Machine learning models, trusted research environments and UK health data

Kerasidou, Charalampia (Xaroula); Malone, Maeve; Daly, Angela; Tava, Francesco

Published in:
Journal of Medical Ethics

DOI:
10.1136/jme-2022-108696

Publication date:
2023

OPEN ACCESS

Machine learning models, trusted research environments and UK health data: ensuring a safe and beneficial future for AI development in healthcare

Charalampia (Xaroula) Kerasidou ,[1] Maeve Malone,[2] Angela Daly,[3] Francesco Tava[4]

[1]School of Medicine, University of Dundee, Dundee, UK
[2]Dundee Law School, School of Humanities Social Sciences and Law, University of Dundee, Dundee, UK
[3]Leverhulme Research Centre for Forensic Science, School of Science and Engineering, University of Dundee, Dundee, UK
[4]School of Social Sciences, UWE Bristol, Bristol, UK

Correspondence to
Dr Charalampia (Xaroula) Kerasidou, University of Dundee, Dundee DD1 4HN, UK; Ckerasidou001@dundee.ac.uk

Received 12 October 2022
Accepted 11 March 2023

To cite: Kerasidou CX, Malone M, Daly A, *et al*. *J Med Ethics* Epub ahead of print: [please include Day Month Year]. doi:10.1136/jme-2022-108696

ABSTRACT

Digitalisation of health and the use of health data in artificial intelligence, and machine learning (ML), including for applications that will then in turn be used in healthcare are major themes permeating current UK and other countries' healthcare systems and policies. Obtaining rich and representative data is key for robust ML development, and UK health data sets are particularly attractive sources for this. However, ensuring that such research and development is in the public interest, produces public benefit and preserves privacy are key challenges. Trusted research environments (TREs) are positioned as a way of balancing the diverging interests in healthcare data research with privacy and public benefit. Using TRE data to train ML models presents various challenges to the balance previously struck between these societal interests, which have hitherto not been discussed in the literature. These challenges include the possibility of personal data being disclosed in ML models, the dynamic nature of ML models and how public benefit may be (re)conceived in this context. For ML research to be facilitated using UK health data, TREs and others involved in the UK health data policy ecosystem need to be aware of these issues and work to address them in order to continue to ensure a 'safe' health and care data environment that truly serves the public.

INTRODUCTION

There is a broad structural shift taking place in the UK and beyond,[i] which ushers in the increasing digitisation of the health and care sector. This is a shift that is balancing between two diverging yet interconnected developments: an increasing appetite for data-driven and machine learning (ML) healthcare technologies, supported by an innovation-driven research, technology and policy sector; and increasing awareness of the importance of legal and ethical safeguards guiding such innovations to ensure that legal rights and obligations, such as confidentiality and privacy, are protected and upheld along with more ethical approaches including the public's continued collaboration in such endeavours.[1 2] In the UK, trusted research environments (TREs) sit at the junction of these developments, attempting to balance differing interests between the public, research and rights.

TREs, also known as 'data enclaves', 'research data centre/centres' or 'safe havens', are physical or virtual analytical environments which can hold various data sets (such as population, census, or healthcare data, etc). Subject to monitoring and access controls, a TRE user can be allowed to work with these data but is prevented from releasing their analysis without permission. The aim of TREs is to provide a secure location for researchers to analyse data, especially personal data, enabling collaborative and transparent research while protecting data confidentiality and privacy.

While TREs have received relatively little attention in academic literature and debate, especially from ethical perspectives,[3 4] they are not new.[ii] Some have been in operation for almost 20 years now. Yet, they have recently come to the limelight as a key service for the UK's National Health Service (NHS) data which can engender public trust by facilitating the intensifying demand for sensitive data for research purposes while ensuring privacy, confidentiality and safe access.[5–7]

Alongside the increasing prominence of TREs sits the drive for health data to feed into artificial intelligence (AI). AI is the prevailing umbrella term to refer to a range of computational techniques that can be used to make machines complete tasks in a way that would be considered intelligent were they to be completed by a human. Here we specifically refer to ML developments. This is a particular form of AI which involves computers 'learning' and adapting without specific instructions, doing so by using algorithms to analyse and draw inferences from data. With the UK's healthcare sector being positioned as a unique data-rich ecosystem that could wield significant medical advances due to its centralised nature and longitudinal population data[8] and as a lucrative business opportunity potentially worth several billions,[9] there is a concerted push to realise the UK's plans to become a global

[i]For example, in June 2022, the Department of Health and Social Care published its data strategy for health and care in England titled 'Data Saves Lives: Reshaping Health and Social Care with Data'.[5] In 2021, Scotland published its updated digital health and care strategy.[46] In Europe, the European Commission is working on the eHealth programme.[47]

[ii]For example, the Secure Research Service run by the ONS (ONS SRS) has been operational since 2003 and has provided the blueprint for many subsequent TREs. In Scotland, a system of safe havens (four regional and one national safe haven) has been in operation for over a decade now. In Wales, the secure research platform SeRP was created to store data for the SAIL Databank that collects and manages all public sector data of Wales, since 2005.

technological superpower[10] via its national healthcare system. This increased interest in the application of ML on sensitive data (ie, special category personal data[iii]), such as healthcare and medical data, means that TREs are increasingly approached with requests to use their data to develop new types of outputs, such as trained ML models. Such developments present new opportunities but also challenges for the secure and trusted function of TREs.

In order to advance these new opportunities for TREs while maintaining high privacy standards, the Data and Analytics Research Environments UK (DARE UK) Sprint Project titled Guidelines and Resources for AI Model Access from TrusTEd Research environments (GRAIMATTER) investigated the additional risk posed for the disclosure of personal data introduced by the release of trained ML models from TREs, and developed a set of technical, legal and ethical recommendations for how TREs should carry out disclosure control on ML models. Reflecting on our work as the Legal and Ethical project subteam, in this article, we focus on key ethical and legal issues stemming from the training of ML models from TRE data and how this impacts on TREs' operation. While the export of research output from TREs is generally regulated through controls such as manual supervision ('eyeballing') to ensure no personal data leave the TRE, the situation changes when TRE exports present more complex configurations such as ML models which then may be released to open source repositories such as GitHub. In such cases, it becomes harder to identify potential risks using conventional manual checks and therefore harder to guarantee the 'T' of TREs. To address these points, we will (1) explain what a TRE is and how it works. In particular, we will discuss how the 'Five Safes' framework contributes to their use and governance. We will then (2) address the relationship between the digitisation of healthcare in the UK and data-intensive innovations such as AI, and specifically ML, while identifying how TREs are being positioned as a way to ensure much needed public trust in such developments. We will conclude (3) by highlighting three legal and ethical critical areas that require further consideration for TREs and others involved in ML healthcare research for public benefit. This is significant for TREs, and for health data research more broadly, as ML research may disturb the current balance struck between facilitating research and protecting privacy in TREs given the risk of disclosure of sensitive personal data once a trained ML model is exported from the TRE, and the lack of clarity in terms of legal responsibility were a data breach to occur. This also relates to the dynamic nature of models compared with the 'static' nature of traditional TRE outputs, which may also require a more dynamic ethics process accompanying the research, and in turn require a rethinking of the public benefit produced by such research.

WHAT IS A TRE

A TRE is a secure physical or virtual environment designed for approved and named researchers to access sensitive pseudonymised data, where access to specific data sets is provided only to approved research projects. TREs differ from other data use models such as the more traditional *data release model*, where data are made available to approved researchers to download and analyse in their own data environments, hence risking losing control of their security and management. Instead, in a TRE, data are not released externally to data users for analysis on their own computers but placed on a server within a restricted, secure information technology environment, where the approved user is given secure access to carry out their project analysis. No row-level data leave the TRE environment. Traditionally, only aggregate-level results (eg, summary tables, graphs, statistical models) are released from a TRE at the end of the project, and only after a range of automatic and manual screening controls are applied to ensure that all outputs are non-disclosive of personal data.

The use of TREs is meant to address the challenges of using health and other forms of sensitive personal data to facilitate research that is assessed to be in the public interest while at the same time protecting privacy and ensuring trustworthiness.[5 6] Importantly, their use does not eliminate the risk of disclosure of sensitive personal data but greatly mitigates it[11] by providing assurances that data are handled securely, as data use can be tracked and technical and organisational measures are in place to check that no data leave the secure environment. Acting as data processors, TREs are meant to maintain a balance between:

- Confidence of data controllers (who determine the purposes for and manner in which any special category personal data are to be processed) through increased security.
- Benefits to the user/researcher (who can be from an academic, commercial or government setting) through improved access to larger data sets.
- Transparency for public and patients (whose personal data are made available in pseudonymised form) as to who has access to the data and for what purposes in order to ensure their continuing confidence and engagement.[7]

Robust data governance is key in achieving and maintaining such a balance. This means meeting all relevant legal obligations (eg, data protection, confidentiality, contracts and intellectual property), technical and cybersecurity requirements, and research ethics and data governance requirements.

There are several related frameworks used for providing guidance on TRE governance[iv] most of which are based on the 'Five Safes' model. 'Five Safes' is an internationally recognised model introduced by the UK Office for National Statistics in 2003. It has been described as an 'explicitly relativistic, subjective and empirical' framework which has proved a 'useful' tool to frame, rather than prescribe, the crucial discussions around governance and management of sensitive data involving data providers, users and regulators.[12] The 'Five Safes' breaks down the decisions surrounding data access and use into five related but separate dimensions:[13v]

[iii]As per Data Protection Act, 2018, section 10 (c).

[iv]For example, the five TREs in Scotland follow the Charter for Safe Havens in Scotland. The charter draws from *the Guiding Principles for Data Linkage* (which in turn draws on human rights legislation, the Data Protection Act, guidance from the Information Commissioner and the Scottish Government Identity Management and Privacy Principles), the *SHIP Blueprint* and associated governance frameworks that define standards and process for the use of non-consented linked data for health informatics research in Scotland. In 2021, the UK Health Data Research Alliance published a set of principles and best practices[7] structured around the 'Five Safes' framework and further inspired by the OECD Guidelines on Human Biobanks and Genetic Research Databases, work of NHSX, NHS Digital, the National Data Guardian and through guidance from the Information Commissioner.

[v]In some cases, the 'Five Safes' model has been extended to include 'Safe Return'[48] which has been coined the 'Five Safes Plus One' approach,[49] and 'Safe Computing' as an extension of 'Safe Setting'.[7]

Safe people
TRE staff and the researchers accessing the data through a TRE are trained and authorised to use the data safely, follow guidelines and report data safety concerns, if any.

Safe projects
Through an initial ethical and data governance approval process, TREs ensure that the research projects are approved by data controllers, and that data are used appropriately and for public benefit.

Safe outputs
TREs screen all outputs thoroughly and approve the release only after ensuring that it does not include personal data.

Safe data
The data are deidentified/pseudonymised before access is granted to researchers. It is ensured that researchers only see the data that they need to.

Safe setting
TREs provide a safe environment to access personal data and prevent any unauthorised use.

AI AND THE DIGITISATION OF HEALTHCARE: IMPROVING SAFETY THROUGH THE USE OF TRES

Healthcare has been identified as 'one of the most important sectors for AI both for better services and for better efficiency'.[14 15] This has paved the way for new, and arguably controversial, public–private–academic partnerships[16] for the development of new AI technologies, including ML, which can be used in several healthcare areas such as diagnostics, therapeutics, population health management and administration, and for providing key infrastructure for the storage, maintenance and management of the data that underpin these technologies.[17]

Such developments align with the ongoing efforts since 2002 towards the digitisation of the NHS—from (missed) aims of achieving a 'paperless' NHS by 2018 to the renewed target for a 'core level of digitisation' by 2024[18]—which has resulted in a rich and valuable wealth of healthcare data. The COVID-19 pandemic has reconfirmed and further accelerated plans for the digitisation of the NHS (ie, NHS apps, virtual appointments, online treatments, etc), along with recent plans to facilitate the more effective sharing of digital health and social care records and data.[19] Further plans to personalise healthcare through the use of wearable technologies and apps will only enrich these data sets.

AI, and in particular ML, technologies for healthcare rely on the availability of big data for their training and development. As such, the extensive medical and healthcare data that result from the ongoing interactions between the UK public and the NHS have long been seen as a prime opportunity for the adoption of innovative AI technologies, for day-to-day patient care and for the further advancement of health research.

While the opportunities that the increasing digitisation of healthcare offers appear exciting, the risks and concerns that such developments entail are considerable. There have been a series of situations where public trust has been eroded in data sharing—some of which have attracted significant media attention and regulatory enforcement, while others may be more 'routine' infractions of contracts. Nevertheless, these instances cumulatively may instil a negative attitude in the public towards data sharing. Among these, past big data health initiatives, such as care.data—an English initiative designed to allow the repurposing of primary care medical data for research and other purposes—and more recently the postponed GP Data for Planning and Research programme,[20] demonstrate the importance of public trust for major projects which seek to aggregate and centralise healthcare and related data, and the costly danger of losing it if legitimate public concerns are not taken seriously.[21 22] Scandals, such as the ongoing case of DeepMind/Google and the Royal Free,[vi] along with a recent report in the *BMJ* that there are hundreds of organisations such as clinical commissioning groups, private companies and universities which have breached patient sharing agreements, some of them with little or no consequences[23] (see also ref 24 25), demonstrate that what is often termed the 'deficit of public trust'[26] is not the result of public ignorance or badly publicised information[21 27] (see also ref 28 29). Instead, it is an appropriate response of a public who, while willing for their data to be used for the benefit of patients and the NHS, are wary of a weak regulatory landscape that allows such data security failures.[2 30] Nevertheless, this emerging picture has received limited attention in the academic literature and limited discussion as to how it may impact on data sharing and governance arrangements and policies more broadly as the NHS seeks to digitise and share more data.[vii]

To address the issue of the protection of personal data and the facilitation of research, TREs are positioned as a way to maintain and restore public trust.[5–7] The technical and organisational safety measures that TREs offer can provide assurances that, not only will data not be leaving their secure environments but every interaction and subsequent analysis will be checked and tracked. Furthermore, their commitment to *Safe Projects* means that each project is assessed by an ethical and data governance committee for their potential to public benefit before a project approval is granted.[viii] In order to ensure public benefit and build public trust, the importance of patient and public participation alongside transparency of decision-making and data use has been further highlighted.[7 31 32]

However, while TREs may be identified as the appropriate way to address public trust concerns, our research shows that the increase in the development and adoption of AI, and in particular ML, in the medical and healthcare fields presents new challenges for the next generation of TREs which may threaten the 'T' of the TREs due to additional risks of disclosure of personal data by ML models trained on TRE data and a lack of clarity about chains of responsibility once the ML model has left the TRE environment.

[vi]Here we are referring to the now infamous case of the Royal Free releasing, in 2015, millions of their patient data to the AI company DeepMind (later acquired by Google) for the development of a medical app without the appropriate legal and ethical safeguards for their protection. In 2017, the UK's ICO sanctioned the Royal Free for breach of UK data protection law. While back then Google avoided sanctions, it is currently being sued in a private litigation class action lawsuit for the unlawful use of patients' confidential medical data[50–53] (see also ref[54-56]).

[vii]The merging of NHS England and NHS Digital in February 2023 will provide ample scope for complicating a complex structure further, https://digital.nhs.uk/about-nhs-digital/nhs-digital-merger-with-nhs-england (accessed 1 Feb 2023).

[viii]The make-up of these committees differs between TREs while depending on the project's design, methods and data needs further governance approvals might be deemed necessary (eg, NHS R&D, or Caldicott approvals).

TRE OUTPUTS AND ML

Typically, TRE outputs take the form of aggregated results, graphs and tables. Before their release, these outputs go through both automatically and manually disclosive controls and are checked to ensure that no identifiable information is attached to them before allowed to leave the TRE. With the increased interest in ML trained on special category data such as healthcare and medical data, TREs are increasingly approached with requests to use their data for such purposes and to disclose new types of outputs, such as ML models trained on TRE data.

As the GRAIMATTER research demonstrates, the release of trained ML models from TREs introduces an additional risk for the disclosure of personal data.[13 33] In other words, while on the one hand models are being constructed using training data, on the other, training data and/or a semblance or subset of it, or information about who was in the training set can also, in certain cases, be reconstructed from a model. This means that trained ML models may be considered as containing personal data and therefore constitute personal data sets, bringing them within the jurisdiction of data protection legislation.[34] Personal data disclosure from trained ML models can happen inadvertently—for example, if the ML algorithm is overtrained, and the weights of the algorithm which are then exported from the TRE correspond to the data underneath—or, there can be malicious intent—for example, when a malicious researcher 'hides' individual-level data within the files (eg, sensitive data could be embedded in the weights of an ML algorithm which are then exported).[33ix]

In order to mitigate any risk of direct or indirect personal data breach from the disclosure of trained ML models, and hence maintain public trust, key aspects of the technical, legal and ethical governance of TREs need to be reconsidered. Our project GRAIMATTER explored these challenges and proposed a range of measures and recommendations that need to be considered for the safe disclosure of trained ML models from TREs.[13] We focused specifically on the ethical and legal governance issues arising from such practices recommending ways that they can be addressed. Here we present some broader issues that informed our thinking.

LEGAL AND ETHICAL CHALLENGES

While TREs have been positioned as the safer response to the riskier and controversial data release model, it is important to highlight that they are not a magic bullet. All their technical controls notwithstanding, they too are complex sociotechnical systems which, each in their own ways, bring together people, technology, regulations, institutional bodies, auditing and organisational procedures in a sophisticated but always precarious balance that seeks to facilitate research access to data while preserving privacy. The introduction of ML in such a setting requires us to rethink carefully whether and how a new balance can be achieved. In the paragraphs that follow, we highlight three critical areas that require further consideration if we want to ensure a 'safe' health and care data environment that truly serves the public, namely the possibility of disclosure of personal data by ML models once they have left the TRE, the dynamic nature of ML models and the impact that these and other factors have on the discussions around public benefit.

Disclosure of personal data in TREs

While TREs only make available pseudonymised data for research purposes, pseudonymisation is a risky process that can lead to reidentification when combined with additional information. This is a known risk that can be mitigated by contractual agreements between TREs and researchers within the legal framework of the Data Protection Act 2018 which covers data that, if processed, could lead to reidentification within the definition of 'personal data'. As per section 171 of the Data Protection Act 2018, fines and criminal penalties are meant to act as a deterrent to any researcher who would attempt to reidentify them.[x]

The case of ML models within TREs complicates matters. As our GRAIMATTER project team has demonstrated, there is indeed a risk that an ML model leaving the TRE could be disclosing data that could lead to reidentification and therefore constitute personal data.[13] However, it is debatable whether the ML model, per se, could be classified as 'personal data', and hence fall under the data protection framework or not.[34 35] If it is personal data, there is a legal responsibility that both the risks of the specific projects and the controls taken to mitigate them should be clearly specified before the release of the model. Currently, there is a lack of guidance from the Information Commissioner's Office on what form these controls should take, and on whom this responsibility falls. While these issues could be addressed by drawing new contractual agreements, or updating existing ones between the data controllers and the researcher, it is important for the relevant regulatory body to provide clear and updated guidance on a national level to address such risks.[xi]

Dynamic nature of ML models

The dynamic nature of ML models means that their life does not come to an end after one application. After the model leaves the TRE it might move between different applications and uses. Its interaction with different data sets might render data identifiable further down the line. However, by then, the chain of legal responsibility may be unclear and existing legal frameworks do not yet provide sufficient guidance on how the changing and dynamic nature of ML algorithms and models can be regulated.[xii]

Besides legal issues, the dynamic nature of these models also raises ethical concerns as it makes it difficult to identify and assess the risks that the TRE export of ML models may entail. Typically, the ethical assessment process conducted by TREs relies on assessing the benefits but also potential risks of the proposed project before judging whether approval should

[ix]Other threats which can result in the recovery or reconstruction of personal data, including special category personal data, after the ML model has left the TRE are membership inference attacks and model inversion attacks.[13 33]

[x]Notably, public bodies especially in healthcare have rarely been the subject of fines from data protection regulators. The incoming UK Information Commissioner announced in June 2022 that fines would only be issued to public authorities 'in the most egregious cases'.[57]

[xi]The UK Government published a Data Protection and Digital Information Bill in June 2022, which represents its vision for reform of data protection law in the UK post-Brexit. With the change of prime ministers, this Bill has now been withdrawn, but it did include provisions relevant to research, which would have made it easier for researchers to use personal data for research purposes with an inverse effect on the privacy and data protection rights of the individuals whose data are being used.[58] If the Government takes up data protection reform again, such reform needs instead to strengthen these rights and adequately address the risks posed by research especially vis-a-vis AI and ML.

[xii]Interventions such as the proposed AI Act in the European Union[59] aim to address this issue but currently there is no similar legislation or legislative proposals in the UK.

be granted. The element of unpredictability introduced by the disclosure of ML models risks undermining this process as it is impossible to determine whether future interactions of the ML model with different data sets after its TRE release might introduce new risks or what their level might be. Therefore, the traditional ethical process governing research using TREs needs to be rethought and new strategies must be developed that can respond to the challenges that ML models pose.[36] For example, instead of limiting the ethical assessment process to the application stage, a more dynamic approach whereby multiple ethical checks are conducted before and after the project is developed, and as an ML model is released, might prove more appropriate.[37] Should these regular checks reveal a modification to the risk of data breach, a new overall ethical assessment should be undertaken in order to minimise future damage.[xiii]

Rethinking public benefit

The concept of public benefit has been identified as the 'critical safeguard' for the safe and appropriate use of health and care data,[31] and in TREs it is key for the delivery of *Safe Projects*. However, if we are to take this concept and our commitment to it seriously, we need to calibrate the public debate to more accurately reflect the risks, difficulties, unknowns and harms that surround data-intensive healthcare research, especially involving AI and ML, along with the asymmetrical ways that these are distributed.

Indeed, despite early warnings of the hype that surrounds AI in healthcare (and beyond), it is not often we hear about the unpredictable ways that AI healthcare technologies can fail.[38] Or about the scarcity of actual clinical trials to prove the safety, the clinical potential or the efficacy of AI medical tools.[39 40] Beyond the excitement, there is little expert knowledge on how operational changes or changes in the diversity and volume of data can impact on the performance of AI algorithms that are already in use with the potential of seriously undermining patient safety,[41] or few public conversations about the trade-off between AI efficacy and data privacy (ie, the more accurate an AI algorithm, the less private it is). A recent policy report warned that 'attention-grabbing' AI technologies can sometimes 'crowd-out', in terms of funding, other conventional but still essential work in a chronically underfunded NHS,[6] while others warn that algorithms are already creating and worsening health inequities.[42 43] There is scarce discussion that the intense computational processes that AI and big data technologies rely on have a big, and unevenly distributed, environmental impact that needs to be factored in,[44] or that the 'essential infrastructures'[45] that they require are to be delivered by the private corporations that the public has repeatedly warned against.

As transparency is central in ensuring public benefit and building public trust, we need an honest, grounded and sophisticated public discussion about what is the public benefit that underwrites these AI and ML developments, how and by whom it is being assessed and at what and whose cost. While this should not mean ignoring the benefits that such innovations could bring forth, it may mean that current ways of assessing and ensuring public benefit in TRE operation and use of TRE data need to be rethought in light of the potential benefits but also broader challenges of AI development and use.

[xiii]Notably, adopting such a dynamic approach to ethics is likely to require significant reform of existing ethics processes and significantly more resources than at present.

CONCLUSION

In this article, we have highlighted developments in health data and ML research and policy vis-a-vis TREs in the UK. We identified how TREs are being positioned at the junction between an increasing appetite for data-driven and ML healthcare technologies and an increasing awareness of the importance of legal and ethical safeguards guiding such innovations. Drawing from our work on the GRAIMATTER project which explored the additional risks when disclosing trained ML models from TREs,[13] we first explained what TREs are and how the 'Five Safes' framework contributes to their use and governance. We then addressed the relationship between the digitisation of healthcare in the UK and data-intensive innovations such as AI, and in particular ML, while identifying how TREs are being positioned as a way to ensure much needed public trust in such developments. We concluded this article by highlighting three broad legal and ethical critical areas that require further consideration if we want to ensure a 'safe' health and care data environment that truly serves the public: (1) the risk of personal data being disclosed in ML models, (2) the dynamic nature of ML models and (3) how public benefit may be (re)conceived in this context. We argue that these broad critical areas require further thought from TREs and others involved in the UK health data policy ecosystem if they want to ensure a truly 'safe' health and care data environment that indeed serves the public while facilitating AI and ML research on UK health data.

Acknowledgements We acknowledge and thank the project's PI Emily Jefferson and the rest of our colleagues and collaborators on the GRAIMATTER project. The views expressed in this paper are the authors' alone and do not necessarily reflect those of the GRAIMATTER team.

Contributors All of the co-authors contributed to the conception of the work. CK wrote the initial draft and led the writing of subsequent drafts. AD, MM and FT reviewed and contributed to the various drafts. All co-authors approved the final draft of the paper and act as guarantors for the overall content.

Funding This study was funded by UK Research and Innovation (grant number: MC_PC_21033).

Competing interests None declared.

Patient consent for publication Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

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

ORCID iD

Charalampia (Xaroula) Kerasidou http://orcid.org/0000-0002-9794-8492

REFERENCES

1 Department for Digital, Culture, Media & Sport. National data strategy. 2020. Available: https://www.gov.uk/government/publications/uk-national-data-strategy/national-data-strategy [Accessed 1 Feb 2023].
2 Ada Lovelace Institute. Who cares what the public think? UK public attitudes to regulating data and data-driven technologies. 2022. Available: https://www.adalovelaceinstitute.org/evidence-review/public-attitudes-data-regulation/ [Accessed 1 Feb 2023].
3 Graham M, Milne R, Fitzsimmons P, *et al*. Trust and the goldacre review: why trusted research environments are not about trust. *J Med Ethics* 2022. 10.1136/jme-2022-108435 [Epub ahead of print 23 Aug 2022].
4 Affleck P, Westaway J, Smith M, *et al*. Trusted research environments are definitely about trust. *J Med Ethics* 2022. 10.1136/jme-2022-108678 [Epub ahead of print 02 Nov 2022].
5 Department of Health and Social Care. Data saves lives: reshaping health and social care with data. 2022. Available: https://www.gov.uk/government/publications/data-saves-lives-reshaping-health-and-social-care-with-data [Accessed 1 Oct 2022].

6 Goldacre B, Morley J. *Better, broader, safer: using health data for research and analysis. A review commissioned by the secretary of state for health and social care*. Department of Health and Social Care, 2022. Available: https://www.gov.uk/government/publications/better-broader-safer-using-health-data-for-research-and-analysis
7 UK Health Data Research Alliance and NHSX. *Building trusted research environments - principles and best practices; towards TRE ecosystems (1.0)*. Zenodo, 2021.
8 Ghafur S, Fontana G, Halligan J, *et al*. *NHS data: maximising its impact on the health and wealth of the united kingdom*. Imperial College London, 2020.
9 Wayman C, Hunerlach N. Realising the value of health care data: a framework for the future. EY. 2019. Available: https://assets.ey.com/content/dam/ey-sites/ey-com/en_gl/topics/life-sciences/life-sciences-pdfs/ey-value-of-health-care-data-v20-final.pdf
10 HM Government. Industrial strategy: building a britain fit for the future. White paper. 2017. Available: https://www.gov.uk/government/publications/industrial-strategy-building-a-britain-fit-for-the-future [Accessed 1 Oct 2022].
11 Lea NC, Nicholls J, Dobbs C, *et al*. Data safe havens and trust: toward a common understanding of trusted research platforms for governing secure and ethical health research [JMIR medical informatics]. *JMIR Med Inform* 2016;4:e22.
12 Desai T, Ritchie F, Welpton R. Five safes: designing data access for research. Working paper. 2016. Available: https://core.ac.uk/download/pdf/323894811.pdf [Accessed 1 Oct 2022].
13 Jefferson E, Liley J, Malone M, *et al*. Green paper: recommendations for disclosure control of trained machine learning (ML) models from trusted research environments. 2022. Available: https://zenodo.org/record/7089491#.Yyt76HbMI2w [Accessed 1 Oct 2022].
14 House Of Lords Select Committee. AI in the UK: ready, willing and able?. House of lords. 2018. Available: https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf [Accessed 1 Oct 2022].
15 Topol EJ. The topol review: preparing the healthcare workforce to deliver the digital future. 2019. Available: https://topol.hee.nhs.uk/the-topol-review/ [Accessed 1 Oct 2022].
16 Sharon T. The googlization of health research: from disruptive innovation to disruptive ethics. *Per Med* 2016;13:563–74.
17 Murdoch B. Privacy and artificial intelligence: challenges for protecting health information in a new era. *BMC Med Ethics* 2021;22:122.
18 Comptroller and Auditor General (2020). Digital transformation in the NHS. National Audit Office. Available: www.nao.org.uk/report/the-use-of-digital-technology-in-the-nhs [Accessed 1 Oct 2022].
19 Hall R. More healthcare to go online in england under digitisation plan. The guardian. 2022. Available: https://www.theguardian.com/society/2022/jun/29/more-healthcare-online-england-nhs-digitisation-plan [Accessed 1 Oct 2022].
20 Macdonald H. Can the NHS successfully deliver its GP data extraction scheme? *BMJ* 2021;374:n2170.
21 Sterckx S, Rakic V, Cockbain J, *et al*. " you hoped we would sleep walk into accepting the collection of our data ": controversies surrounding the UK care.Data scheme and their wider relevance for biomedical research. *Med Health Care Philos* 2016;19:177–90.
22 Carter P, Laurie GT, Dixon-Woods M. The social licence for research: why care.Data Ran into trouble. *J Med Ethics* 2015;41:404–9.
23 Oxford E. Hundreds of patient data breaches are left unpunished. *BMJ* 2022;377:o1126.
24 NHS Digital. Data sharing remote audit: small area health statistics unit at Imperial College London. 2021. Available: https://digital.nhs.uk/services/data-access-request-service-dars/data-sharing-audits/2021/data-sharing-remote-audit-icl [Accessed 1 Feb 2023].
25 NHS Digital. Data sharing remote audit: glaxosmithkline. 2021. Available: https://digital.nhs.uk/services/data-access-request-service-dars/data-sharing-audits/2021/data-sharing-remote-audit-gsk [Accessed 1 Feb 2023].
26 Morley J, Taddeo M, Floridi L. Google health and the NHS: overcoming the trust deficit. *Lancet Digit Health* 2019;1:e389.
27 McCartney M. Care.data: why are scotland and wales doing it differently? *BMJ* 2014;348:g1702.
28 Felt U, Wynne B. Taking european knowledge society seriously. Report prepared for european commission, directorate-general for research and innovation. 2007. Available: https://op.europa.eu/en/publication-detail/-/publication/5d0e77c7-2948-4ef5-aec7-bd18efe3c442 [Accessed 29 Apr 2021].
29 Kerasidou CX, Kerasidou A, Buscher M, *et al*. Before and beyond trust: reliance in medical AI. *J Med Ethics* 2022;48:852–6.
30 Banner N. NHS data breaches: a further erosion of trust. *BMJ* 2022;377:1187.
31 National Data Guardian. Putting good into practice: a public dialogue on making public benefit assessments when using health and care data. 2021. Available: https://www.gov.uk/government/publications/putting-good-into-practice-a-public-dialogue-on-making-public-benefit-assessments-when-using-health-and-care-data [Accessed 1 Oct 2022].
32 Ada Lovelace Institute. How we work with people. Available: https://www.adalovelaceinstitute.org/about/how-to-work-with-us/publics/ [Accessed 1 Oct 2022].
33 Mansouri-Benssassi E, Rogers S, Smith J, *et al*. Machine learning models disclosure from trusted research environments (TRE), challenges and opportunities. *ArXiv Preprint ArXiv* 2021:2111.05628.
34 Veale M, Binns R, Edwards L. Algorithms that remember: model inversion attacks and data protection law. *Philos Trans A Math Phys Eng Sci* 2018;376:20180083.
35 Leiser MR, Dechesne F. Governing machine-learning models: challenging the personal data presumption. *International Data Privacy Law* 2020;10:187–200.
36 Ada Lovelace Institute. Looking before we leap: expanding ethical review processes for AI and data science research. 2022. Available: https://www.adalovelaceinstitute.org/report/looking-before-we-leap/ [Accessed 1 Feb 2023].
37 Floridi L, Holweg M, Taddeo M, *et al*. A procedure for conducting conformity assessment of AI systems in line with the EU artificial intelligence act. 2022. 10.2139/ssrn.4064091
38 Liu X, Glocker B, McCradden MM, *et al*. The medical algorithmic audit. *Lancet Digit Health* 2022;4:e384–97.
39 Topol EJ. High-Performance medicine: the convergence of human and artificial intelligence. *Nat Med* 2019;25:44–56.
40 van Leeuwen KG, Schalekamp S, Rutten MJCM, *et al*. Artificial intelligence in radiology: 100 commercially available products and their scientific evidence. *Eur Radiol* 2021;31:3797–804.
41 Ross C. AI gone astray: how subtle shifts in patient data send popular algorithms reeling, undermining patient safety. STAT. 2022. Available: https://www.statnews.com/2022/02/28/sepsis-hospital-algorithms-data-shift/ [Accessed 1 Oct 2022].
42 Naumova EN. Public health inequalities, structural missingness, and digital revolution: time to question assumptions. *J Public Health Policy* 2021;42:531–5. 10.1057/s41271-021-00312-y Available: https://doi.org/10.1057/s41271-021-00312-y
43 Moore CMakeda, Candace Makeda M. The challenges of health inequities and AI. *Intelligence-Based Medicine* 2022;6:100067.
44 Crawford Kate. *The atlas of AI: power, politics, and the planetary costs of artificial intelligence*. Yale University Press, 2021.
45 Srnicek N. *Platform capitalism*. Cambridge: Polity Press, 2017.
46 Scottish Government. Digital health and care strategy. 2021. Available: https://www.gov.scot/publications/scotlands-digital-health-care-strategy/pages/2/[Date [Accessed 1 Oct 2022].
47 European Commission. Shaping Europe's digital future: ehealth programme. Available: https://digital-strategy.ec.europa.eu/en/policies/ehealth [Accessed 1 Oct 2022].
48 UK Health Data Research Alliance (UKHDRA). Trusted research environments (TRE): a strategy to build public trust and meet changing health data science needs. Green paper v2.0. 2020. Available: https://ukhealthdata.org/wp-content/uploads/2020/07/200723-Alliance-Board_Paper-E_TRE-Green-Paper.pdf [Accessed 1 Oct 2022].
49 Boniface M, Carmichael L, Hall W, *et al*. The social data foundation model: facilitating health and social care transformation through datatrust services. *Data & Policy* 2022;4. 10.1017/dap.2022.1 Available: https://doi.org/10.1017/dap.2022.1
50 Powles J, Hodson H. Google deepmind and healthcare in an age of algorithms. *Health Technol (Berl)* 2017;7:351–67. 10.1007/s12553-017-0179-1 Available: https://doi.org/10.1007/s12553-017-0179-1
51 Mishcon de Reya. New claim against google and deepmind technologies for unauthorised use of confidential medical records. 2022. Available: https://www.mishcon.com/news/new-claim-against-google-and-deepmind-technologies-for-unauthorised-use-of-confidential-medical-records [Accessed 1 Feb 2023].
52 Martin A. Google sued for using the NHS data of 1.6 million britons 'without their knowledge or consent. Sky news. 2022. Available: https://news.sky.com/story/google-sued-for-using-the-nhs-data-of-16-million-brits-without-their-knowledge-or-consent-12614525?dcmp=snt-sf-twitter [Accessed 1 Oct 2022].
53 Lomas N. Google faces fresh class action-style suit in UK over deepmind NHS patient data scandal. Techcrunch. 2022. Available: https://techcrunch.com/2022/05/16/google-deepmind-nhs-misuse-of-private-data-lawsuit/?guccounter=1&guce_referrer=aHR0cHM6Ly9kdWNrZHVja2dvLmNvbS8&guce_referrer_sig=AQAAAJ5rtQy9-bdx02EnXBuS_ksK76xcnSEqxGo3qmySBAVH1Bfw24xs4WG9ww8eQ4iFktXsOXuxEchtW0FhG5hDqc3fDjb51ZvEhpf_STC5mpFT-tAy9_ay1qWLPhuzyLmdxxslPrSA0YeVDukaqt72odpAwRIkZjvHesbTNkD8tYfW [Accessed 1 Feb 2023].
54 medConfidential. Major health data breaches and scandals. Available: https://medconfidential.org/for-patients/major-health-data-breaches-and-scandals/ [Accessed 1 Feb 2023].
55 Ghafur S, Grass E, Jennings NR, *et al*. The challenges of cybersecurity in health care: the UK national health service as a case study. *Lancet Digit Health* 2019;1:e10–2.
56 ICO. Open letter from UK information commissioner John Edwards to public authorities. 2022. Available: https://ico.org.uk/about-the-ico/media-centre/news-and-blogs/2022/06/open-letter-from-uk-information-commissioner-john-edwards-to-public-authorities/ [Accessed 1 Oct 2022].
57 Amberhawk. Expansive RAS exemption in DPDI bill encourages unethical research. Hawtalk. 2022. Available: https://amberhawk.typepad.com/amberhawk/2022/09/expansive-ras-exemption-in-dpdi-bill-encourages-unethical-research.html [Accessed 1 Oct 2022].
58 European Commission. Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts. 2021. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206 [Accessed 1 Oct 2022].
59 Dyer C. Government faces legal action over £23m deal involving patient data. *BMJ* 2021;372:n587.