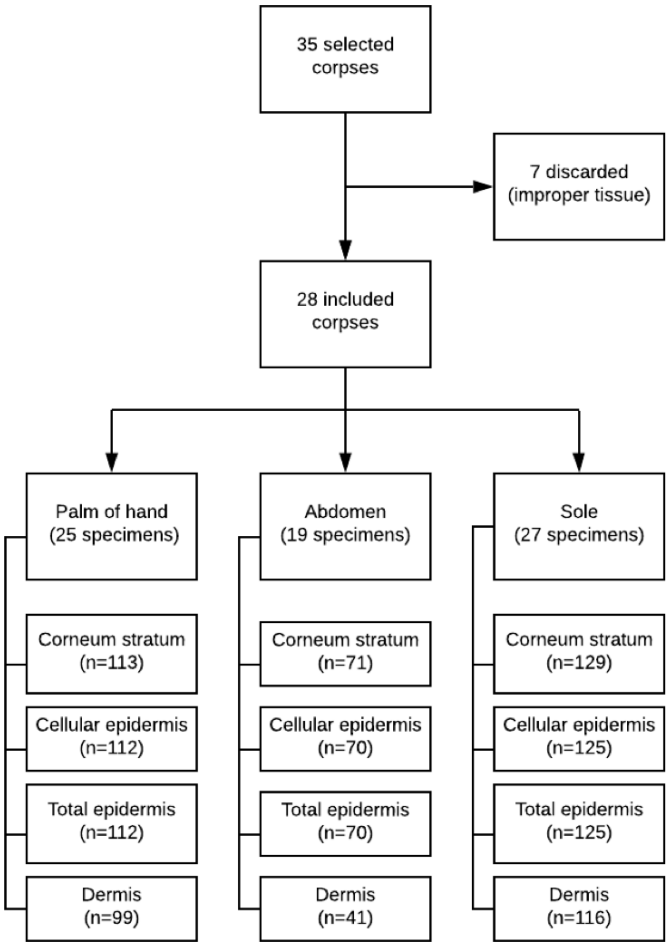


Figure 2 Flowchart diagram detailing the number of analysed images, according to the skin over body areas.

histological preparation included a 10% neutral formalin fixation and 5 µm tissue sections of blocks embedded in paraffin. In addition, the histological slides were stained by Gomori's trichrome.

Morphometric analysis of the skin

The thickness of the epidermis, dermis, area of the sebaceous and sweat glands were measured, as well as the area of fibrous tissue. A3DHISTECH Pannoramic MIDI (Budapest, Hungary) scanner and Pannoramic Viewer software captured images of the slides. From each slide, 2–5 frames with an objective magnification of ×10 were selected according to image quality criteria, tissue integrity and presence of all skin layers and part of the hypodermis. We set algorithms in the KS300 software of analysis contained in the Carl Zeiss image analyzer (Oberkochen, Germany) to semi-automatically explore the image, based on Caliari procedures.[21] Epidermal measurements included the thickness of the epidermal layer and the corneum stratum, with the boundary in the image delineated by the observer. The epidermis was identified by its darker colour and stratified keratinocytes, figure 1. Dermal layer thickness corresponded to the measurement from the epidermal–dermal junction to the dermal–hypodermal limits. The average of five smaller and five larger measures were obtained interactively to average represents the thickness and within variance.

A dermal sector with 7.7×10^5 µm^2 was obtained by selecting pixels with shades of green, creating a binary image and using digital processing to calculate the dermal fibrous connective tissue area. We set algorithms in the KS300 software of analysis, based on Prata *et al*.[22] Interactive measurements of each sweat or sebaceous glands were obtained separately, within a dermal and hypodermal sector with 7.27×10^5 µm^2, based on procedures described by Costa *et al*.[23]

Statistical analyses

Descriptive statistics assessed the clinical characteristics of the newborns and skin morphometry variables. Depending on the data distribution, quantitative variables were presented as averages, SDs, medians (minimum and maximum) or IQRs. The coefficient of variation and the 95% CI were calculated by bootstrap to allow inference based on the skin morphometry sample data. Qualitative variables were presented as absolute values and percentages. Univariate and multivariate regression analyses assessed the correlation between gestational age and skin morphometry for each area on the body where skin biopsy was performed. Using the stepwise approach, multiple regression analysis included significant ($p<0.05$) predictor variables from the univariate models. Durbin-Watson test of residuals evaluated the fit of the models. Coefficient of determination (adjusted R^2) was carried based on the hypothesis that it was zero. The SPSS V.22.0 was used for the analysis. P values of less than 0.05 were considered to be significant.

RESULTS

From 35 enlisted corpses, seven did not meet the quality criteria of the skin tissues during histological analysis. Twenty-eight selected newborns gathered 12 (57.14%) after birth and 16 (42.86%) stillbirths. Figure 2 presents details from the enrollment of the newborns to the imagery, according to the assessed segment of the body.

Gestational age ranged from 20.3 to 41.2 weeks of gestation. Clinical characteristics of newborns are described in table 1. The main cause of death was major malformation, accounting for 16 (57.1%—line 3) newborns. There was no difference between stillborn and deaths after childbirth newborns, in relation to the cause of death ($p=0.313$, line 2), gestational age ($p=0.252$, line 7), birth weight ($p=0.252$, line 8), birth weight centile ($p=0.840$, line 9) and sex ($p=0.215$, line 10). Among 21 fetuses with gender determination and gestational age at birth equal or above 24 weeks, seven had birth weight below the 10th percentile for gestational age, according to the Intergrowth 21st standard,[24] three of them stillbirths and four dead after delivery. Two stillbirths had birth weights below the third percentile for gestational age.

The thickness of the newborn's skin layers

One thousand hundred and eighty-three skin images were analysed from 71 slides. The dimensions of the

Table 1 Clinical characteristics of newborns

Characteristics	Stillbirths (n=12)	Dead after delivery (n=16)	P value
Causes of death			0.313*
Major malformation, n (%)	5 (17.86)	11 (39.28)	
Fetal distress, n (%)	2 (7.14)	1 (3.57)	
Diabetes, n (%)	0 (0)	1 (3.57)	
Unknown or others, n (%)	7 (58.3)	3 (18.75)	
Gestational age (weeks), average (SD)	33.1 (17.53)	35.2 (19.8)	0.252†
Birth weight (g), average (SD)	1237.5 (2770)	1935 (3175)	0.252†
Birth weight centile, average (SD)‡	36.1 (39.2)	32.9 (32.6)	0.840†
Sex			0.215*
Male, n (%)	6 (21.43)	3 (10.71)	
Female, n (%)	5 (17.86)	11 (39.28)	
Undetermined, n (%)	1 (3.57)	2 (7.14)	

*χ^2 test.
†Mann-Whitney test.
‡According to the Intergrowth 21st standard for gestational age ≥24 weeks.[24]

skin layers, their intrinsic variations and comparisons between areas of the body are presented in table 2. The median epidermal thickness on the skin over the palm was similar to that of the sole: 152.1 (43.9–251.9) µm and 146.2 (56.2–276.4) µm (p=0.618), respectively, lines 11 and 12. However, the median thickness of the dermal layer was higher over the periumbilical abdominal area 724.0 (287.0–1107.0) µm, line 16, than sole 396.3 (174.0–493.2) µm, line 15 and palm 384.1 (166.0–751.0) µm, line 14, p<0.001. The standardised variability of measurements for layers of the skin had high value in skin layers over the periumbilical abdominal area, lines 11, 12 and 13.

The area of fibrous connective tissue of the skin over periumbilical area 0.259×10^6 µm^2 (0.093–0.526) had a median value similar to that of the sole 0.235×10^6 µm^2

Table 2 Dimensions of the skin layers at birth, with comparisons between the assessed areas of the body

	Median (95% CI)	Min–Max	CV* (%)	Comparisons		
				P value†	P value‡	P value§
Thickness of the corneum stratum (µm)						
Palm	63.6 (21.3 to 81.9)	6.1–154.5	32.9	0.707		
Sole	72.4 (7.6 to 176.0)	7.6–176.0	34.1			0.002
Periumbilical abdominal area	18.0 (8.0 to 43.4)	8.0–43.4	46.6		0.010	
Epidermal thickness (µm)						
Palm	72.0 (33.0 to 101.7)	33.0–101.7	44.2	0.701		
Sole	78.8 (41.2 to 128.5)	41.2–128.5	41.6			<0.001
Periumbilical abdominal area	44.3 (19.0 to 61.2)	19–61.2	41.7		<0.001	
Epidermal total thickness (µm)						
Palm	152.1 (43.9 to 251.9)	43.9–251.9	77.1	0.618		
Sole	146.2 (122.6 to 170.3)	56.2–276.4	74.5			<0.001
Periumbilical abdominal area	66.0 (28.1 to 99.5)	28.1–99.5	85.9		<0.001	
Dermal thickness (µm)						
Palm	384.1 (166.0 to 751.0)	166.0–751.0	21.9	0.977	0.002	<0.001
Sole	396.3 (174.0 to 493.2)	174.0–493.2	20.3			
Periumbilical abdominal area	724.0 (287.0 to 1107.0)	287.0–1107.0	18.7			

*CV: average of the coefficient of variation obtained for each image.
†Difference between palm and sole areas.
‡Difference between palm and periumbilical abdominal area.
§Difference between a sole and periumbilical abdominal area.

Table 3 Concentration of fibrous tissue and glands of the skin at birth, with comparisons between the assessed areas of the body

	Median (Min–Max)	Comparisons		
		P value*	P value†	P value‡
Area of fibrous connective tissue (10^6 μm^2)				
Palm	0.248 (0.069–0.346)	1		
Sole	0.235 (0.008–0.524)			0.708
Periumbilical abdominal area	0.259 (0.093–0.526)		0.817	
Area of sweat glands (10^6 μm^2)				
Palm	0.097 (0.028–0.173)	0.718		
Sole	0.088 (0.033–0.242)			<0.001
Periumbilical abdominal area	0.025 (0.010–0.061)		<0.001	
Area of sebaceous glands (10^6 μm^2)				
Periumbilical abdominal area	0.294 (0.020–3.652)	–	–	–

*Difference between palm and sole.
†Difference between palm and periumbilical area.
‡Difference between sole and periumbilical area.

(0.008–0.524) and palm 0.248×10^6 μm^2 (0.069–0.346), p=0.708 and p=0.817, respectively (table 3, lines 4, 5 and 6). However, the median value of the area of the sweat glands in the skin over periumbilical area, 0.294×10^6 μm^2 (0.020–3.651), was higher than that in the palm 0.097×10^6 μm^2 (0.028–0.172) or sole 0.088×10^6 μm^2 (0.033–0.242), lines 8 and 9, p<0.001, for both comparisons.

The correlation between the gestational age and morphometry of the skin at birth is presented in table 4. Scatter plots with the linear correlation of each morphometric variable with the gestational age are in online supplemental file S1 to S13. In the univariate analysis, the epidermal thickness layer highlighted as the dimension strongly associated with gestational age: in the skin over palm (r=0.867, p<0.001, line 3), periumbilical abdominal area (r=0.806, p<0.001, line 8) and sole (r=0.712, p<0.001, line 14). The fibrous connective tissue (lines 5, 10 and 16), sweat or sebaceous glands areas had mild or absent correlations with the gestational age (lines 6, 11, 12 and 17). However, compositions of

Table 4 Predictive models for gestational age, based on morphometry values of the skin at birth

	Univariate analysis	Multivariate analysis	
	Linear coefficient (P value)	Adjusted coefficient of correlation	P value of the model
Skin over palm			
Epidermal thickness (μm)	0.867 (<0.001)	0.655	0.94 (p<0.001)
Dermal thickness (μm)	0.805 (<0.001)	0.256	
Area of fibrous connective tissue (μm^2)	0.518 (0.014)	0.169	
Area of sweat glands (μm^2)	–0.143 (0.515)	–	–
Skin of periumbilical abdominal area			
Epidermal thickness (μm)	0.806 (<0.001)	0.559	0.99 (p<0.001)
Dermal thickness (μm)	0.579 (0.038)	–0.216	
Area of fibrous connective tissue (μm^2)	0.538 (0.071)	–	–
Area of sweat glands (μm^2)	0.441 (0.131)	–	–
Area of sebaceous glands (μm^2)	–0.845 (0.001)	–0.646	
Skin over sole			
Epidermal thickness (μm)	0.712 (<0.001)	0.540	0.83 (p<0.001)
Dermal thickness (μm)	0.660 (<0.001)	0.456	
Area of fibrous connective tissue (μm^2)	–0.266 (0.189)	–	–
Area of sweat glands (μm^2)	–0.266 (0.189)	–	–

R-square of multivariate models: 0.87 (palm), 0.97 (abdomen), 0.69 (sole). Durbin-Watson analysis: 1.94 (palm), 1.90 (abdomen), 1.45 (sole).

the morphometric parameters fitted multivariate models better explained the variability of the gestational age than univariate correlations. Considering the skin of the periumbilical area, the composition formed by the thickness of the epidermis, dermis and the area of sebaceous glands showed an excellent correlation with gestational (r=0.99, p<0.001, line 8). Concern the skin over the hand and sole, the multivariate model grouping morphometry parameters also enhanced the model of prediction of gestational age, concerning the univariate models: adjusted r=0.94, p<0.001 (line 3), and r=0.99, p<0.001 (line 8).

DISCUSSION

Main findings

In this study, the main contribution was to correlate dimensions measured by morphometry of the skin of a newborn with its gestational age, a new knowledge that can objectively estimate the chronology of pregnancy from histology. The processing of images and the synthesis of values with inferential statistics on the measurements of layers, sublayers, gland area and fibrous connective tissue allowed the development of mathematical models of prediction. In addition, the study documented the intra-subject variability of these measures, numerically reflecting the ripple of the skin layers, guided by the dermal papillae. Regarding the external validity, the selected sample gathered newborns with a wide range of gestational age from extreme prematurity, 20.3 weeks, to term, 41.6 weeks. Although major malformations were responsible for most deaths (57.1%), conditions associated with changes in skin structure were excluded in the recruitment phase.

Regarding morphometric measurements, the results fill a knowledge gap in the study of human skin in this age group, including samples of premature births. In a systematic review published by De-Souza *et al*,[4] similar studies that provide measurements of newborn skin thickness were considered insufficient to describe morphometry in a reproducible and detailed manner. In addition to the care with microscopic measurements, the chronology of pregnancy was calculated based on early obstetric ultrasound examination, considered a reference standard for pregnancy dating.[10]

There are numerous challenges of inaccurate calculation of pregnancy chronology by available clinical methods,[19] and this is also a motivation using of fetal skin histology in pregnancy dating. The proposed models of prediction of gestational age may support the investigation of perinatal death and support non-invasive studies with similar applications.[7 20] Infant mortality has at preterm birth, one of the major current challenges of obstetric and neonatal care.[19 25] Although the approach is invasive, using skin biopsy in the corpses of newborns, the process brought an opportunity to estimate the chronology of pregnancy, at the time of death, from the morphometry of the skin of specific regions and technique. The histological analysis of the skin, through the visual analysis of architectural patterns, the tissues already proved predictive of gestational age in a previous study,[26] without, however, presenting quantitative elements that allow the dating.

Comparison with prior studies

In relation to the magnitude of the measurements, the thickness of the epidermis was greater in the region of the palm and sole of the foot, in relation to the periumbilical region. This finding confirms previous reports that in these places, the stratification of the epidermis is earlier and more intense than in other regions of the body.[26 27] The early and progressive multiplication of the epidermis in these places may explain the strong correlation found between the thickness of the skin layers and the chronology of pregnancy, even as an isolated marker. However, the comparability of the values found with previous reports is hampered by the incomplete description of the various measures and techniques already published in the scientific literature. Measurements of part of the sublayers, for example, the thickness of the epidermis without including the stratum corneum, only dermis thickness[9] and measurements made in different places of the body and ages of the children studied.[27–29] Besides, the measurement of epidermal thickness, according to Kakasheva-Mazhenkovska *et al*,[30] was 193.2 µm in the sole of the foot, 161.6 µm in the abdomen and 142.0 µm in the hand, comparable to the present study. The measurements of the epidermis described here also corroborate the findings of a non-invasive study that performed measurements of different sites of the body of newborns through high-frequency ultrasound,[7] which showed values of the thickness of the epidermis in the region of the sole of the foot were 175.4 (17.6) µm. In the dermal layer, we obtained values apparently lower than 873.0 µm in the palm, 719.9 µm in the sole and 1297.0 µm in the abdomen.[30] We attribute these differences to variations in technique and gestational age of the samples.

More recently, Dhingra *et al* analysed four regions of the body of 30 fetuses from 11 to 40 weeks of gestation. The epidermal thickness had a significant positive correlation with gestational age.[31] Our study corroborated such results of a strong correlation with gestational age in the skin over the abdomen and palm. However, this study did not combine variables and nor assess gland area and fibrous connective tissue in the prediction as to the current approach.

Limitations and highlights

The main limitation of this study was the strict eligibility criteria for pregnancy dating and tissue quality, which made it challenging to obtain the postmortem specimen, considered rare.[32] On the other hand, we emphasise that the multivariate models achieved high correlation coefficients for groups of morphometric measures, 0.94 in the palm region, 0.99 in the abdomen region and 0.83 in the sole, table 4. In addition, the objective measurement of several tissue components such as the area of connective

tissue and glands, to estimate gestational age, is unprecedented. Therefore, the mathematical models have the potential to automate the analysis process and may facilitate in the future the obtaining of gestational age information from the systematised analysis of a histological image of the skin. In addition, we believe that future studies may find utility in the results presented in this analysis in tissue engineering, simulation models of the skin, mainly subsidising more appropriate care with the newborn's skin.

Besides, seven corpses had birth weight below the 10th percentile for gestational age and two below the third percentile. Even fetal growth reference standards are suboptimal for stillbirths,[33] the influence of fetal malnutrition in the dimensions of deep layers of the skin is possible. However, the skin surface seems not to be influenced by fetal nutrition. In a prior study, Vitral *et al* analysed 222 alive newborns at birth, with gestational age ranging from 24 to 41 weeks of gestation, using high-frequency ultrasound, and epidermal thickness was not fetal growth standard dependent.[7]

CONCLUSIONS

Skin morphometry, especially the measurement of layer thickness, proved to be an essential marker of gestational age at birth. The representation of structural changes in the skin in composite mathematical models involving various elements of this tissue proved to be promising automating of the pregnancy dating process from histological images.

Contributors IMFdS and GLNV collected clinical data, interpreted and analysed the images, and wrote and revised the article. ZSNR and MVC designed the study, interpreted and analysed the data, and wrote and revised the paper. ZSNR is the guarantor.

Funding This research was supported by the Grand Challenges Exploration from the Bill & Melinda Gates Foundation (Grant number OPP1128907, Contract) http://www.gatesfoundation.org/ and Fundação de Amparo a Pesquisa de Minas Gerais (Grant number AUC00015-17), Brazil http://wwwfapemigbr/en/.

Disclaimer The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests These authors declare a patent deposit number BR1020200215736, on behalf of the Universidade Federal de Minas Gerais and Fundação de Amparo a Pesquisa de Minas Gerais. The inventors were ZSNR, IMFdS and MVC.

Patient consent for publication Consent obtained from parent(s)/guardian(s)

Ethics approval The Research Council of Universidade Federal de Minas Gerais, Hospital das Clínicas, and Hospital Sofia Feldman, Brazil reviewed and independently approved the study protocol. The ethical approval number is CAAE 44834915.3.0000.5149. Parents were enlightened regarding this scientific investigation and signed an informed consent form on behalf of their newborns.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available upon reasonable request. Raw data were generated at Universidade Federal de Minas Gerais, Bazil. Derived data supporting the findings of this study are available from the corresponding author ZSNR on request.

ORCID iD

Zilma Silveira Nogueira Reis http://orcid.org/0000-0001-6374-9295

REFERENCES

1 Roger M, Fullard N, Costello L, *et al*. Bioengineering the microanatomy of human skin. *J Anat* 2019;234:438–55.
2 Byrne C, Hardman M, Nield K. Covering the limb–formation of the integument. *J Anat* 2003;202:113–23.
3 Menon GK, Cleary GW, Lane ME. The structure and function of the stratum corneum. *Int J Pharm* 2012;435:3–9.
4 de-Souza IMF, Vitral GLN, Reis ZSN. Skin thickness dimensions in histological section measurement during late-fetal and neonatal developmental period: a systematic review. *Skin Res Technol* 2019;25:793–800.
5 Gambichler T, Matip R, Moussa G, *et al*. In vivo data of epidermal thickness evaluated by optical coherence tomography: effects of age, gender, skin type, and anatomic site. *J Dermatol Sci* 2006;44:145–52.
6 Steven B, Hoath HIM. *Neonatal skin: structure and function*. Second ed, Revised and Expanded ed, 1982.
7 Vitral GLN, Aguiar RAPL, de Souza IMF, *et al*. Skin thickness as a potential marker of gestational age at birth despite different fetal growth profiles: a feasibility study. *PLoS One* 2018;13:e0196542.
8 Oriá RB, Ferreira FVA, ÉN S, *et al*. Study of age-related changes in human skin using histomorphometric and autofluorescence approaches. *An Bras Dermatol* 2003;78:425–34.
9 Smith LT, Holbrook KA, Byers PH. Structure of the dermal matrix during development and in the adult. *J Invest Dermatol* 1982;79 Suppl 1:93–104.
10 American College of Obstetricians and Gynecologists. Committee opinion no 611: method for estimating due date. *Obstet Gynecol* 2014;124:863–6.
11 Kalia YN, Nonato LB, Lund CH, *et al*. Development of skin barrier function in premature infants. *J Invest Dermatol* 1998;111:320–6.
12 King A, Balaji S, Keswani SG. Biology and function of fetal and pediatric skin. *Facial Plast Surg Clin North Am* 2013;21:1–6.
13 Hardman MJ, Sisi P, Banbury DN, *et al*. Patterned acquisition of skin barrier function during development. *Development* 1998;125:1541–52.
14 Oranges T, Dini V, Romanelli M. Skin physiology of the neonate and infant: clinical implications. *Adv Wound Care* 2015;4:587–95.
15 Waaijer MEC, Gunn DA, Catt SD, *et al*. Morphometric skin characteristics dependent on chronological and biological age: the Leiden longevity study. *Age* 2012;34:1543–52.
16 World Health Organization. *Born too soon: the global action report on preterm birth*, 2012.
17 Karl S, Li Wai Suen CSN, Unger HW, *et al*. Preterm or not--an evaluation of estimates of gestational age in a cohort of women from Rural Papua New Guinea. *PLoS One* 2015;10:e0124286.
18 Silveira MF, Santos IS, Barros AJD, *et al*. Aumento dA prematuridade no Brasil: revisão de estudos de base populacional. *Revista de Saúde Pública* 2008;42:957–64.
19 Lee AC, Panchal P, Folger L, *et al*. Diagnostic accuracy of neonatal assessment for gestational age determination: a systematic review. *Pediatrics* 2017;140:e20171423.
20 Reis ZSN, Vitral GLN, de Souza IMF, *et al*. Newborn skin reflection: proof of concept for a new approach for predicting gestational age at birth. A cross-sectional study. *PLoS One* 2017;12:e0184734.
21 Caliari MV. *Princípios de Morfometria digital: KS300 para iniciantes: Editora UFMG*, 1997.
22 Prata LO, Rodrigues CR, Martins JM, *et al*. Original research: ACE2 activator associated with physical exercise potentiates the reduction of pulmonary fibrosis. *Exp Biol Med* 2017;242:8–21.
23 Costa CAX, Brito KNO, Gomes MA, *et al*. Morphometric study of the hepatic lesions experimentally induced in hamsters by Entamoeba dispar and E. histolytica. *Parasite* 2007;14:329–34.

24 Papageorghiou AT, Kennedy SH, Salomon LJ. International Fetal and Newborn Growth Consortium for the 21(st) Century (INTERGROWTH-21(st)). The INTERGROWTH-21st fetal growth standards: toward the global integration of pregnancy and pediatric care. *Am J Obstet Gynecol* 2018;218:S630–40.
25 Howson CP, Kinney MV, McDougall L, *et al*. Born too soon: preterm birth matters. *Reprod Health* 2013;10 Suppl 1:S1.
26 Ersch J, Stallmach T. Assessing gestational age from histology of fetal skin: an autopsy study of 379 fetuses. *Obstet Gynecol* 1999;94:753–7.
27 Hashimoto K, Gross BG, DiBella RJ, *et al*. The ultrastructure of the skin of human embryos. IV. The epidermis. *J Invest Dermatol* 1966;47:317–35.
28 Fairley JA, Rasmussen JE. Comparison of stratum corneum thickness in children and adults. *J Am Acad Dermatol* 1983;8:652–4.
29 Albideri A. Histological and Cytoarchitectural measurements of human epidermis in different anatomical sites of embryonic, fetal and neonatal Iraqi subjects in Al-Hilla/Iraq maternity hospital. *J Pharm Sci Res* 2018;10:812–8.
30 Kakasheva-Mazhenkovska L, Milenkova L, Kostovska N, *et al*. Histomorphometrical characteristics of human skin from capillitium in subjects of different age. *Prilozi* 2011;32:105–18.
31 Dhingra S, Asghar A. *The epidermal growth pattern in human fetuses*, 2020.
32 Hoath SB, Maibach HI. *Neonatal skin: structure and function*. CRC Press, 2003.
33 Xun Z, Joseph KS, Cnattingiuset S. Birth weight differences between preterm stillbirths and live births: analysis of population-based studies from the US and Sweden. *BMC Pregnancy Childbirth* 2012;1:1–7.

BMJ Health & Care Informatics

Are child health information services a viable source of accurate vaccination data for clinicians working in paediatric emergency departments in England?

Rachel Isba ,[1,2] Nigel Davies,[3] Jo Knight [1]

To cite: Isba R, Davies N, Knight J. Are child health information services a viable source of accurate vaccination data for clinicians working in paediatric emergency departments in England? *BMJ Health Care Inform* 2021;**28**:e100486. doi:10.1136/bmjhci-2021-100486

Received 27 September 2021
Accepted 11 December 2021

[1]Lancaster Medical School, Lancaster University, Lancaster, UK
[2]Emergency Department, North Manchester General Hospital, Manchester University NHS Foundation Trust, Manchester, UK
[3]School of Computing and Communication, Lancaster University, Lancaster, UK

Correspondence to
Professor Rachel Isba; rachel.isba@nhs.net

ABSTRACT

Vaccination is a global success story, yet UK coverage remains undertarget for a number of diseases. The paediatric emergency department (PED) offers the potential for opportunistic vaccination interventions.

Objectives To map the Greater Manchester (GM) Child Health Information System network to see if it was a viable source of vaccination data for clinicians working in the PED as a case study.

Methods Postprimary care vaccination management systems for GM were visualised using a systems mapping approach, with data obtained from the Office for National Statistics and commissioners in the GM Health and Social Care Partnership.

Results Once vaccination data left primary care, it passed through 1 of 10 local child health information services (CHISs), using an assortment of different information technology systems, after which it shed individual identifiers and was aggregated within national systems. None of the existing GM CHISs were accessible to PED practitioners.

Conclusion More work needs to be done to explore possible alternative sources of accurate vaccination data during a PED consultation.

INTRODUCTION

Vaccination remains one of the great global public health successes. Since their discovery more than 300 years ago, vaccines have saved countless millions of lives,[1] reduced the incidence of dozens of diseases and even led to the eradication of smallpox.[2] However, in the UK, uptake of routine childhood vaccinations (provided by the National Health Service (NHS) at no cost to the parent/carer) has fluctuated over recent years and remains below WHO targets for a number of vaccinations (eg, the measles, mumps and rubella (MMR) vaccine).[3] This finding is on a background of global changes in the pattern of vaccination and an associated increase in outbreaks of vaccine-preventable diseases, further compounded by disruptions to delivery of routine vaccination programmes during the SARS-CoV-2/COVID-19 pandemic.

Every year in England, millions of children and young people (CYP) attend the paediatric emergency department (PED)[4] and may sometimes have a long wait to see a healthcare professional. In addition to their primary reason for presentation, CYP attending the hospital may have unmet health need (eg, sexual health) or may not be able to access preventive elements of routine healthcare (eg, vaccination) for a myriad of reasons. A hospital attendance might therefore be an opportunity to improve health, beyond the initial reason for presentation, and early work has shown that this would be an acceptable approach to parents/carers.[5]

If any child or young person who have not had their age-appropriate vaccinations is identified during a PED attendance, clinicians may (should it be clinically/situationally appropriate) be able to offer one or more tailored interventions to address this.[6] The benefits of such an approach are numerous and include ensuring appropriate management, for example, in the case of a tetanus-prone wound (where management depends on vaccination status), and increasing community coverage in case of an outbreak of a vaccine-preventable disease, for example, measles.

However, in order to be able to intervene with those at greatest risk of being undervaccinated, it is first necessary to be able to identify them in a timely and accurate way, given the time-limited interaction in the department and departmental pressures. Guidance recommends that professionals 'Check the immunisation status of CYP at every appropriate opportunity'.[7] In the PED, therefore, all practitioners should routinely enquire of parents/carers accompanying a child or young person if they have had all their age-appropriate vaccinations. However, past work has shown that often no question

is asked or recorded in the notes,[8] and if the enquiry is made, it is usually done in a superficial way via a question such as 'Have they had all their vaccinations?'. When asked, parents/carers tended to overestimate vaccination coverage.[5]

In contrast, in primary care, if a child attends a general practitioner (GP) appointment, the clinician is alerted, via the presence of a 'pop-up', if the child is not up to date with his or her vaccinations. The difference here is that the vaccination data are held within the same system as the GP records, but the hospital systems are separate. In the UK, the majority of routine childhood immunisations are offered in community locations, commonly delivered via settings such as a GP surgery. Administration of one or more vaccines will be recorded in the GP electronic system, with returns sent from these systems to the local child health information service (CHIS) and then on to the central surveillance system.

The objective of this work was to map the CHIS network in Greater Manchester (GM) to assess its potential as a source of accurate vaccination data for clinicians working in PEDs across the region, given the issues with obtaining information from parents/carer. This work was carried out as part of a bigger project looking at the potential for a PED-delivered vaccination intervention.

METHODS

The work was carried out in GM, England. The Office for National Statistics (ONS) mid-2019 estimates were used to describe the GM population of CYP <16 years old.[9] Names of local authorities (LAs) and associated CHISs, the provider organisations for each CHIS and the data management systems used were obtained via requests to GM Health and Social Care Partnership (GMHSCP), the organisation responsible for commissioning vaccination services in GM. Lists of LAs, CHIS and provider organisations (where relevant) were compiled and then combined with ONS data using systems mapping,[10] an approach commonly used in public health. The map in figure 1 (which represents the structure of the system in GM in mid-2020) was created using Microsoft Visio V.2016 and fact-checked by GMHSCP before the names of individual organisations and information technology (IT) systems were removed (to protect commercially sensitive information).

RESULTS

In GM, a population of around 582 000 CYP had their vaccination data held by 10 different CHISs, provided by four different organisations, using three different national IT management systems commissioned in GM (although this has recently been reduced to two). Figure 1 shows the population served (by LA), the CHIS holding and managing data for each population, and the provider organisations commissioned to manage multiple CHISs (where relevant). Flow of vaccination data is represented by directional arrows (labelled with the IT system used).

No CHIS was accessible to practitioners working in secondary care (each system is password protected and only accessible to those working in community-based services), nor was there a focal point for GM that would have acted as a meaningful target for connecting the CHISs to secondary care data systems (aside from issues of interoperability) as none of the CHISs were connected to each other (even if managed by the same provider organisation). Once the vaccination data left GM CHISs, they shed individual identifiers and progressed up the national system in an aggregated anonymised format.

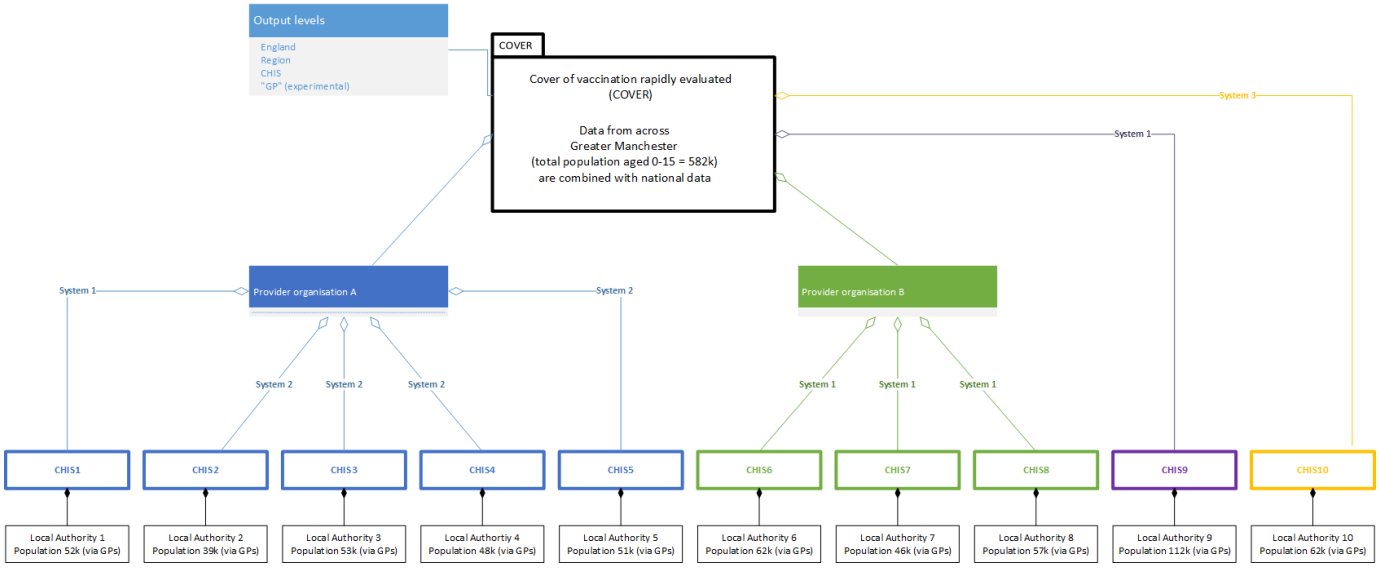


Figure 1 Management of data relating to vaccination in children and young people (aged <16 years old) in Greater Manchester. The names of the local providers and systems have been anonymised. Population=Office for National Statistics 2019 mid-year estimate for those aged 0–15 years inclusive to the nearest 1000. CHIS, child health information service; GP, general practitioner.

DISCUSSION

CYP attending settings such as the PED may benefit from interventions to improve vaccination coverage; however, it is not currently possible to reliably identify those who are not up to date. Although parent/carer recall remains the most common source of vaccination data during a PED consultation, clinicians often do not take a (meaningful) vaccination history and parent/carer recall tends towards overestimation.[5 8 11] An alternative approach is needed for checking vaccination status for all CYP as part of routine care but would also add value in special circumstances, such as those where subsequent medical management might be altered by the child's vaccination status (eg, tetanus) or in controlling outbreaks of vaccine-preventable diseases (eg, measles).

Potential alternative sources of data include the Red Book (a handheld paper or electronic record of child health), the GP summary care record (where available and accessible), phoning GP surgeries (on an individual patient basis) and local CHIS. This work has used a system mapping to approach to show that, while an individual CHIS may contain accurate vaccination data, it is inaccessible to hospital-based clinicians and also part of a prohibitively complex system with no single focal point, so it does not represent a viable option in GM at the current time. The simplest solution might be a unified regional CHIS, but that is a commissioning decision beyond the influence of secondary care clinicians. A limitation of the study is that it used only a single mapping approach to visualise the data. Another potential limitation is that GM has a commissioning structure which may not be replicated elsewhere, so collating the CHIS data may be more complex in other settings.

Future work will look at the potential for accessing primary care-held vaccination data (eg, via summary care records) as an alternative. However, preliminary work suggests that while these records are technically accessible, extracting relevant data takes a disproportionate amount of time as the vaccination data are unstructured and only interpretable by someone with an extensive working knowledge of the NHS childhood vaccination schedule.

Until a viable (in terms of time and effort for clinicians), accurate and real-time alternative to parent/carer recall is available, it is not going to be possible to progress to delivering an intervention to those CYP who are undervaccinated at the time of their attendance to the PED.

CONCLUSIONS

The PED offers an underused opportunity to deliver interventions to improve the wider health and well-being of patients, with vaccination being an example of such an intervention. However, the lack of access to reliable vaccination data in a timely fashion, during a PED attendance, means that it is not currently possible to identify those CYP in need of an intervention. The complex structures of postprimary care data management mean that in GM, the CHISs, while considered the definitive source of vaccine data, are wholly inaccessible in their current form and are therefore not a viable source of vaccination information for clinicians working in the PED.

Twitter Rachel Isba @dr_asbo

Acknowledgements Thanks to colleagues on the Great Manchester Screening and Immunisation Team for their invaluable input into the system mapping work.

Contributors RI conceived of the study, collected and mapped the data, and was involved in all stages of preparation of this manuscript. JK and ND were involved in all stages of the design of the study and in writing and revising the manuscript.

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

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval This study does not involve human participants.

Provenance and peer review Not commissioned; externally peer reviewed.

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

ORCID iDs
Rachel Isba http://orcid.org/0000-0002-2896-4309
Jo Knight http://orcid.org/0000-0002-7148-1660

REFERENCES

1 Plotkin S. History of vaccination. *Proc Natl Acad Sci U S A* 2014;111:12283–7.
2 World Health Organization. Smallpox. Available: https://www.who.int/westernpacific/health-topics/smallpox [Accessed 06 Sep 2021].
3 European Centre for Disease Prevention and Control. Vaccination coverage for the second dose of measles-containing vaccine, EU/EEA, 2018. Available: https://www.ecdc.europa.eu/en/publications-data/vaccination-coverage-second-dose-measles-containing-vaccine-eueea-2018 [Accessed 06 Sep 2021].
4 NHS Digita. Hospital Accident & Emergency Activity 2019-20, 2021. Available: https://digital.nhs.uk/data-and-information/publications/statistical/hospital-accident-emergency-activity/2019-20 [Accessed 06 Sep 2021].
5 Isba R, Edge R. Delivery of a multi-focus public health intervention in the paediatric emergency department: a feasibility and acceptability pilot study. *BMJ Open* 2021;11:e047139.
6 Edge R, Isba R. Interventions delivered in secondary or tertiary medical care settings to improve routine vaccination uptake in children and young people: a scoping review protocol. *JBI Evid Synth* 2020;18:1566–72.
7 National Institute for Health and Care Excellence. Immunisations: reducing differences in uptake in under 19, 2021. Available: https://www.nice.org.uk/guidance/ph21 [Accessed 06 Sep 2021].
8 Newell K, Rousseva C, Slade C, *et al*. Should we offer opportunistic vaccination in the paediatric emergency department? *Emerg Med J* 2015;32:1007.2–8.
9 Office for National Statistics. Population estimates for the UK, England and Wales, Scotland and Northern Ireland: mid-2019, 2020. Available: https://www.ons.gov.uk/releases/populationestimatesforthe ukenglandandwalesscotlandandnorthernirelandmid2019 [Accessed 06 Sep 2021].
10 Egan M, McGill E, Penney T. *NIHR SPHR guidance on systems approaches to local public health evaluation. Part 1: introducing systems thinking*. London: National Institute for Health Research Scholl for Public Health Research, 2019. https://sphr.nihr.ac.uk/wp-content/uploads/2018/08/NIHR-SPHR-SYSTEM-GUIDANCE-PART-1-FINAL_SBnavy.pdf
11 Conway SP. Opportunistic immunisation in hospital. *Arch Dis Child* 1999;81:422–5.