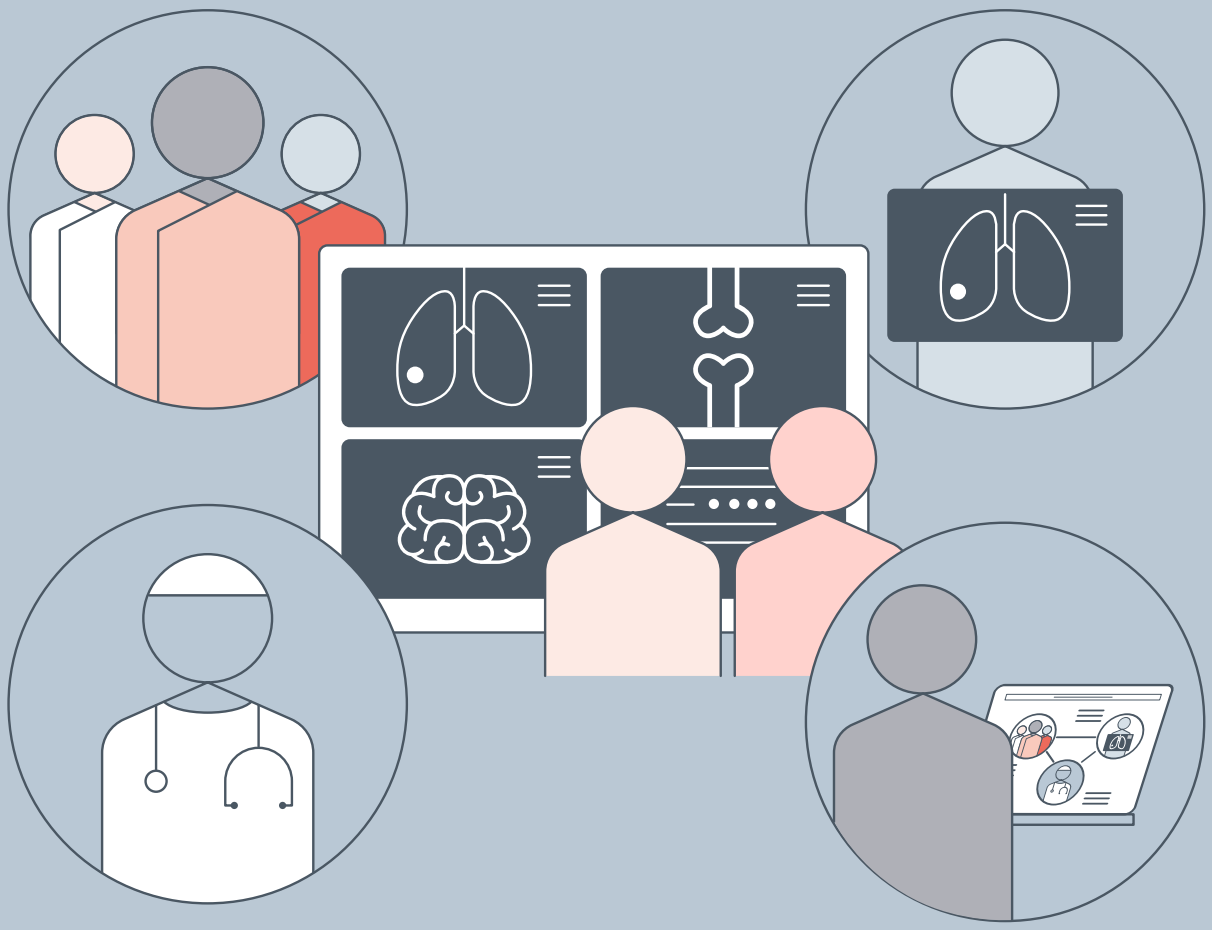
Ada Lovelace Institute

Ethics and accountability in practice

Algorithmic impact assessment: a case study in healthcare

February 2022



Contents

Executive summary

Governments, public bodies and developers of artificial intelligence (AI) systems are becoming interested in algorithmic impact assessments (referred to throughout this report as 'AIAs') as a means to create better understanding of and accountability for potential benefits and harms from AI systems. At the same time – as a rapidly growing area of AI research and application – healthcare is recognised as a domain where AI has the potential to bring significant benefits, albeit with wide-ranging implications for people and society.

This report offers the first-known detailed proposal for the use of an algorithmic impact assessment for data access in a healthcare context – the UK National Health Service (NHS)'s proposed National Medical Imaging Platform (NMIP). It includes actionable steps for the AIA process, alongside more general considerations for the use of AIAs in other public and private-sector contexts.

There are a range of algorithmic accountability mechanisms being used in the public sector, designed to hold the people and institutions that design and deploy AI systems accountable to those affected by them.[1] AIAs are an emerging mechanism, proposed as a method for building algorithmic accountability, as they have the potential to help build public trust, mitigate potential harm and maximise potential benefit of AI systems.

Carrying out an AIA involves assessing possible societal impacts of an AI system before implementation (with ongoing monitoring often advised).[2]

1 Ada Lovelace Institute, AI Now Institute, Open Government Partnership. (2021). *Algorithmic accountability for the public sector.* Open Government Partnership. Available at: https://www.opengovpartnership.org/documents/algorithmic-accountability-public-sector/

2 Ada Lovelace Institute and DataKind UK. (2020). *Examining the black box: tools for assessing AI systems.* Ada Lovelace Institute. Available at: https://www.adalovelaceinstitute.org/report/examining-the-black-box-tools-for-assessing-algorithmic-systems/

This report offers a proposal for the use of an algorithmic impact assessment for data access in a healthcare context

AIAs are not a complete solution for accountability on their own: they are best complemented by other algorithmic accountability initiatives, such as audits or transparency registers.

AIAs are currently largely untested in public-sector contexts. This project synthesises existing literature with new research to propose both a use case for AIA methods and a detailed process for a robust algorithmic impact assessment. This research has been conducted in the context of a specific example of an AIA in a healthcare setting, to explore the potential for this accountability mechanism to help data-driven innovations to fulfil their potential to support new practices in healthcare.

In the UK, the national Department for Health and Social Care and the English National Health Service (NHS) are supporting public and private-sector AI research and development, by enabling access for developers and researchers to high-quality medical imaging datasets to train and validate AI systems. However, data-driven healthcare innovations also have the potential to produce harmful outcomes and exacerbate existing health and social inequalities, by undermining patient consent to data use and public trust in AI systems. These impacts can result in serious harm to both individuals and groups who are often 'left behind' in provision of health and social care.[3]

Because of the risk and scale of harm, it is vital that developers of AI-based healthcare systems go through a process of assessing the potential impacts of their system throughout its lifecycle. This can help mitigate possible risks to patients and the public, reduce legal liabilities for healthcare providers who use their system, and build understanding of how the system can be successfully integrated and used by clinicians.

This report offers a proposal for the use of an algorithmic impact assessment for data access in a healthcare context – the proposed National Medical Imaging Platform (NMIP) from the NHS AI Lab. Uniquely, the focus of this research is a context where the public and private sector use of AIAs intersect – a public health body that has created a database of medical imaging records and, as part of the process for granting access, has requested private sector and academic researchers and developers complete an AIA.

3 Ada Lovelace Institute. (2021). *The data divide*. Available at: https://www.adalovelaceinstitute.org/wp-content/uploads/2021/03/The-data-divide_25March_final-1.pdf

This report proposes a seven-stage process for algorithmic impact assessments

Building on Ada's existing work on assessing AI systems,[4] the project evaluates the literature on AIA methods and identifies a model for their use in a particular context. Through interviews with NHS stakeholders, experts in impact assessments and potential 'users' of the NMIP, this report explores how an AIA process can be implemented in practice, addressing three questions:

1. As an emerging methodology, what does an AIA process involve, and what can it achieve?

2. What is the current state of thinking around AIAs and their potential to produce accountability, minimise harmful impacts, and serve as a tool for the more equitable design of AI systems?

3. How could AIAs be conducted in a way that is practical, effective, inclusive and trustworthy?

The report proposes a process for AIAs, which aims to ensure that algorithmic uses of public-sector data are evaluated and governed to produce benefits for society, governments, public bodies and technology developers, as well as the people represented in the data and affected by the technologies and their outcomes.

The report findings include actionable steps to help the NHS AI Lab establish this process, alongside more general considerations for the use of AIAs in other public and private-sector contexts.

The proposed process this report recommends the NHS AI Lab adopts includes seven steps (see p. 44):

1. **AIA reflexive exercise:** an impact-identification exercise is completed by the applicant team(s) and submitted to the NMIP Data Access Committee (DAC) as part of the NMIP filtering. This templated exercise prompts teams to detail the purpose, scope and intended use of the proposed system, model or research, and who will be affected. It also provokes reflexive thinking about common ethical concerns, consideration of intended and unintended consequences and possible measures to help mitigate any harms.

4 Ada Lovelace Institute. (2021). *Technical methods for regulatory inspection of algorithmic systems*. Available at: https://www.adalovelaceinstitute.org/report/technical-methods-regulatory-inspection/

2. **Application filtering:** an initial process of application filtering is completed by the NMIP DAC to determine which applicants proceed to the next stage of the AIA.

3. **AIA participatory workshop:** an interactive workshop is held, which equips participants with a means to pose questions and pass judgement on the harm and benefit scenarios identified in the previous exercise (and possibly uncovering some further impacts), broadening participation in the AIA process.

4. **AIA synthesis:** the applicant team integrates the workshop findings into the template.

5. **Data-access decision:** the NMIP DAC makes a decision about whether to grant data access. This decision is based on criteria relating to the potential risks posed by this system and whether the product team has offered satisfactory mitigations to potentially harmful outcomes.

6. **AIA publication:** the completed AIAs are published externally in a central, easily accessible location, probably the NMIP website.

7. **AIA iteration:** the AIA is revised on an ongoing basis by project teams, and at certain trigger points, such as a process of significant model redevelopment.

Alongside the AIA process detail, this report outlines seven 'operational questions' for policymakers, developers and researchers to consider before beginning to develop or implement an AIA:

1. How to navigate the immaturity of the wider assessment ecosystem?

2. What groundwork is required prior to the AIA?

3. Who can conduct the assessment?

4. How to ensure meaningful participation in defining and identifying impacts?

5. What is the artefact of the AIA and where can it be published?

The report offers a clear roadmap towards the implementation of an AIA

6. Who will act as a decisionmaker about the suitability of the AIA and the acceptability of the impacts it documents?

7. How will trials be resourced, evaluated and iterated?

In conclusion, the report offers a clear roadmap towards the implementation of an AIA. It will be of value to policymakers, public institutions and technology developers interested in algorithmic accountability mechanisms who need a high-level understanding of the process and its specific uses, alongside generalisable findings. It will also be useful for people interested in participatory methods for data governance (following on from our *Participatory data stewardship* report).[5]

In addition, for technology developers with an AI system that needs to go through an AIA process or data controllers requiring external applicants to complete an AIA as part of data-access process, the report offers a detailed understanding of the process through supporting documentation.

This documentation includes a step-by-step guide to completing the AIA for applicants to the NMIP, and a sample AIA output template, modelled on the document NMIP applicant teams would submit with a data-access application.

5 Ada Lovelace Institute. (2021). *Participatory data stewardship*. Available at: https://www.adalovelaceinstitute.org/report/participatory-data-stewardship/

How to read this report

If you are developing AI for healthcare

- This report explores in detail how a public health body might consider implementing an algorithmic impact assessment process, and explores what this process could achieve for developers, healthcare professionals and patients.

- We intend this report to deepen understanding of AIAs as a mechanism for creating more accountability, and how commercial companies or research labs might conduct an AIA to thoroughly and meaningfully assess possible benefits and harms of a proposed system in the early stages of its lifecycle.

- Start with page 44 for AIA process detail recommendations for start-ups and research labs that may be required to complete an AIA, and 'Annex 1' for supplementary resources, including the AIA template, where evidence of the AIA activity is captured.

If you are a policymaker

- This report provides a practical guide for how a public health body could apply AIAs to create more accountability. This paper and process guide may offer some generalisable considerations that could be applied to other contexts in which policymakers wish to pilot the use of AIAs.

- This paper highlights some of the practical limitations of AIAs as a methodology for algorithmic accountability, and includes a discussion of what benefits AIAs might bring to a particular context.

- See page 30 for 'The utility of AIAs in health policy: complementing existing governance processes in the UK healthcare space', and the accompanying table in the Annex, for detail on the applicability of AIAs to other domains and how AIAs complement other UK healthcare governance frameworks. Also see 'Annex 1: Proposed process in detail'

to explore our findings, challenges and uncertainties with applying AIAs in a particular context.

- Consider the relevance of this research to thinking about participatory data governance, and developing processes to ensure data subjects and people affected by technologies are considered in data-access requests to public-sector datasets.

If you are a researcher

- This report translates approaches that may be familiar in computer or social science to a specific context involving the governance of AI systems in healthcare. We intend it to support thinking about applying theoretical approaches in practice, or existing research approaches in a policy context.

- There are still remaining questions, and valuable work to be done by researchers in this space: on page 77 we outline 'Seven operational questions for AIAs' signposting areas for future study, and considerations for those interested in adopting AIAs. We also provide recommendations for how researchers can engage with policymakers and healthcare contexts.

This project explores the potential for the use of AIAs in a real-world case study: AI in medical imaging

Introduction

Rapid innovation in the use of analytics and data-driven technology (including AI) is shaping almost every aspect of our daily lives. The healthcare sector has seen significant growth in applications of data and AI, from automated diagnostics and personalised medicine to the analysis of medical imaging for screening, diagnosis and triage. The healthcare sector has seen a substantial surge in attempts to utilise AI and data-driven techniques to make existing tasks like diagnostic prediction more efficient and reimagine new ways of delivering more personalised forms of healthcare.[6]

However, while data-driven innovation holds the potential to revolutionise healthcare, it also has the potential to exacerbate health inequalities and increase demand on an already overstretched health and social care system. The risks of deploying AI and data-driven technologies in the health system include, but are not limited to:

- The perpetuation of 'algorithmic bias',[7] exacerbating health inequalities by replicating entrenched social biases and racism in existing systems.[8, 9, 10]

- Inaccessible language or lack of transparent explanations can make it hard for clinicians, patients and the public to understand the technologies and their uses, undermining public scrutiny and accountability.

6 Bohr, A. and Memarzadeh. K. (2020). 'The rise of artificial intelligence in healthcare applications'. *Artificial Intelligence in Healthcare*, pp.25-60. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7325854/

7 Angwin, J., Larson, J., Mattu, S. and Kirchnir, L. (2016). 'Machine bias'. *ProPublica*. Available at: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

8 Barocas, S. and Selbst, A. D. (2016). 'Big data's disparate impact'. *California Law Review*, 104, pp. 671- 732. [online] Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2477899

9 Buolamwini, J. and Gebru, T. (2018). 'Gender shades: intersectional accuracy disparities in commercial gender classification'. *Conference on Fairness, Accountability and Transparency*, pp.1-15.[online] Available at: https://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf

10 Miller, C. (2015). 'When algorithms discriminate'. *The New York Times*. Available at: https://www.nytimes.com/2015/07/10/upshot/when-algorithms-discriminate.html

By exploring the applicability of AIAs toward a healthcare case study of medical imaging, we hope to gain a richer understanding of how AIAs should be adopted in practice

- The collection of personal data, tracking and the normalisation of surveillance, creating risks to individual privacy.

This project explores the potential for use of one approach to algorithmic accountability, algorithmic impact assessments or 'AIAs' (see: 'What is an algorithmic impact assessment?' page 14), in a real-world case study: AI in medical imaging. AIAs are an emerging approach for holding the people and institutions that design and deploy AI systems accountable to those who are affected by them, and a way to pre-emptively identify potential impacts arising from the design, development and deployment of algorithms on people and society.

The site of research is unique among existing uses of AIAs, being located in the domain of healthcare, which is significantly regulated with a strong tradition of ethical awareness and the importance of public participation. It is also likely to produce 'high-risk' applications.

While many AIA proposals have focused on public-sector uses of AI[11,12,13] (AIAs have not yet been adopted in the private sector), and there may be a health-related AIA completed under the Canadian AIA framework, this study looks at applications at the intersection of a public and private-sector data-access process. Applications in this context are developed on data originating in the public sector, by a range of mainly private actors, but with some oversight from a public-sector department (the NHS).

This new AIA is proposed as part of a data-access process for a public-sector dataset – the National Medical Imaging Platform (NMIP). This is, to our knowledge, unique in AIAs so far. Where other proposals for AIAs have used legislation or independent assessors, this model uses a Data Access Committee (DAC) as a forum for holding developers accountable – to require the completion of the AIA, to evaluate the AIA and to prevent a project proceeding (or at least, proceeding with NHS data) if the findings are not satisfactory.

11 Reisman, D., Schultz, J., Crawford, K. and Whittaker, M. (2018). *Algorithmic impact assessments: a practical framework for public agency accountability*. AI Now Institute. Available at: https://ainowinstitute.org/aiareport2018.pdf

12 Government of Canada. (2020). Directive on Automated Decision-Making. Available at: https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592

13 Ada Lovelace Institute, AI Now Institute, Open Government Partnership.(2021). *Algorithmic accountability for the public sector*. Open Government Partnership

These properties provide a unique context, and also have implications for the design of this AIA, which should be considered by anyone looking to apply parts of this process in another domain or context. It is expected that elements of this process, such as the AIA template and exercise formats, to prove transferrable.

Some aspects, including using a DAC as the core accountability mechanism, and the centralisation of publication and resourcing for the participatory workshops, will not be directly transferable to all other cases but should form a sound structural basis for thinking about alternative solutions.

The generalisable findings to emerge from this research should be valuable to the regulators, policymakers and healthcare providers like the NHS, who will need to use a variety of tools and approaches to assess the potential and actual impacts of AI systems operating in the healthcare environment. In *Examining the Black Box*, we surveyed the state of the field in data-driven technologies and identified four notable methodologies under development, including AIAs,[14] and our study of algorithmic accountability mechanisms for the public sector identifies AIAs as forming part of the typology of other policies currently in use globally, including transparency mechanisms, audits and regulatory inspection, and independent oversight bodies.[15]

These tools and approaches are still very much in their infancy, with little consensus on how and when to apply them and what their stated aims should be, and few examples of these tools in practice. Most evidence for the usefulness of AIAs at present has come from examples of impact assessments in other sectors, rather than practical implementation. Accordingly, AIAs cannot be assumed to be ready to roll out.

By exploring the applicability of AIAs toward a healthcare case study of medical imaging – namely, the use of AIAs as part of the data release strategy of the forthcoming National Medical Imaging Platform (NMIP) from the NHS AI Lab, we hope to gain a richer understanding of how AIAs should be adopted in practice, and how such tools can be translated into

14 Ada Lovelace Institute and DataKind UK. (2020). *Examining the Black Box: tools for assessing algorithmic systems*. Available at: https://www.adalovelaceinstitute.org/report/examining-the-black-box-tools-for-assessing-algorithmic-systems/

15 Ada Lovelace Institute, AI Now Institute and Open Government Partnership. (2021). *Algorithmic accountability for the public sector*. Open Government Partnership. Available at: https://www.opengovpartnership.org/documents/algorithmic-accountability-public-sector/

meaningful algorithmic accountability and, ultimately, better outcomes for people and society.

AI in medical imaging has the potential to optimise existing processes in clinical pathways, support clinicians with decision-making and allow for better use of clinical data, but some have urged developers to adhere to regulation and governance frameworks to assure safety, quality and security and prioritise patient benefit and clinician support.[16]

Leveraging AIAs in healthcare AI has the potential to unlock better health outcomes and reduced health inequalities, as well as for building better-quality imaging products by providing developers with early insight into successful integration of their technology in complex healthcare environments, and early-stage feedback from doctors, patients, nurses and others.

16 Royal College of Radiologists. *Policy priorities: Artificial Intelligence*. Available at: https://www.rcr.ac.uk/press-and-policy/policy-priorities/artificial-intelligence

Understanding algorithmic impact assessments

What is an algorithmic impact assessment?

Algorithmic impact assessments (referred to throughout this report as 'AIAs') are a tool for assessing possible societal impacts of an AI system before the system is in use (with ongoing monitoring often advised).[17]

They have been proposed by researchers, policymakers and developers as one algorithmic accountability approach – a way to create greater accountability for the design and deployment of AI systems.[18] The intention of these approaches is to build public trust in the use of these systems, mitigate their potential to cause harm to people and groups,[19] and maximise their potential for benefit.[20]

AIAs build on the broader methodology of impact assessments, a type of policy assessment with a long history of use in other domains, such as finance, cybersecurity and environmental studies.[21] Other closely related types of impact assessments include data protection impact assessments (DPIAs), which evaluate the impact of a technology or policy on individual data privacy rights, and human rights impact assessments (HRIAs), originating in the development sector but increasingly used to assess the human rights impacts of

17 Ada Lovelace Institute and DataKindUK. (2020). *Examining the Black Box: tools for assessing algorithmic systems*. Available at: https://www.adalovelaceinstitute.org/report/examining-the-black-box-tools-for-assessing-algorithmic-systems/

18 Knowles, B. and Richards, J. (2021). 'The sanction of authority: promoting public trust in AI'. *Computers and Society*. Available at: https://arxiv.org/abs/2102.04221

19 Raji, D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D. and Barnes, P. (2020). 'Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing'. *Conference on Fairness, Accountability, and Transparency*, pp.33–44. Barcelona: ACM. Available at: https://doi.org/10.1145/3351095.3372873

20 Leslie, D. (2019). *Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector*. The Alan Turing Institute. Available at: https://www.turing.ac.uk/sites/default/files/2019-06/understanding_artificial_intelligence_ethics_and_safety.pdf

21 Reisman, D., Schultz, J., Crawford, K. and Whittaker, M. (2018). *Algorithmic impact assessments: a practical framework for public agency accountability*. AI Now Institute. Available at: https://ainowinstitute.org/aiareport2018.pdf

AIAs encourage developers of AI systems to consider the potential impacts of the development and implementation of their system

business practices and technologies.[22]

Conducting an impact assessment provides actors with a way to assess and evaluate the potential economic, social and environmental impacts of a proposed policy or intervention.[23] Some impact assessments are conducted prior to launching a policy or project as a way to foresee potential risks, known as *ex ante* assessments, while others are launched once the policy or project is already in place, to evaluate how the project went – known as *ex post*.

Unlike other impact assessments, AIAs specifically encourage developers of AI systems to consider the potential impacts of the development and implementation of their system. Will this system affect certain individuals disproportionately more than others? What kinds of socio-environmental factors – such as stable internet connectivity or a reliance on existing hospital infrastructure – will determine its success or failure? AIAs provide an ex ante assessment of these kinds of impacts and potential mitigations at the earliest stages of an AI system's development.

Current AIA practice in the public and private sectors

AIAs are currently not widely used in either public or private sector contexts and there is no single accepted standard, or 'one size fits all', methodology for their use.

AIAs were first proposed by the AI Now Institute as a detailed framework for underpinning accountability in public sector agencies that engages communities impacted by the use of public sector algorithmic decision-making,[24] building from earlier scholarship that proposed the use of 'algorithmic impact statements' as a way to

22 Recent examples include Facebook's *ex post* HRIA of their platform's effects on the genocide in Myanmar, and Microsoft's HRIA of its use of AI. See: Latonero, M. and Agarwal, A. (2021). *Human rights impact assessments for AI: learning from Facebook's failure in Myanmar*. CARR Center for Human Rights Policy Harvard Kennedy School. Available at: https://carrcenter.hks.harvard.edu/files/cchr/files/210318-facebook-failure-in-myanmar.pdf; Article One. Challenge: From 2017 to 2018, Microsoft partnered with Article One to conduct the first-ever Human Rights Impact Assessment (HRIA) of the human rights risks and opportunities related to artificial intelligence (AI). Available at: https://www.articleoneadvisors.com/case-studies-microsoft

23 Adelle, C. and Weiland, S. (2012). 'Policy assessment: the state of the art'. *Impact Assessment and Project Appraisal* 30.1, pp. 25-33 Available at: https://www.tandfonline.com/doi/full/10.1080/14615517.2012.663256

24 Reisman, D., Schultz, J., Crawford, K. and Whittaker, M. (2018). *Algorithmic impact assessments: a practical framework for public agency accountability*. AI Now Institute. Available at: https://ainowinstitute.org/aiareport2018.pdf

manage predictive policing technologies.[25]

Though consensus is growing over the importance of principles for the development and use of AI systems like accountability, transparency and fairness, individual priorities and organisational interpretation of these terms differ. The lack of consistency with these concepts means not all AIAs are designed to achieve the same ends, and the process for conducting AIAs will depend on the specific context in which they are implemented.[26]

Recent scholarship from Data & Society identifies 10 'constitutive components' as common to different types of impact assessment, and that are necessary for inclusion in any AIA. These include a 'source of legitimacy', the idea that an impact assessment must be legally mandated and enforced through another institutional structure such as a government agency, and a relational dynamic between stakeholders, the accountable actor and an accountability forum that describe how accountability relationships are formed.

In an 'actor – forum' relationship, an actor should be able to explain and justify conduct to an external forum, who are able to pass judgement.[27] Other components include 'public consultation', involving gathering feedback from external perspectives for evaluative purposes, and 'public access', which gives members of the public access to crucial material about the AIA, such as its procedural elements, in order to further build accountability.[28]

While varied approaches to AIAs have been proposed in theory, only one current model of AIA exists in practice, authorised by the Treasury Board of Canada Secretariat's Directive on Automated Decision-Making,[29]

25 Selbst, A.D. (2017). 'Disparate impact in big data policing'. 52 *Georgia Law Review* 109, pp.109-195. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2819182

26 Metcalf, J., Moss, E., Watkins, E.A., Ranjit, S. and Elish, M.C. (2021). 'Algorithmic impact assessments and accountability: the co-construction of impacts'. *Conference on Fairness Accountability, and Transparency* [online] Available at: https://dl.acm.org/doi/pdf/10.1145/3442188.3445935

27 Wieringa, M. (2020). 'What to account for when accounting for algorithms: a systematic literature review on algorithmic accountability'. *Conference on Fairness, Accountability, and Transparency*, pp.1-18 [online] Barcelona: ACM. Available at: https://dl.acm.org/doi/10.1145/3351095.3372833

28 Moss, E., Watkins, E.A., Singh, R., Elish, M.C. and Metcalf, J. (2021). *Assembling accountability: algorithmic impact assessment for the public interest*. Data & Society. Available at: https://datasociety.net/library/assembling-accountability-algorithmic-impact-assessment-for-the-public-interest/

29 Government of Canada. (2020). *Directive on Automated Decision-Making*. Available at: https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592

aimed at Canadian civil servants and used to manage public-sector AI delivery and procurement standards. The lack of more practical examples of AIAs is a known deficiency in the literature.

The lack of real-world examples and practical difficulty for institutions implementing AIAs remains a concern for those advocating for their widespread adoption, particularly as part of policy interventions.

An additional consideration is the inclusion of a diverse range of perspectives in the process of its development. Most AIA processes are controlled and determined by decision-makers in the algorithmic process, with less emphasis on the consultation of outside perspectives, including the experiences of those most impacted by the algorithmic deployment. As a result, AIAs are at risk of adopting an incomplete or incoherent view of potential impacts, divorced from these lived experiences.[30] To practically seek and integrate those perspectives into the final AIA output has proven to be a difficult and ill-defined undertaking, with the required guidance being largely unavailable.

Canadian algorithmic impact assessment model

At the time of writing, the Canadian AIA is the only known and recorded AIA process implemented in practice. The Canadian AIA is a procurement management tool adopted under the Directive on Automated Decision-Making, aiming to guide policymakers into best practice use and procurement of AI systems that might be used to help govern service delivery at the federal level.

The Directive draws from administrative law principles of procedural fairness, accountability, impartiality and rationality,[31] and is aimed at all AI systems that are used to make a decision about an individual.[32] One of the architects of the AIA, Noel Corriveau, considers a merit of impact assessments is to facilitate

30 Katell, M., Young, M., Dailey, D., Herman, B., Guetler, V., Tam, A., Bintz, C., Raz, D. and Krafft, P. M. (2020). 'Toward situated interventions for algorithmic equity: lessons from the field'. *Conference on Fairness, Accountability, and Transparency* pp.44-45 [online] ACM: Barcelona. Available at: https://dl.acm.org/doi/abs/10.1145/3351095.3372874

31 Scassa, T. (2020). *Administrative law and the governance of automated decision-making: a critical look at Canada's Directive on Automated Decision-Making*. Forthcoming, University of British Columbia Law Review. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3722192

32 Government of Canada. (2020). *Directive on Automated Decision-Making*. Available at: https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592

compliance with legal and regulatory requirements.[33]

The AIA itself consists of an online questionnaire of eight sections containing 60 questions related to technical attributes of the AI system, the data underpinning it and how the system designates decision-making, and frames 'impacts' as the 'broad range of factors' that may arise because of a decision made by, or supported by, an AI system. Four categories of 'impacts' are utilised in this AIA: **the rights of individuals**, **health and wellbeing of individuals**, **economic interests of individuals** and **impacts on the ongoing sustainability of an environmental ecosystem**.

Identified impacts are ranked according to a sliding scale, from little to no impact to very high impact, and weighted to produce a final impact score. Once complete, the AIA is exported to PDF format and published on the Open Canada website. At the time of writing, there are four completed Canadian AIAs, providing useful starting evidence for how AIAs might be documented and published.

Many scholars and practitioners consider AIAs to hold great promise in assessing the possible impacts of the use of AI systems within the public sector, including applications that range from law enforcement to welfare delivery.[34] For instance, the AI Now Institute's proposed AIA sets out a process intended to build public agency accountability and public trust.[35] As we explored in *Algorithmic accountability for the public sector*, AIAs can be considered part of a wider toolkit of algorithmic accountability policies and approaches adopted globally, including algorithm auditing,[36] and algorithm transparency registers.[37]

33 Karlin, M. and Corriveau, N. (2018). 'The Government of Canada's Algorithmic Impact Assessment: Take Two'. *Supergovernance*. Available at: https://medium.com/@supergovernance/the-government-of-canadas-algorithmic-impact-assessment-take-two-8a22a87acf6f

34 Margetts, H. and Dorobantu, C. (2019). 'Rethink government with AI'. *Nature*. Available at: https://www.nature.com/articles/d41586-019-01099-5

35 Reisman, D., Schultz, J., Crawford, K. and Whittaker, M. (2018). *Algorithmic impact assessments: a practical framework for public agency accountability*. AI Now Institute. Available at: https://ainowinstitute.org/aiareport2018.pdf

36 Raji, D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D. and Barnes, P. (2020). 'Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing'. *Conference on Fairness, Accountability, and Transparency*, pp.33–44. Barcelona: ACM. Available at: https://doi.org/10.1145/3351095.3372873

37 Transparency registers document and make public the contexts where algorithms and AI systems are in use in local or federal Government, and have been adopted in cities including Helsinki, see: City of Helsinki AI register. *What is the AI register?* Available at: https://ai.hel.fi/ and Amsterdam, see: Amsterdam Algorithm Register Beta. *What is the algorithm register?* Available at: https://algoritmeregister.amsterdam.nl/

Other initiatives have been devised as 'soft' self-assessment frameworks, to be used alongside an organisation or institution's existing ethics and norms guidelines, or in deference to global standards like the IEEE's AI Standards or the UN Guiding Principles on Business and Human Rights. These kinds of initiatives often relay some flexibility on recommendations to suit specific use cases, as seen in the European Commission's High-level Expert Group on AI's assessment list for trustworthy AI.[38]

While many proponents of AIAs from civil society and academia see them as a method for improving *public* accountability,[39] AIAs also have scope for adoption within private-sector institutions, under the condition of regulators and public institutions incentivising their adoption and compelling their use in certain private sector contexts. Conversely, AIAs also help provide a lens for regulators to view, understand and pass judgement on institutional cultures and practices.[40] The proposed US Algorithm Accountability Act sets out requirements for large private companies to undertake impact assessments in 2019,[41] with progress on the Act beginning to regain momentum.[42]

The focus of this case study is on a context where the public and private sector use of AIAs intersect – a public health body has created a database of medical imaging records and, as part of the process for granting access, has requested private-sector and academic researchers and developers complete an AIA. This is a novel context that presents its own unique challenges and learnings (see: 'Annex 1: Proposed process in detail' p. 92), but has also yielded important considerations that we believe are pertinent and timely for other actors interested in AIAs (see: 'Seven operational questions for AIAs', p. 77)

38 European Commission. (2020). *Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment*. Available at: https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment

39 Binns, R. (2018). 'Algorithmic accountability and public reason'. *Philosophy & Technology*, 31, pp.543-556. [online] Available at: https://link.springer.com/article/10.1007/s13347-017-0263-5

40 Selbst, A.D. (2021). 'An institutional view of algorithmic impact assessments', *Harvard Journal of Law & Technology* (forthcoming). Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3867634

41 Congress.Gov. (2019). *H.R.2231 - Algorithmic Accountability Act of 2019*. Available at: https://www.congress.gov/bill/116th-congress/house-bill/2231#:~:text=Introduced%20in%20House%20(04%2F10%2F2019)&text=This%20bill%20requires%20specified%20commercial,artificial%20intelligence%20or%20machine%20learning

42 Johnson, K. (2021). 'The movement to hold AI accountable gains more steam'. *Ars Technica*. Available at: https://arstechnica.com/tech-policy/2021/12/the-movement-to-hold-ai-accountable-gains-more-steam/3/

Goals of the NHS AI Lab NMIP AIA process

This report aims to outline a practical design of the AIA process for the NHS AI Lab's NMIP project. To do this, we reviewed the literature to uncover both areas of consensus and uncertainty among AIA scholars and practitioners, in order to build on and extend existing research. We also interviewed key NHS AI Lab and NMIP stakeholders, employees at research labs and healthtech start-ups who would seek access to the NMIP and experts in algorithmic accountability issues in order to guide the development of our process (see: 'Methodology' p. 90).

As discussed above, AIAs are context-specific and differ in their objectives and assumptions, and their construction and implementation. It is therefore vital that the NMIP AIA has clearly defined and explained goals in order to both communicate the purpose of an AIA for the NMIP context, and ensure the process works, enabling a thorough, critical and meaningful *ex ante* assessment of impacts.

This information is important for developers who undertake the AIA process to understand the assumptions behind its method, as well as policymakers interested in algorithmic accountability mechanisms, in order to usefully communicate the value of this AIA and distinguish it from other proposals.

In this context, this AIA process is designed to achieve the following goals:

1. accountability
2. reflection/reflexivity
3. standardisation
4. independent scrutiny
5. transparency.

These goals emerged both from literature review and interviews, enabling us to identify areas where the AIA would add value, complement existing governance initiatives and contribute to minimising harmful impacts.

1. Accountability

It's important to have a clear understanding of what accountability means in the context of the AIA process. The definition that is most helpful here understands accountability as a depiction of the social relationship between an 'actor' and a 'forum', where being accountable describes an obligation of the actor to explain and justify conduct to a forum.[43] An actor in this context might be a key decision-maker within an applicant team, such as a technology developer and project principal investigator. The forum might comprise the arrangement of external stakeholders, such as clinicians who might use the system, members of the Data Access Committee (DAC) and members of the public. The forum must have the capacity to deliberate on the actor's actions, ask questions, pass judgement and enforce sanctions if necessary.[44]

To create a more accountable relationship between developers and individuals affected by their systems, the AIA process equips a forum of clinicians and patients to request the information they need. If successful, this will allow them to pose questions about an AI system and be given the agency to deliberate on social impacts of AI systems, providing alternate expertise and insight. The result of their deliberations is then shared with the DAC, who have the power to ask further questions and pass judgement, as well as enforcing sanctions by denying a request for access from an applicant. Finally, members of the public are another actor in this accountability chain. Once the results of the AIA are published externally (see 'Transparency' below), the public has the ability to scrutinise and evaluate the impacts documented in the AIA.

2. Reflection/reflexivity

An AIA process should prompt reflection from developers and critical dialogue with individuals who would be affected by this process about how the design and development of a system might result in certain harms and benefits – to clinicians, patients, and society. Behaving

43 Bovens, M. (2006). Analysing and assessing public accountability. A conceptual framework. *European Governance Papers* (EUROGOV) No. C-06-01. Available at: https://www.ihs.ac.at/publications/lib/ep7.pdf

44 Metcalf, J., Moss, E., Watkins, E.A., Ranjit, S. and Elish, M.C. (2021). 'Algorithmic impact assessments and accountability: the co-construction of impacts'. *Conference on Fairness Accountability, and Transparency* [online] Available at: https://dl.acm.org/doi/pdf/10.1145/3442188.3445935

reflexively means examining or responding to one's – or that of a teams' – own practices, motives and beliefs during a research process.

Reflexivity is an essential principle for completing a thorough, meaningful and critical AIA, closely related to the concept of positionality, which has been developed through work on AI ethics and safety in the public sector.[45] Our reflexive exercise enables this practice among developers by providing an actionable framework for discussing ethical considerations arising from the deployment of AI systems, and a forum for exploration of individual biases and ways of viewing and understanding the world.

The broad participation of a range of perspectives is therefore a critical element of increased awareness in a reflection that includes some level of awareness to positionality. The AIA exercises were built with continual reflexivity in mind, which provide a means for technology developers to examine ethical principles thoroughly during design and development phases.

3. Standardisation

Our literature review revealed that while many scholars have proposed possible approaches and methods for an AIA, these tend to be higher-level recommendations for an overall approach. There is little discussion around how individual activities of the AIA should be structured, captured and recorded. A notable exception is the Canadian AIA, which makes use of a questionnaire to capture the impact assessment process, providing a format for the AIA 'users' to follow in order to complete the AIA, and for external stakeholders to view once the AIA is published.

Some existing data/AI governance processes were confusing for product and development teams. One stakeholder interviewee commented:

'Not something I'm an expert in – lots of the forms written in language I don't understand, so was grateful that our information governance chaps took over and made sure I answered the right things within that.'
Expert stakeholder

45 Leslie, D. (2019). *Understanding artificial intelligence ethics and safety*. Available at: https://www.turing.ac.uk/sites/default/files/2019-08/understanding_artificial_intelligence_ethics_and_safety.pdf

This underscored the need for a clear and coherent, standardised AIA process to ensure that applicant teams were able to engage fully with the task and that completed AIAs are of a consistent standard.

To ensure NMIP applicants find the AIA as effective and practical as possible, and to build consistency between applications, it is important they undergo a clearly defined process that leads to an output that can be easily compared and evaluated. To this end, our AIA process provides a standard template document, both to aid the process and keep relative uniformity between different NMIP applications.

Over time, once this AIA has been trialled and tested, we envisage that standardised and consistent applications will also help the DAC and members of the public to begin to develop paradigms of the kinds of harms and benefits that new applicants should consider.

4. Independent scrutiny

The goal of independent scrutiny is to provide external stakeholders with the powers to scrutinise, assess and evaluate AIAs and identify any potential issues with process. Many proposed AIAs argue for multistakeholder collaboration,[46] but there is a notable gap in procedure for how participation would be structured in an AIA, and how external perspectives would be included in the process.

We sought to address these gaps by building a participatory initiative as part of the NMIP AIA (for more information on the participatory workshop, see: 'Annex 1: Proposed process in detail' p. 92). Independent scrutiny helps to build robust accountability, as it helps to formalise the actor-forum relationship, providing further opportunity for judgement and deliberation among the wider forum.[47] AIAs should be routinely scrutinised to ensure they are used and adopted effectively, that teams are confident and critical in their approach to examining impacts, and that AIAs provide continual value.

46 Reisman, D., Schultz, J., Crawford, K. and Whittaker, M. (2018). *Algorithmic impact assessments: a practical framework for public agency accountability*. AI Now Institute. Available at: https://ainowinstitute.org/aiareport2018.pdf

47 Wieringa, M. (2020). 'What to account for when accounting for algorithms: a systematic literature review on algorithmic accountability'. *Conference on Fairness, Accountability and Transparency*, p.1-18. ACM: Barcelona. Available at: https://dl.acm.org/doi/abs/10.1145/3351095.3372833

5. Transparency

In this context, we consider AIA transparency as building in critical oversight of the AIA process itself, focusing on making the AIA, as a mechanism of governance, transparent. This differs to making transparent details about the AI system and its logic – what has been referred to as 'first-order transparency'.[48] This AIA aims to improve transparency via both internal and external visibility, by prompting applicant teams to document the AIA process and findings, which are then published centrally for members of the public to view. Making this information publicly available provides more information for regulators, civil society organisations and members of the public about what kinds of systems are being developed in the UK healthcare context, and how their societal impacts are understood by those who develop or research them.

In order to achieve these goals, the AIA process and output make use of two principal approaches: **documentation** and **participation**.

1. Documentation

Thorough recordkeeping is critical to this AIA process and can produce significant benefits for developers and external stakeholders.

Teams who have access to documentation stating ethical direction are more likely to address ethical concerns with a project at the outset.[49] Documentation can change internal process and practice, as it necessitates reflexivity, which creates opportunities to better identify, understand and question assumptions and behaviours.

This shift in internal process may also begin to influence *external* practice: it has been argued that good AIA documentation process may create what sociologists call 'institutional isomorphism', where industry practice begins to homogenise owing to social and normative pressures.[50]

48 Kaminski, M. (2020). 'Understanding transparency in algorithmic accountability'. *Cambridge Handbook of the Law of Algorithms*, e.d. Woodrow Barfield. Cambridge: Cambridge University Press [online] Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3622657

49 Boyd, K.L (2021). 'Datasheets for datasets help ML engineers notice and understand ethical issues in training data'. *Proceedings of theACM on Human-Computer Interaction*, 5, 438, pp.1-27. [online] Available at: https://dl.acm.org/doi/abs/10.1145/3479582

50 Selbst, A. (2021). 'An institutional view of algorithmic impact assessments'. 35 *Harvard Journal of Law & Technology* (forthcoming), pp.1-79. [online] Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3867634

Through consistent documentation, teams gain a richer context for present and future analysis and evaluation of the project.

2. Participation

Participation is the mechanism for bringing a wider range of perspectives to the AIA process. It can take various forms – from soliciting written feedback through to deliberative workshops – but should always aim to bring the lived experiences of people and communities who are affected by an algorithm to bear on the AIA process.[51]

When carried out effectively, participation supports teams in building higher quality, safer and fairer products.[52] The participatory workshop in the NMIP AIA (see: 'Annex 1: Proposed process in detail' p. 92 for a full description) enables the process of impact identification to go beyond the narrow scope of the applicant team(s).

Building participation into the AIA process brings external scrutiny of an AI healthcare system from outside the teams' perspective, provides alternate sources of knowledge and relevant lived experience and expertise. It also enables independent review of the impacts of an AI system, as participants are unencumbered by the typical conflicts of interest that may interfere with the ability of project stakeholders to judge their system impartially.

51 See: Ada Lovelace Institute. (2021). *Participatory data stewardship*. Available at: https://www.adalovelaceinstitute.org/report/participatory-data-stewardship/ for a framework of different approaches to participation in relation to data-driven technologies and systems.

52 Madaio, M.A. et al (2020) 'Co-designing checklists to understand organizational challenges and opportunities around fairness in AI' *CHI Conference on Human Factors in Computing Systems*, pp.1-14 [online]. Available at: https://doi.org/10.1145/3313831.3376445

The context of healthcare AI

There is a surge in the development and trialling of AI systems in healthcare.[53] A significant area of growth is the use of AI in medical imaging, where AI imaging systems assist clinicians in cancer screening, supporting diagnosis/prognosis, patient triage and patient monitoring.[54]

The UK Department of Health and Social Care (DHSC) has set out national commitments to support public and private sector AI research and development in healthcare by ensuring that developers and researchers have access to high-quality datasets to train and validate AI models, underlining four guiding principles that steer this effort:

1. user need
2. privacy and security
3. interoperability and openness
4. inclusion.[55]

In the current NHS Long Term Plan, published in 2019, AI is described as a means to improve efficiency across service delivery by supporting clinical decisions, as well as a way to 'maximise the opportunities for use of technology in the health service'.[56] Current initiatives to support this drive for testing, evaluation and scale of AI-driven technologies include the AI in Health and Care Award, run by the Accelerated Access Collaborative, in partnership with NHSX (now part of the NHS Transformation Directorate)[57]

53 Davenport, T. and Kalakota, R. (2019). 'The potential for artificial intelligence in healthcare'. *Future Healthcare Journal*, 6,2, pp.94-98. [online] Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6616181/

54 NHS AI Lab. *AI in imaging*. Available at: https://www.nhsx.nhs.uk/ai-lab/ai-lab-programmes/ai-in-imaging/

55 Department of Health and Social Care. (2018). *The future of healthcare: our vision for digital, data and technology in health and care*. UK Government. Available at: https://www.gov.uk/government/publications/the-future-of-healthcare-our-vision-for-digital-data-and-technology-in-health-and-care/the-future-of-healthcare-our-vision-for-digital-data-and-technology-in-health-and-care

56 NHS. (2019). *The NHS Long Term Plan*. Available at: https://www.longtermplan.nhs.uk/wp-content/uploads/2019/08/nhs-long-term-plan-version-1.2.pdf

57 NHSX is now part of the NHS Transformation Directorate. More information is available at: https://www.nhsx.nhs.uk/blogs/nhsx-moves-on/ At the time of research and writing NHSX was a joint unit of NHS England and the UK Department of Health and Social Care that reported directly to the Secretary of State and the Chief Executive of NHS England and NHS Improvement. NHSX was also the parent organisation of the NHS AI Lab.

and the National Institute for Health Research (NIHR).

However, while data-driven healthcare innovation holds the potential to support new practices in healthcare, careful research into the integration of AI systems in clinical practice is needed to ground claims of model performance and to uncover where systems would be most beneficial in the context of particular clinical pathways. For example, a recent systematic review of studies measuring test accuracy of AI in mammography screening practice has revealed that radiologists still outperform the AI in detection of breast cancer.[58]

To ensure healthcare AI achieves the benefits society hopes for, it is necessary to recognise the possible risks of harmful impacts from these systems. For instance, concerns have been raised that AI risks further embedding or exacerbating existing health and social inequalities – a risk that is evidenced in both systems that are working as designed,[59] and in those that are producing errors or are failing.[60, 61]

Additionally, there are concerns around the kinds of interactions that take place between clinicians and AI systems in clinical settings: the AI system may contribute to human error, override much-needed human judgement, or lead to overreliance or misplaced faith in the accuracy metrics of the system.[62]

The NHS has a longstanding commitment to privacy and processing personal data in accordance with the General Data Protection Regulation (GDPR)[63] which may create tension with the more recent

58 Freeman, K., Geppert, J., Stinton, C., Todkill, D., Johnson, S., Clarke, A. and Taylor-Phillips, S. (2021). 'Use of artificial intelligence for image analysis in breast cancer screening programmes: systematic review of test accuracy'. *British Medical Journal* 2021, 374 [online] Available at: https://pubmed.ncbi.nlm.nih.gov/34470740/

59 Wen, D., Khan, S., Ji Xu, A., Ibrahim, H., Smith, L., Caballero, J., Zepeda, L., de Blas Perez, C., Denniston, A., Lui, X. and Martin, R. (2021). 'Characteristics of publicly available skin cancer image datasets: a systematic review'. *The Lancet: Digital Health* [online]. Available at: https://www.thelancet.com/journals/landig/article/PIIS2589-7500(21)00252-1/fulltext

60 Banerje, I et al. (2021). 'Reading race: AI recognises patient's racial identity in medical images'. *Computer Vision and Pattern Recognition*. Available at: https://arxiv.org/abs/2107.10356

61 Antun, V., Renna, F., Poon, C., Adcock, B., Hansen, A. C. (2020). 'On instabilities of deep learning in image reconstruction and the potential costs of AI'. *Proceedings of the National Academy of Sciences of the United States of America*, p. 117, 48 [online] Available at: https://www.pnas.org/content/117/48/30088

62 Topol, E. (2019). 'High performance medicine: the convergence of human and artificial intelligence'. *Nature Medicine*, 25, pp.45-56. [online] Available at: https://www.nature.com/articles/s41591-018-0300-7

63 NHSX. *How NHS and care data is protected*. Available at: https://www.nhsx.nhs.uk/key-tools-and-info/data-saves-lives/how-nhs-and-care-data-is-protected

commitment to make patient data available for companies.[64] Potential harmful impacts arising from use of these systems are myriad, from both healthcare-specific concerns around violating patient consent over the use of their data, to more generic risks such as creating public mistrust of AI systems and the institutions that develop or deploy them.

It is important to understand impacts do not have parity across people and groups: for example, a person belonging to a marginalised group may experience even greater mistrust around use of AI, owing to past discrimination.

These impacts can result in serious harm to both individuals and groups, who are often 'left behind' in provision of health and social care.[65] Harmful impacts can arise from endemic forms of bias during AI design and development, from error or malpractice at the point of data collection, to over-acceptance of model output, and reducing vigilance at the point of end use.[66] Human values and subjectivities such as biased or racist attitudes or behaviours can become baked-in to AI systems,[67] and reinforce systems of oppression once in use, resulting in serious harm.[68] For example, in the USA, an algorithm commonly used in hospitals to determine which patients required follow-up care was found to classify White patients as more ill than Black patients even when their level of illness was the same, affecting millions of patients for years before it was detected.[69]

Because of the risk and scale of harm, it is vital that developers of AI-based healthcare systems go through a process of assessing potential impacts of their system throughout its lifecycle. Doing so can help developers mitigate possible risks to patients and the public, reduce legal liabilities for healthcare providers who use their system, and consider how their system can be successfully integrated and used by clinicians.

64 NHS Digital. *How NHS Digital makes decisions about data access*. Available at: https://digital.nhs.uk/services/data-access-request-service-dars/how-nhs-digital-makes-decisions-about-data-access

65 Ada Lovelace Institute. (2021). *The data divide*. Available at: https://www.adalovelaceinstitute.org/report/the-data-divide/

66 Data Smart Schools. (2021). *Deb Raji on what 'algorithmic bias' is (...and what it is not)*. Available at: https://data-smart-schools.net/2021/04/02/deb-raji-on-what-algorithmic-bias-is-and-what-it-is-not/

67 Balayn, A and Gürses, S. (2021). *Beyond debiasing: regulating AI and its inequalities*. European Digital Rights. Available at: https://edri.org/wp-content/uploads/2021/09/EDRi_Beyond-Debiasing-Report_Online.pdf

68 Noble, S.U. (2018). *Algorithms of oppression: how search engines reinforce racism*. NYU Press

69 Chakradhar, S. (2019). 'Widely used algorithm in hospitals is biased, study finds'. *STAT*. Available at: https://www.statnews.com/2019/10/24/widely-used-algorithm-hospitals-racial-bias/

Impacts arising from development and deployment of healthcare AI systems

AI systems are valued by their proponents for their potential to support clinical decisions, monitoring of patient health, freeing resources and improving patient outcomes. These impacts, if realised, would hopefully result in beneficial, tangible outcomes, but there may also be consequences arising from when the AI system is used as intended or when it is producing errors or failing.

Many of these technologies are in their infancy, and often only recently adopted into clinical settings, so there is a real risk of these technologies producing adverse effects, causing harm to people and society in the near and long term. Given the scale that these systems operate at and the high risk of significant harm if they do fail in a healthcare setting, it is essential for developers to consider the impacts of their system before they are put in use.

Recent evidence provides examples of some kinds of impacts (intended or otherwise) that have emerged from the development and deployment of healthcare AI systems:

- A study released in July 2021 found that algorithms used in healthcare are able to read a patient's race from medical images including chest and hand X-rays and mammograms.[70] Race is not an attribute normally detectable from scans. Other evidence shows that Black patients and patients from other marginalised groups may receive inferior care than White patients.[71] Being able to identify race from a scan (with any level of certainty) raises the risk of introducing an unintended system impact that causes harm to both individuals and society, reinforcing systemic health inequalities.

- A 2020 study of the development, implementation and evaluation of Sepsis Watch, an AI 'early-warning system' for assisting hospital clinicians in the early diagnosis and treatment of sepsis uncovered unintended consequences.[72] Sepsis Watch was successfully integrated with clinical practice after close engagement with nurses and hospital staff to ensure it triggered an alarm in an appropriate way and led to a meaningful response. But the adoption of the system had an unanticipated impact of clinicians taking on an intermediary role between the AI system and other clinicians in order to successfully integrate the tool for hospital use. This demonstrates that developers should

70 Gichoya, J.W. et al. (2021). 'Reading race: AI recognises patient's racial identity in medical images'. *arXiv*. Available at: https://arxiv.org/abs/2107.10356

71 Frakt, A. (2020). 'Bad medicine: the harm that comes from racism'. *The New York Times*. [online] Available at: https://www.nytimes.com/2020/01/13/upshot/bad-medicine-the-harm-that-comes-from-racism.html

72 Sendak, M. et al. (2020). '"The human body is a black box": supporting clinical decision-making with deep learning'. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM:, New York, NY, USA, pp. 99–109. Available at: https://doi.org/10.1145/3351095.3372827

take into account the socio-environmental requirements to successfully implement and run an AI system.

- A study released in December 2021 revealed underdiagnosis bias in AI-based chest X-ray (CXR) prediction models among marginalised populations, particularly in intersectional subgroups.[73] This example shows that analysis of how an AI system performs on certain societal groups may be missed, so careful consideration of user populations *ex ante* is critical to help mitigate harms *ex post*. It also demonstrates how some AI systems may result in a reduced quality of care that may result in injury to some patients.

- A study on the implementation of an AI-based retinal scanning tool in Thailand for detecting diabetic eye disease found that its success depended on socio-environmental factors like whether the hospital had a stable internet connection and lighting conditions for taking photographs – when these were insufficient, the use of the AI system caused delays and disruption.[74] They found that clinicians unexpectedly created 'work-arounds' for the intended study design use of the AI system. This reflected unanticipated needs that affected how the process worked, in particular that patients may struggle to attend distant hospitals for further examination, which made hospital referral a bad fallback for when the AI system failed. This concern was identified through researchers' discussions with clinicians, showing the potential value of participation early in the design and development process.

The utility of AIAs in health policy: complementing existing governance processes in the UK healthcare space

The AIA process is intended to complement and build from existing regulatory requirements imposed on proposed medical AI products, recognising the sanctity of well-established regulation. As a result, it is essential to survey that regulatory context before diving into the specifics of what an AIA requires, and where an AIA can add value.

73 Seyyad-Kalantari, L., Zhang, H., McDermott, M., Chen, I. Y., Ghassemi, M. (2021). 'Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in underserved patient populations'. *Nature Medicine*, 27, pp. 2176-2182. Available at: https://www.nature.com/articles/s41591-021-01595-0

74 Beede, E., Elliott Baylor, E., Hersch, F., Iurchenko, A., Wilcox, L., Ruamviboonsuk, P. and Vardoulakis, L. (2020). 'A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy'. In: *CHI Conference on Human Factors in Computing Systems (CHI '20),* April 25-30, 2020, Honolulu, HI, USA. ACM, New York, NY, USA. Available at: https://dl.acm.org/doi/fullHtml/10.1145/3313831.3376718

Compared to most other domains, the UK's healthcare sector already has in place relatively mature regulatory frameworks for the development and deployment of AI systems with a medical purpose. The UK Government has indicated that further updates to regulation are forthcoming, in order to be more responsive to data-driven technologies like AI.[75] There is in a complex ecosystem of regulatory compliance, with several frameworks for risk assessment, technical, scientific and clinical assurance and data protection that those adopting or building these systems must navigate.

This AIA process is therefore proposed as one component in a broader accountability toolkit, which is intended to provide a standardised, reflexive framework for assessing impacts of AI systems on people and society. It was designed to complement – not replicate or override – existing governance processes in the UK healthcare space. Table 1 below compares the purpose, properties and evidence required by some of these processes, to map how this AIA adds value.

75 Medicines and Healthcare products Regulatory Agency (2020). *Regulating medical devices in the UK*. UK Government. Available at: https://www.gov.uk/guidance/regulating-medical-devices-in-the-uk

Table 1: How does this AIA complement some existing processes in the healthcare space?

Name of initiative	Medical devices regulation	NHS code of conduct for digital and data-driven health technologies (DHTs)	NICE evidence standards frameworks for DHTs	Data protection impact assessments (DPIAs)	ISO clinical standards: 14155 & 14971
Initiative details	**Legislation** Follows the EU risk-based classification of medical devices implemented and enforced by a competent authority: in the UK, this is the Medicines & Healthcare products Regulatory Agency (MHRA). MHRA's medical device product registration, known as a CE marking process, is a requirement under the UK's Medical Device Regulations 2002. Higher-risk products will have conformity assessments carried out by third-parties: notified bodies.[76]	**Non-mandatory, voluntary best-practice standards** The NHS outlines 12 key principles of good practice for innovators designing and developing data-driven healthcare products, including 'how to operate ethically', 'usability and accessibility', and technical assurance. There is considerable emphasis on 'good data protection practice, including data transparency'.	**Non-mandatory, voluntary best-practice standards** Outlines a set of standards for innovation, grouping DHTs into tiers based on functionality for a proportionate, streamlined framework. The framework's scope covers DHTs that incorporate AI using fixed algorithms (but not DHTs using adaptive algorithms).	**Mandatory impact assessment (with a legal basis under the GDPR)** Completed as a guardrail against improper data handling and to protect individual data rights (DPIAs are not specific to healthcare).	**Non-mandatory clinical standards for medical devices (including devices with an AI component)** From the International Standards Organisation, and considered gold standard, is internationally recognised, and can be used as a benchmark for regulatory compliance.
Which part of project lifecycle?	Whole lifecycle, particularly development, and including post-deployment.	Development and procurement.	Development and procurement.	Ideation to development.	Whole lifecycle.

76 Medicines and Healthcare products Regulatory Agency (MHRA). (2020). *Medical devices: conformity assessment and the UKCA mark. UK Government*. Available at: https://www.gov.uk/guidance/medical-devices-conformity-assessment-and-the-ukca-mark

Purpose	To demonstrate the product meets regulatory requirements and to achieve a risk classification, from Class I (lowest perceived risk) to Class III (highest) that provides a quantified measure of risk.	To help developers understand NHS motivations and standards for buying digital and data-driven technology products.	To help developers collect the appropriate evidence to demonstrate clinical effectiveness and economic impact for their data-driven product.	To ensure safe and fair handling of personal data and minimise risks arising from improper data handling, and as a legal compliance exercise.	To provide 'presumption of conformity' of good clinical practice during design, conduct, recording and reporting of clinical investigations, to assess the clinical performance or effectiveness and safety of medical devices.
Output?	Classification of device, e.g. Class IIb, to be displayed outwardly. Technical documentation on metrics like safety and performance. Declaration of conformity resulting in CE/UKCA mark.	No specific output.	No specific output.	Completed DPIA document, probably a Word document or PDF saved as an internal record. While there is a general obligation to notify a data subject about the processing of their data, there is no obligation to publish the results of the DPIA.[77]	No specific output.
What evidence is needed?	Chemical, physical and biological properties of the product, and that the benefits outweigh risks and achieve claimed performance (proven with clinical evidence). Manufacturers must also ensure ongoing safety by carrying out post-market surveillance under guidance of MHRA.	Value proposition, mission statement, assurance testing of product, and asks users to think of data ethics frameworks.	Evidence of effectiveness of the technology and evidence of economic impact standards. Uses contextual questions to help identify 'higher-risk' DHTs, e.g those with users from 'vulnerable groups'.	Evidence of compliance with the GDPR regulation on data categories, data handling, redress procedures, scope, context and nature of processing. Asks users to identify source and nature of risk on individuals, with an assessment of likelihood and severity of harm. The DPIA also includes questions on consultations with 'relevant stakeholders'.	Evidence of how rights, safety and wellbeing of subjects are protected, scientific conduct, and responsibilities of principal investigator. The ISO 14971 requires teams to build a risk-management plan, including a risk-assessment to identify possible hazards.

77 Kaminski, M.E. and Malgieri, G. (2020). 'Algorithmic impact assessments under the GDPR: producing multi-layered explanations'. *International Data Privacy Law*, 11,2, pp.125-144. Available at: https://doi.org/10.1993/idpl/ipaa020

How does the AIA differ from, and complement this process?	Building off the risk-based approach, the AIA encourages further reflexivity on who gets to decide and define these risks and impacts, broadening out the MHRA classification framework. It also helps teams better understand impacts beyond risk to the individual. This AIA proposes a DAC to assess AIAs; in future, this could be a notified body (as in the MHRA initiative).	The code of conduct mentions DPIAs; this AIA would move beyond data-processing risk. The guide considers impacts, such as impact on patient outcomes: the AIA adds weight by detailing procedure to achieve this impact: e.g. improving clinical outcomes because of the comprehensive assessment of negative impacts, producing a record of this information to build evidence, and releasing it publicly for transparency.	Our impact identification exercise uses similar Q&A prompts to help developers assess risk, but the AIA helps interrogate the 'higher-risk' framing: higher risk for who? Who decides? The participatory workshop broadens out the people involved in these discussions, to help build a more holistic understanding of risk.	AIAs and DPIAs differ in scope and procedure, and we therefore recommend a copy of the DPIA also be included as part the NMIP data access process. AIAs seek to encourage a reflexive discussion among project teams to identify and mitigate a wider array of potential impacts, including environmental, societal or individual harms. DPIAs are generally led by a single data-controller processor, legal expert or information-governance team, limiting scope for broader engagement. The AIA encourages engagement of individuals who may be affected by an AI system even if they are not subjects of that data.	The process of identifying possible impacts and building into a standardised framework is confluent between the ISO 14971 and the AIA. However, the AIA does not measure for quality assurance or clinical robustness to avoid duplication. Instead, it extends these proposals by helping developers better understand the needs of their users through the participatory exercise.

There is no single body responsible for regulation for data-driven technologies in healthcare. Some of the key regulatory bodies for the development of medical devices in the UK that include an AI component are outlined in Table 2 below:

Table 2: Key regulatory bodies for data-driven technologies in healthcare

Regulatory body	Medicines and Healthcare products Regulatory Agency (MHRA)	Health Research Authority (HRA)	Information Commissioner's Office (ICO)	National Institute for Health & Care Excellence (NICE)
Details	The MHRA regulates medicine, medical devices and blood components in the UK. It ensures regulatory requirements are met and has responsibility for setting post-market surveillance standards for medical devices.[78] AI systems that are regulated by the MHRA as medical devices.	If AI systems are developed within the NHS, projects will need approval from the Health Research Authority, who oversee responsible use of medical data, through a process that includes seeking ethical approval from an independent Research Ethics Committee (REC).[79] The REC evaluates for ethical concerns around research methodology but does not evaluate for the potential broader societal impacts of research.	The ICO is the UK's data protection regulator. AI systems in health are often trained on, and process individual patients' health data. There must be a lawful basis for use of personal data in the UK,[80] and organisations are required to demonstrate understanding of and compliance with data security policies, usually by completing a data protection impact assessment (DPIA). The ICO assurance team may conduct audits of different health organisations to ensure compliance with the Data Protection Act.[81]	NICE supports developers and manufacturers of healthcare products, including data-driven technologies like AI systems, to be able to produce robust evidence for their effectiveness. They have produced comprehensive guidance pages for clinical conditions, quality standards and advice pages, including the NICE evidence standards framework for digital health technologies (see 'Table 1' above).[82]

It is important to emphasise that this proposed AIA process is not a replacement for the above governance and regulatory frameworks. NMIP applicants expecting to build or validate a product from NMIP data are likely to go on to complete (or in some cases, have already completed), the processes of product registration and risk classification, and are likely to have experience working with frameworks such as the 'Guide to good practice' and NICE evidence standards framework.

78 Health Research Authority (HRA). *Research Ethics Service and Research Ethics Committees*. Available at: https://www.hra.nhs.uk/about-us/committees-and-services/res-and-recs/

79 Health Research Authority (HRA). *Research Ethics Service and Research Ethics Committees*. Available at: https://www.hra.nhs.uk/about-us/committees-and-services/res-and-recs/

80 Health Research Authority (HRA). *Research Ethics Service and Research Ethics Committees*. Available at: https://www.hra.nhs.uk/about-us/committees-and-services/res-and-recs/

81 Information Commissioner's Office (ICO). *Findings from ICO audits of NHS Trusts under the GDPR*. Available at: https://ico.org.uk/media/action-weve-taken/audits-and-advisory-visits/2618960/health-sector-outcomes-report.pdf

82 National Institute for Care Excellence (NICE). *Evidence standards framework for digital health technologies*. Available at: https://www.nice.org.uk/about/what-we-do/our-programmes/evidence-standards-framework-for-digital-health-technologies

Similarly, DPIAs are widely used across multiple domains because of their legal basis and are critical in healthcare, where use of personal data is widespread across different research and clinical settings. As Table 1 shows, we recommend to the NHS AI Lab that NMIP applicant teams should be required to submit a copy of their DPIA as part of the data access process, as it specifically addresses data protection and privacy concerns around the use of NMIP data, which have not been the focus of the AIA process.

The AIA process complements these processes by providing insights into potential impacts through a participatory process with patients and clinicians (see 'What value can AIAs offer developers of medical technologies?' p. 38) The AIA is intended as a tool for building robust accountability by providing additional routes to participation and external scrutiny: for example, there is no public access requirement for DPIAs, so we have sought to improve documentation practice to provide stable records of the process.

This project also made recommendations to the NHS AI Lab around best practice for documenting the NMIP dataset itself, using a datasheet that includes information about the dataset's sources, what level of consent it was collected under, and other necessary information to help inform teams looking to use NMIP data and conduct AIAs – because datasets can have downstream consequences for the impacts of AI systems developed with them.[83, 84]

Where does an AIA add value among existing processes?

Viewing impacts of AI systems with a wider lens

Given the high-stakes context of healthcare, many accountability initiatives use matrices of technical assurance, like accuracy, safety and quality. Additionally, technologies that build from patient data would need to be assessed for their impacts on individual data privacy and security.

83 Boyd, K.L. (2021). 'Datasheets for datasets help ML engineers notice and understand ethical issues in training data'. *Proceedings of the ACM on Human-Computer Interaction*, 5, 438. [online] Available at: http://karenboyd.org/blog/wp-content/uploads/2021/09/Datasheets_Help_CSCW-5.pdf

84 Gebru, T., Mogenstern, J., Vecchione, B., Wortman Vaughan, J., Wallach, H., Daumé III, H. and Crawford, K. (2018). Datasheets for datasets. *ArXiv* [online] Available at: https://arxiv.org/abs/1803.09010

This AIA process encourages project teams to consider a wider range of impacts on individuals, society and the environment in the early stages of their work. It encourages a reflexive consideration of common issues that AI systems in healthcare may face, such as considerations around the explainability and contestability of decisions, potential avenues for misuse or abuse of a system, and where different forms of bias may appear in the development and deployment of a system.

Broadening the range of perspectives in a governance process

Beyond third-party auditing, there is little scope in the current landscape for larger-scale public engagement activity to deliberate on governance or regulation of AI in the healthcare space. Public and patient participation in health processes is widespread, but many organisations lack the resources or support to complete public engagement work at the scale they'd like to. It emerged from stakeholder interviews that our AIA would need to include a bespoke participatory process, to provide insight into potential algorithmic harm in order to build meaningful, critical AIAs, which in turn will help to build better products.

Standardised, publicly available documentation

Many risk assessments, including other impact assessments like DPIAs, do not have a requirement for completed documentation to be published or for other evidence about how the process was undertaken to be evidenced.[85] It has been demonstrated that the varied applications of AI in healthcare worldwide have led to a lack of consensus and standardisation of documentation around AI systems and their adoption in clinical decision-making settings, which has implications both for evaluation and auditing of these systems, and for ensuring harm prevention.[86] For the NMIP context, the intention was to introduce a level of standardisation across all AIAs to help address this challenge.

85 Gebru, T., Mogenstern, J., Vecchione, B., Wortman Vaughan, J., Wallach, H., Daumé III, H. and Crawford, K. (2018). Datasheets for datasets. *ArXiv* [online] Available at: https://arxiv.org/abs/1803.09010

86 Sendak, M., Gao, M., Brajer, N. and Balu, S. (2020). 'Presenting machine learning model information to clinical end users with model facts labels'. *npj Digital Medicine*, 3,41, p1-4. [online] Available at: https://www.nature.com/articles/s41746-020-0253-3

What value can AIAs offer developers of medical technologies?

With over 80 AI ethics guides and guidelines available, developers express confusion about how to translate ethical and social principles into practice that leads to inertia. To disrupt this cycle, it is vital that technology developers and organisations adopting AI systems have access to frameworks and step-by-step processes to proceed with ethical design.

We interviewed several research labs and private firms developing AI products to identify where an AIA would add value (see 'Methodology' p.90). Our research uncovered that academic research teams, small health-tech start-ups and more established companies all have different considerations, organisational resources and expertise to bring to the table, but there are still common themes that underscore why a developer benefits from this AIA process:

1. Clearer frameworks for meeting NHS expectations. Developers see value in considering societal impacts at the outset of a project, but lack a detailed and actionable framework for thinking about impacts. This kind of AIA exercise can identify potential failure modes within the successful implementation of a medical technology, and can help developers meet the NHS's compliance requirements.

2. Early insights can support and improve patient care outcomes. Some technology developers we interviewed reported a struggle with reaching and engaging patients and representatives of the public at the scale they would like. The AIA enables this larger-scale, meaningful interaction, resulting in novel insights. For applicant teams early on in the development process, the participatory workshop provides important context for how an applicant's AI system might be received. Better understanding patient needs before the majority of system development or application is underway allows for further consideration in design decisions that might have a tangible effect on the quality of patient care in settings supported by an AI system.

3. Building on AI system risk categorisation. Applicants hoping to use NMIP data to build and validate products will also have to undertake the MHRA medical device classification, which asks organisations to assign a category of risk to the product. It can be challenging for AI developers to make a judgement on the risk level of their system, and so the framework requires developers to assign a pre-determined risk category using a flowchart for guidance. It may still be challenging for developers to understand why and how certain attributes or more detailed design decisions correspond to a higher level of risk.

The AIA's reflexive impact identification exercise and participatory workshop move beyond a process of mapping technical details and help build a comprehensive understanding of possible impacts. It also provides space for applicant teams to explore risks or impacts that they feel may not be wholly addressed by current regulatory processes, such as considering societal risk in addition to individual risk of harm.

Case study: NHS AI Lab's National Medical Imaging Platform

In this research, the NHS AI Lab's National Medical Imaging Platform (NMIP) operates as a case study: a specific research context to test the applicability of algorithmic impact assessments (AIAs) within the chosen domain of AI in healthcare. It should be emphasised that this is not an *implementation* case study – rather, it is a case study of designing and building an AIA process. Further work will be required to implement and trial the process, and to evaluate its effectiveness once in operation.

The NHS AI Lab – part of the NHS Transformation Directorate driving the digital transformation of care – aims to accelerate the safe, ethical and effective adoption of AI in healthcare, bringing together government, health and care providers, academics and technology companies to collaborate on achieving this outcome.[87]

The NMIP is an initiative to bring together medical-imaging data from across the NHS and make it available to companies and research groups to develop and test AI models.[88]

It is envisioned as a large medical-imaging dataset, comprising chest X-ray (CXR), magnetic resonance imaging (MRI) and computed tomography (CT) images from a national population base. It is being scoped as a possible initiative after a precursor study, the National COVID Chest Imaging Database (NCCID), which was a centralised database that contributed to the early COVID-19 pandemic response.[89] The NMIP was designed with the intention of broadening the

87 NHS AI Lab. *The NHS AI Lab: accelerating the safe adoption of AI in health and care*. Available at: https://www.nhsx.nhs.uk/ai-lab/

88 NHS AI Lab. *National Medical Imaging Platform (NMIP)*. Available at: https://www.nhsx.nhs.uk/ai-lab/ai-lab-programmes/ai-in-imaging/national-medical-imaging-platform-nmip/

89 NHS AI Lab. *The National COVID Chest Imaging Database*. Available at: https://www.nhsx.nhs.uk/covid-19-response/data-and-covid-19/national-covid-19-chest-imaging-database-nccid/

geographical base and diagnostic scope of the original NCCID platform. At the time of writing, the NMIP is still a proposal and does not exist as a database.

How is AI used in medical imaging?

When we talk about the use of AI in medical imaging, we mean the use of machine-learning techniques on images for medical purposes – such as CT scans, MRI images or even photographs of the body. Medical imaging can be used in medical specialisms including radiology (using CT scans or X-rays) and ophthalmology (using retinal photographs). Machine learning describes when computer software 'learns' to do a task from data it is given instead of being programmed explicitly to do that task. The use of machine learning with images is often referred to as 'computer vision'. The field of computer vision – the use of machine learning (i.e. AI tools) to better process information about images – has had an impact in the medical field over a long period.[90]

For example, AI in medical imaging may be used to make a diagnosis from a radiology image. The machine learning model will be trained on many radiology images ('*training data*') – some which exhibit the clinical condition, and some which don't – and from this will 'learn' to recognise images with the clinical condition, with a particular level of accuracy (they won't always be correct). This model could then be used in a radiology department for diagnosis. Other uses include identifying types or severity of a clinical condition. Currently, these models are mostly intended for use alongside clinicians' opinions.

An example of AI in medical imaging is a software that uses machine learning to read chest CT scans, to detect possible early-stage lung cancer. It does this by identifying lung (pulmonary) nodules, a kind of abnormal growth that forms in the lung. Such products are intended to speed up the CT reading process and claim to lower the risk of misdiagnosis.

The NMIP, as part of the NHS AI Lab, is intended to collect medical images and associated data that could be used to train and validate machine learning models.[91]

90 Esteva, A., Chou, K., Yeung, S., Naik, N., Madani, A., Mottaghi, A., Liu, Y., Topol, E., Dean, J., and Socher, R. (2021). 'Deep learning-enabled medical computer vision'. *npj Digital Medicine*, pp.1-9 [online]. Available at: https://www.nature.com/articles/s41746-020-00376-2

91 NHS AI Lab. *National Medical Imaging Platform*. Available at: https://www.nhsx.nhs.uk/ai-lab/ai-lab-programmes/ai-in-imaging/national-medical-imaging-platform-nmip/

An example product that might be built from a dataset like the NMIP would be a tool that helps to detect the presence of a cardiac tumour by interpreting images, after training on thousands of MRI images that show both presence and no presence of a tumour. As well as detection, AI imaging products may help with patient diagnosis for clinical conditions like cancer, and may also help triage patients based on the severity of abnormality detected from a particular set of images. The developers of these products claim they have the potential to improve health outcomes – by speeding up waiting times for patient diagnosis, for example – and to ease possible resourcing issues at clinical sites.

The NMIP will be available, on application, for developers to test, train and validate imaging products. Organisations with a project that would benefit from access to the NMIP dataset would need to make an application to access the dataset, describing the project and how it will use NMIP data.

From interviews with stakeholders, we envisage that applicants will be seeking access to the NMIP for one of three reasons:

1. To conduct academic or corporate research that uses images from the NMIP dataset.

2. To train a new commercial medical product that uses NMIP data.

3. To analyse and assess existing models or commercial medical products using NMIP data.

This AIA process is therefore aimed at both private *and* public-sector researchers and firms.

In this proposed process, access to the NMIP will be decided by an NHS-operated Data Access Committee (DAC). DACs are already used for access to other NHS datasets, such as the University College London Hospital (UCLH) DAC, which manages and controls access to COVID-19 patient data.[92] There is also a DAC process in place for the NCCID, which will help inform the process for the NMIP.

92 UCL. (2020). *UCLH Covid-19 data access committee set up*. Available at: https://www.ucl.ac.uk/joint-research-office/news/2020/jun/uclh-covid-19-data-access-committee-set

For the NCCID, the DAC evaluates requests for access on criteria such as scientific merit of the project, its technical feasibility, the track record of the research team, reasonable evidence that access to data can benefit patients and the NHS, compliance with the GDPR and NHS standards of information governance and IT security. We anticipate the NMIP will evaluate for similar criteria, and have structured this process so that the AIA complements these other criteria by encouraging research teams to think reflexively about the potential benefits and harms of their project, engage with patients and clinicians to surface critical responses, and present a document outlining those impacts to the DAC.

DACs can deliberate on a number of ethical and safety issues around use of data, as shown in the detailed process outlined below. For example, in the NMIP context, the DAC will be able to review submitted AIAs and make judgements about the clarity and strength of the process of impact identification, but they may also be required to review a DPIA, which we recommend would be a requirement of access. This would provide a more well-rounded picture of how each applicant has considered possible social impacts arising from their project. However, evidence suggests DACs often deliberate predominately around issues of data privacy and the rights of individual data subjects[93, 94] which is not the sole focus of our AIA. Accordingly, the NMIP DAC will be expected to broaden their expertise and understanding of a range of possible harms and benefits from an AI system – a task that we acknowledge is essential but may require additional resource and support.

93 Cheah, P.Y. and Piasecki, J. (2020). 'Data access committees'. *BMC Medical Ethics*, 21, 12 [online] Available at: https://link.springer.com/article/10.1186/s12910-020-0453-z

94 Thorogood A., and Knoppers, B.M. (2017). 'Can research ethics committees enable clinical trial data sharing?'. *Ethics, Medicine and Public Health*, 3,1, pp.56-63.[online] Available at: https://www.sciencedirect.com/science/article/abs/pii/S2352552517300129

The proposed AIA process

Summary

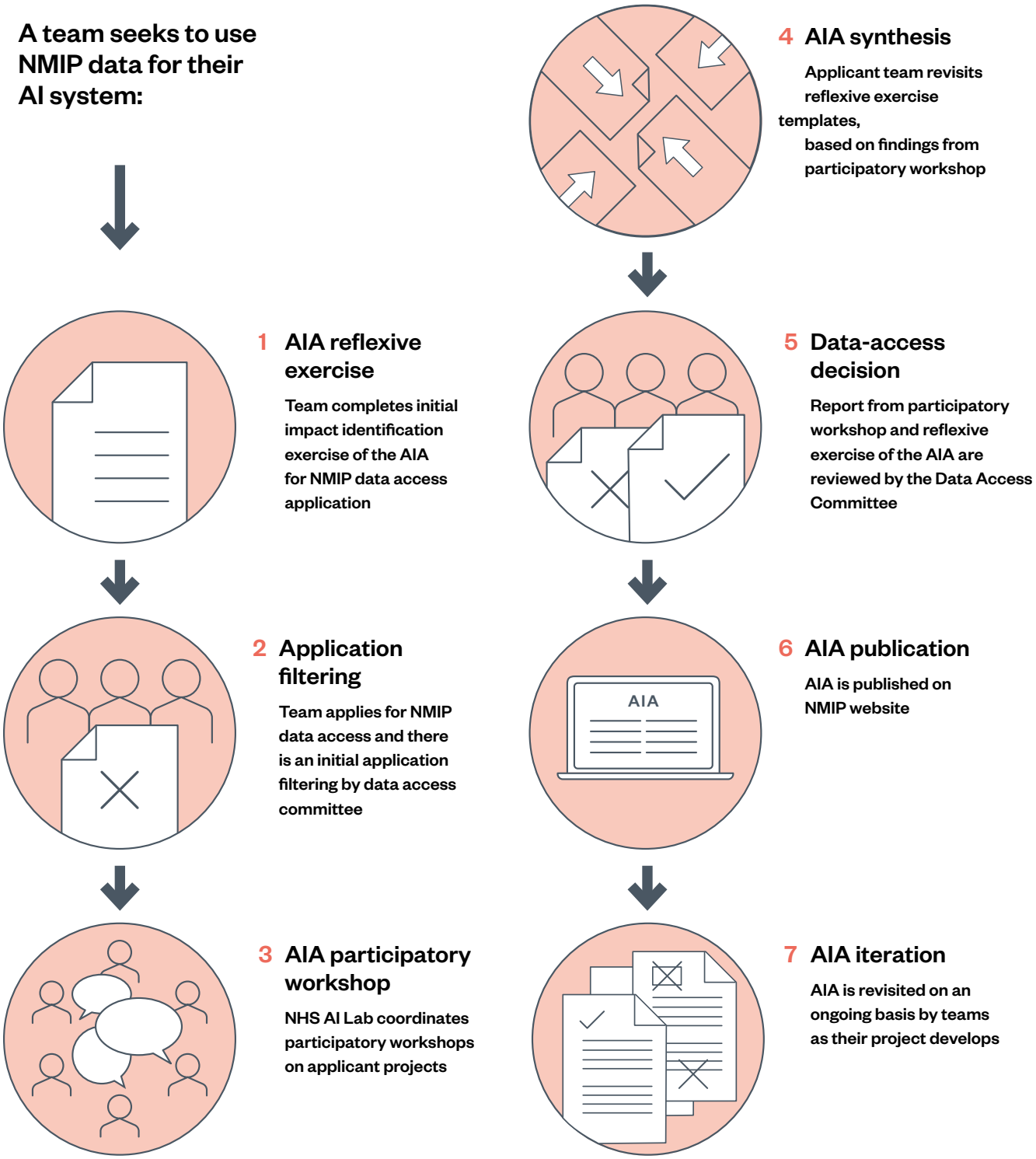
Our AIA process is designed to ensure that National Medical Imaging Platform (NMIP) applicants have demonstrated a thorough and thoughtful evaluation of possible impacts, in order to be granted access to the platform. The process presented here is the final AIA process we recommend the NHS AI Lab implements and makes requisite for NMIP applicants.

While this process is designed specifically for NHS AI Lab and NMIP applicants, we expect it to be of interest to policymakers, AIA researchers and those interested in adopting algorithmic accountability mechanisms.

As the first draft of this process, we expect the advice to develop over time as teams trial the process and discover its strengths and limitations, as the public and research community provide feedback on this process, and as new AIA practical frameworks emerge.

The process consists of seven steps, with three main exercises, or points of activity, from the NMIP applicant perspective: a reflexive impact identification exercise, a participatory workshop, and a synthesis of the two (AIA synthesis). See figure 1 (below) for an overview of the process.

Figure 1: Proposed AIA process



The described AIA process is initiated by a request from a team of technology developers to access the NMIP database. It is the project that sets the conditions for the AIA – for example, the dataset might be used to build a completely new model or, alternatively, the team may have a

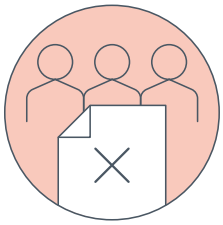
pre-existing functioning model that the team would like to be retrained or validated on the NMIP. At the point that the applicant team decides the project would benefit from NMIP data access, they will be required to begin the AIA process as part of their data-access request.

1. AIA reflexive exercise

A reflexive impact identification exercise submitted to the NMIP DAC as part of the application to access the NMIP database.

The exercise uses a questionnaire format, drawing from best-practice methodologies for impact assessments. It prompts teams to answer a set of questions that consider common ethical considerations in AI and healthcare literature, and potential impacts that could arise, based on the best-case and worst-case scenarios for their project. It then asks teams to discuss the potential harms arising from uses based on the identified scenarios, and who is most likely to be harmed.

Applicants are required to consider harms in relation to their perceived importance or urgency, i.e. weight of the consequence, difficulty to remediate and detectability of the impact. Teams are then asked to consider possible steps to mitigate these harms. These responses will be captured in the AIA template.

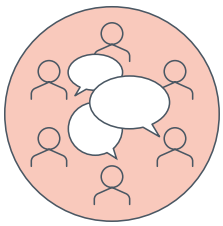


2. Application filtering

At this stage, the NMIP DAC filters initial applications.

Applications are judged according to the engagement shown toward the exercise: whether they have completed all the prompts set out in the AIA template, and whether the answers to the AIA prompts are written in an understandable format, reflecting serious and careful consideration to the potential impacts of this system.

Those deemed to have met the above criteria will be invited to take part in the participatory workshop, and those that have not are rejected until the reflexive exercise is properly conducted.

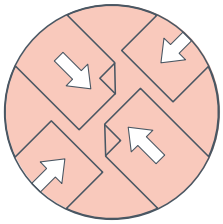


3. AIA participatory workshop

Step three is a participatory process designed as an interactive workshop, which would follow a 'citizen's jury' methodology,[95] equipping patients and members of the public with a means to pose questions and pass judgement on the harm and benefit scenarios identified in the previous exercise (and possibly uncovering some further impacts).

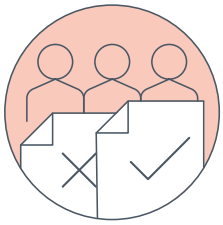
The workshop would be an informal setting, where participants should feel safe and comfortable to ask questions and receive support from the workshop facilitator and other experts present. An NHS AI Lab rapporteur would be present to document the workshop's deliberations and findings on behalf of the patient and public participants.

After the exercise has concluded, the participants will asynchronously review the rapporteur's account and the list of impacts identified, and review any mitigation plans the applicant team has devised in this window.



4. AIA synthesis

The applicant team(s) revisit the template, and work the new knowledge back into the template document, based on findings from the participatory workshop.



5. Data-access decision

This updated template is re-submitted to the DAC, who will also receive the account of the participatory workshop from the NHS AI Lab rapporteur.

The DAC then makes a decision on whether to grant access to the data, based on a set of criteria relating to the potential risks posed by this system, and whether the product team has offered satisfactory mitigations to potentially harmful outcomes.

95 Gastil, J. (ed.) (2005). *The deliberative democracy handbook: strategies for effective civic engagement in the twenty-first century*. 1. ed., 1. impr. Hoboken, N.J: Wiley.

6. AIA publication

The completed AIAs are then published in a central, easily accessible location – probably the NMIP website – for internal record-keeping and the potential for external viewing on request.

7. AIA iteration

The AIA is then revisited on an ongoing basis by project teams, and at certain trigger points.

Such reviews may be prompted by notable events, such as changes to the proposed use case or a significant model update. In some cases, the DAC may, as part of its data access decision, mandate selected project teams to revisit the AIA after a certain period of time to determine if they are allowed to retain access, at its discretion.

Learnings from the AIA process

Building on the outline process, in this section we describe detailed learnings and recommendations. For each stage of the process we provide justification for the recommendation, details for implementation, as well as outlining some uncertainties and challenges surfaced during this study, including some practical constraints associated with implementation.

1. AIA reflexive exercise

Recommendation

For this first step we recommend a reflexive impact identification and analysis exercise to be run within teams applying for the NMIP. This exercise enables teams to identify possible impacts, including harms, arising from development and deployment of the applicant team's AI system by working together through a template of questions and discussion prompts.

Implementation detail

1. Applicant teams should identify a lead for this exercise (we recommend the project team lead, principal investigator or product lead) and a notetaker (for small teams, these roles may be combined).

2. Once identified, the lead should organise and facilitate a meeting with relevant team members to work through the prompts (estimated time: two-to-three hours). The notetaker will be responsible for writing up the team's answers in the template document (estimate one-to-two hours).

3. Teams will first give some **high-level project information**: the purpose, the intended uses of the system, model of research; the project team/organisation; the inputs and outputs for the system, and the stakeholders affected by the system, including users and the people it serves.

4. The template then guides applicants through some **common ethical considerations** in the context of healthcare, AI and the algorithmic literature, including whether the project could exacerbate health inequalities, increase surveillance, impact the relationship between stakeholders, have environmental effects or whether it could be intentionally or unintentionally misused.

5. In the next section, **impact identification and scenarios**, teams reflect on some possible scenarios arising from use of the system and what impacts they would have, including the best-case scenario when the system is working as designed and the worst-case scenario, when not working in some way. This section also asks for some likely challenges and hurdles encountered on the way to achieving the best-case scenario, and the socio-environmental requirements necessary to achieve success, such as a stable connection to the internet, or training for doctors and nurses.

6. In the final section, teams undertake **potential harms analysis** – based on the scenarios identified earlier in the exercise, teams should consider what the potential harms resulting from implementation that should be designed for, and who is at risk of being harmed. Teams should also make a judgement on the perceived importance, urgency, difficulty and detectability of the harms.

7. Teams are given space to detail some possible mitigation plans to minimise identified harms.

All thinking is captured by the notetaker in the AIA template document. It is estimated that this exercise will take three-to-five hours in total (discussion and documentation) to complete.

Frictions and learnings

The impact assessment design:
This exercise is designed to encourage critical dialogue and reflexivity within teams. By stipulating that evidence of these discussions should be built into a template, the AIA facilitates records and documentation for internal and external visibility.

This format of the exercise draws from an approach often used in impact assessments, including AIAs, adapting a Q&A or questionnaire format to

prompt teams to consider possible impacts, discuss the circumstances in which impacts might arise and who might be affected. This exercise was also built to be aligned with traditional internal auditing processes – a methodical, internally conducted process with the intention to enrich understanding of possible risk once a system or product is in deployment.[96]

Other impact assessments request consideration of high-level categories of impact, such as privacy, health or economic impact.[97] In this process, we chose to prompt consideration of impacts by asking teams to consider what they perceive to be the best and worst-case arising from the use of the system. Our hope is this will make the exercise easier to digest and engage with for those less familiar with adopting ethics discussions into their work. Impacts should relate to the kinds of challenges that are associated with AI systems, such as concerns around bias, misuse, explainability of findings, contestability; but may also include several of the 'Generic data and digital health considerations' outlined by the NHS such as concerns around patient involvement and ownership of health and care data.[98]

Some other formats of impact assessment ask an assessor to assign a risk level (e.g. low to high) for a product, which may in turn dictate additional follow-up actions from a developer. The regulatory framework adopted by the MHRA classifies AI as a medical device using a risk-level system. We therefore expect most applicants to be familiar with this format, with some projects having a system that will have already undergone this process at the point of NMIP application. This process is intended to complement risk categorisation, giving developers and project leads a richer understanding of potential harmful impacts of an AI system, to better inform this self-assessment of risk.

The current MHRA framework is focused primarily on risks to the individual, i.e. the risk of harm to a patient if a technology fails (similar to a DPIA, which focuses on the fundamental rights of an individual and the attendant risks of improper personal data handling). Assessment

96 Raji, D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D. and Barnes, P. (2020). 'Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing'. *Conference on Fairness, Accountability, and Transparency*, pp.33–44. Barcelona: ACM. Available at: https://doi.org/10.1145/3351095.3372873

97 Government of Canada. (2020). *Algorithmic impact assessment tool*. Available at: https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html

98 NHSX. (2019). *Artificial Intelligence: how to get it right*. Available at: https://www.nhsx.nhs.uk/media/documents/NHSX_AI_report.pdf

of individual risk is an important component of ensuring safe, fair and just patient outcomes, but we designed the reflexive exercise to go further than this framing of risk. By asking developers to reflexively examine impacts through a broader lens, they are able to consider some possible impacts of their proposed system on society, such as whether it might reinforce systemic biases and discrimination against certain marginalised groups. This process is not about identifying a total measure or quantification of risk, as incorporated in other processes such as the MHRA medical device classification system, but about better understanding impacts and broadening the range of impacts given due consideration.

It is possible that as a precedent of completed AIAs develops, the NHS may in future be able to ascribe particular risk categories based on common criteria or issues they see. But at this stage, we have intentionally chosen not to ask applicants to make a value judgement on risk severity. As described in '5. Data-access decision' p. 60, we recommend that applicants should also submit their DPIA as part of the application process, concurrently with the reflexive exercise. However, if the NHS team determined that not all project teams would have to undertake a DPIA prior to receiving full assessment by the DAC, then we would recommend a template amendment to reflect more considerations around data privacy.

Making complex information accessible:
We also experienced a particular challenge in this exercise of couching the language of ethical values like accountability and transparency into practical recommendations for technology developers to understand and comply with, given that many may be unfamiliar with AIAs and wider algorithmic accountability initiatives.

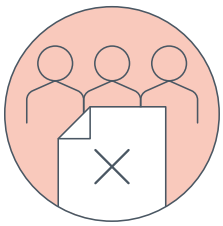
Expert stakeholder interviews with start-ups and research labs revealed enthusiasm for an impact assessment, but less clarity on and understanding of the types of questions that would be captured in an AIA and the kinds of impacts to consider. To address these concerns, we produced a detailed AIA template that helps applicant teams gain understanding of possible ethical considerations and project impacts, how they might arise, who they might impact the most, and which of the impacts are likely to result in harm. We also point to data, digital, health

and algorithmic considerations from NHSX's AI report,[99] which many teams may be familiar with, and instruct applicants on how best to adopt plain language in their answers.

Limitations of the proposed process

We note that different NMIP applicants may have different interactions with the exercise once trialled: for example, an applicant developing an AI system from scratch may have higher expectations of what it might achieve or what its outcomes will be once deployed, than an applicant who is seeking to validate an existing system, and may be armed with prior evidence. Once these sorts of considerations emerge after the AIA process is trialled, we may be able to make a more robust claim about the utility of a reflexive exercise as a quality component of this AIA.

This exercise (in addition to the participatory workshop, below) is probably applicable to other AIA-process proposals, with some amendments. However, we emphasise that domain expertise should be used to ensure the reflexive exercise operates as intended, as a preliminary exploration of potential benefits and harms. Further study will be required to see how well this exercise works in practice for both project teams and the NHS DAC, how effective it proves at foreseeing which impacts may arise, and whether any revisions or additions to the process are required.



2. Application filtering

Recommendation

We recommend that the DAC conducts an application-filtering exercise once the reflexive exercise has been submitted, to remove applications that are missing basic requirements, or will not meet the criteria for reasons other than the AIA.

Depending on the strength of the application, the DAC can choose to either reject NMIP applications at this stage, or invite applicants to

99 NHSX. (2019). *Artificial Intelligence: how to get it right*. Available at: https://www.nhsx.nhs.uk/media/documents/NHSX_AI_report.pdf

proceed to the participatory workshop. Most of these criteria will be established by the NMIP team, and we anticipate they will be similar to those for the National COVID Chest Imaging Database (NCCID), which sees an administrator and a subset of the DAC members involved in application screening for completeness, technical and scientific quality, and includes safeguards for conflicts of interest.[100]

Implementation detail

Based on the review process for the National COVID Chest Imaging Database (NCCID), the precursor platform to the NMIP, the NHS AI Lab is likely to adopt the following filtering procedure:

1. After the applicant team completes the reflexive exercise and builds the evidence into the template document, the teams submit the exercise as part of their application to the NMIP dataset.

2. The person(s) fielding the data-access requests, such as the administrator, will check that all relevant information required in the AIA template has been submitted. In the instance that some is missing, the administrator will go back to the applicant to request it. If the initial submission is very incomplete, the application is declined at this step.

3. The administrator passes the acceptable applications to members of the DAC

4. The members of the DAC chosen to filter the application are given opportunity to declare any conflict of interests with the applicant's project. If one or both has a conflict of interest, they should select another expert to review the AIA at this stage.

5. The selected members assess the AIA and make a judgement call about whether the applicant team can proceed on to the participatory workshop. The decision will be based on technical and scientific quality criteria established by the DAC, as well as review of the AIA, for which we suggest the following initial filtering criteria:

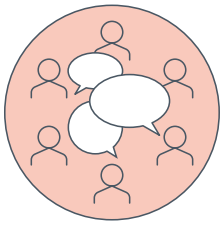
100 Based on NHS AI Lab documentation reviewed in research.

a. The project team has completed the reflexive exercise.
b. The answers to the AIA prompts are written in an understandable format, avoiding jargon and other technical language.
c. The answers given do not identify any unacceptable impacts that would place severe risk on the health and wellbeing of patients and clinicians.

Frictions and learnings

Scale and resource:
The NHS AI Lab raised a challenge of scale for this process and suggested that a triage phase might be needed to identify applicants to prioritise for the participatory workshop in the event of a high volume of applications. However, in prioritising some applications over others for the participatory workshop, there is a risk of pre-empting what the findings would be; implicitly judging which applications might be of greater risk. Without a history of participatory impact identification, that judgement is a challenging one for the DAC or NHS AI Lab to make, and risks prioritising impacts and harms already well understood by established processes.

Consequently, we have recommended that *all* applicants should undertake the participatory workshop but that, as the process begins to be trialled, the DAC will develop a paradigm based on previous applications, which may enable them to make a judgement call for applicant suitability.



3. AIA participatory workshop

Recommendation

We recommend that the NHS AI Lab runs centralised participatory impact assessment workshops, in order to bring a diverse range of lived experience and perspectives to the AIA process, and to support NMIP applicants who may not have the resources or skills to run a participatory process of this size. We believe this process will also help project teams identify risks and mitigation measures for their project and provide valuable feedback on how their project might be successfully integrated into the UK's healthcare system.

Implementation detail

1. We recommend the NHS AI Lab sets up a paid panel of 25–30 patients and members of the public who represent traditionally underrepresented groups who are likely to be affected by algorithms that interact with NMIP data, across dimensions such as age, gender, region, ethnic background, socio-economic background. This includes members of the impacted groups, in addition to adequate representatives of certain communities (e.g. a representative from a grassroots immigrant support organisation being able to speak to migrant experience and concerns).

2. All panellists will be briefed on their role at an induction session, where they will be introduced to each other, learn more about AI and its uses in healthcare, and about the NMIP and its aims and purpose. Participants should also be briefed on the aims of the workshops, how the participatory process will work and what is expected of them.

3. The panellists will be invited to discuss the applicant team's answers to the reflexive exercise, possibly identifying other harms and impacts not already addressed by applicant teams. This is designed as an interactive workshop following a 'citizen's jury' methodology,[101] equipping participants with a means to deliberate on the harms and benefit scenarios identified in the previous exercises (and possibly uncovering some further impacts). The workshop would be designed as an informal setting, where participants should feel safe and comfortable to ask questions and receive support from the workshop facilitator and other experts present. The workshops will involve a presentation from the developers of each applicant team on what their system does or will do, what prompted the need for it, how the system uses NMIP data, what outputs it will generate, how the AI system will be deployed and used and what benefits and impacts it will bring and how these were considered (reporting back evidence from the reflexive exercise).

101 For more information on citizen's jury methodology, see Involve. *Citizen's jury*. Available at: https://www.involve.org.uk/resources/methods/citizens-jury

4. The panellists will then deliberate on the impacts identified to consider whether they agree with the best, worst and most-likely scenarios produced, what other considerations might be important and possible next steps. The facilitator will support this discussion, offering further questions and support where necessary.

5. The Lab would have ownership over this process, and contribute staffing support by supplying facilitators for the workshops, as well as other miscellaneous resources such as workshop materials. The facilitator will coordinate and lead the workshop, with the responsibility for overseeing the impact identification tasks, fielding questions, and for leading on the induction session.

6. 8–12 panellists will be present per workshop, to avoid the same people reviewing every application (8–12 participants per applicant project suggests a different combination for each workshop if there are six or more applications to review). We recommend one workshop per application, and that workshops are batched, so they can run quarterly.

7. Also present at these workshops would be two 'critical friends': independent experts in the fields of data and AI ethics/computer science and biomedical sciences, available to judge the proposed model with a different lens and offer further support. An NHS rapporteur will also be present to provide an account of the workshop on behalf of the patient and public panellists that is fed back to the NHS AI Lab. The rapporteur's account will be reviewed by the panellists to ensure it is an accurate and full representation of the workshop deliberations.

8. Members of the applicant team will be present to observe the workshop and answer any questions as required, and will then return to their teams to update the original AIA with the new knowledge and findings.

9. This updated AIA is then re-submitted to the NMIP DAC.

For the full participatory AIA implementation detail, see Annex 3.

Frictions and learnings

The bespoke participatory framework:
The impetus for producing a tailor-made participatory impact assessment framework came from combining learning from AIA literature with challenges with public and patient involvement (PPI) processes that were raised in our stakeholder interviews.

There is consensus within the AIA literature that public consultation and participation are an important component of an AIA, but little consensus as to what that process should involve. Within the UK healthcare context, there is an existing participatory practice known as public and patient involvement (PPI). These are frameworks that aim to improve consultation and engagement in how healthcare research and delivery is designed and conducted.[102] PPI is well-supported and a common feature in healthcare research: many research funders now require evidence of PPI activity from research labs or companies as a condition of funding.[103] Our interviews revealed a multitude of different approaches to PPI in healthcare, with varying levels of maturity and formality. This echoes research findings that PPI activities, and particularly reporting and documentation, can often end up as an 'optional extra' instead of an embedded practice.[104]

Our stakeholder interviews highlighted that PPI processes are generally supported among both public and private-sector organisations and are in use across the breadth of organisations in the healthcare sector, but many expressed challenges with engagement from patients and the public. One interviewee lamented the struggle to recruit participants: *'Why would they want to talk to us? [...] it might be that we're a small company: why engage with us?'*

In an earlier iteration of our AIA, we recommended applicants design and run the participatory process themselves, but our interviews identified varying capacity for such a process, and it was decided that this would

102 Health Research Authority (HRA). *What is public involvement in research?* Available at: https://www.hra.nhs.uk/planning-and-improving-research/best-practice/public-involvement/

103 University Hospital Southampton. *Involving patients and the public.* Available at: https://www.uhs.nhs.uk/ClinicalResearchinSouthampton/For-researchers/PPI-in-your-research.aspx

104 Price, A., Schroter, S., Snow, R., et al. (2018). 'Frequency of reporting on patient and public involvement (PPI) in research studies published in a general medical journal: a descriptive study'. *BMJ Open 2018*;8:e020452. [online] Available at: https://bmjopen.bmj.com/content/8/3/e020452

be too onerous for individual organisations to manage on their own – particularly for small research labs. One organisation, a healthtech start-up, reported that having more access to funding enabled them to increase the scope and reach of their activity: '*We would always have been keen to do [PPI work], but [funding] is an enabler to do it bigger than otherwise*'. These interviews demonstrated the desire to undertake public participation, but also showcased a lack of internal resources to do so effectively.

There is also the risk that having individual applicants run this process themselves may create perverse incentives for 'participation washing', in which perspectives from the panellists are presented in a way that downplays their concerns.[105] It will be preferable for this process to be run by a centralised body that is independent from applicants, as well as independent from the NHS, and can provide a more standardised and consistent experience across different applications. This led to the proposal that NHS AI Lab run the participatory process centrally, to ease the burden on applicants, reduce any conflicts of interests, and to gain some oversight over the quality of the process.

Lack of standardised method:
To address the challenge of a lack of standardised methods for how to run public engagement, we decided building a bespoke methodology for participation in impact assessment was an important recommendation for this project. This would provide a way to stabilise the differing PPI approaches currently in use in healthcare research, align the NMIP with best practice for public deliberation methods and ameliorate some concerns over a lack of standard procedure.

This process also arose from an understanding that developing a novel participatory process for an AIA requires a large amount of both knowledge and capacity for the process to operate meaningfully and produce high-quality outputs. To address this challenge, we have drawn from the Ada Lovelace Institute's experience and expertise in designing and delivering public engagement in the data/AI space, as well as best practice from the field in order to co-ordinate an approach to a participatory AIA.

105 Sloane, M., Moss, E., Awomolo, O., & Forlano, L. (2020). 'Participation is not a design fix for machine learning'. *arXiv*. [online] Available at: https://arxiv.org/abs/2007.02423

Resource versus benefit:
The participatory workshop is an extension of many existing participatory procedures in operation, and consequently is time and resource intensive for the stakeholders involved, but has significant benefits.

Beyond bringing traditionally underrepresented patients into the process, which is an important objective, we believe that the workshop offers the potential to build more intuitive, higher-quality products that understand and can respond to the needs of end users.

For applicants early on in the project lifecycle, the participatory workshop is a meaningful opportunity to engage with the potential beneficiaries of their AI system: patients (or patient representatives). It means possible patient concerns around the scope, applicability or use case for the proposed system can be surfaced while there is still opportunity to make changes or undertake further reflection before the system is in use. In this way, the participatory workshop strengthens the initial internally conducted exercise of impact identification.

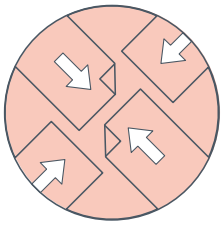
Researchers have argued that, to support positive patient outcomes in clinical pathways in which AI systems are used to administer or support care, evaluation metrics must go beyond measures of technical assurance and look at how use of AI might impact on the quality of patient care.[106] The process provides a useful forum for communication between patients and developers, in which developers may be able to better understand the needs of the affected communities, and therefore build products better suited to them.

The process at scale:
Given that the NMIP is purported as a national initiative, challenges and uncertainties have arisen from the NHS AI Lab around how this process would operate at scale. We have sought to address this by recommending the NHS AI Lab run the workshops in batches, as opposed to on a rolling basis. We have also suggested that over time the NHS AI Lab may be able to use previous data-access decisions to triage future applications, and possibly even have applicants with similar projects sharing the same workshop.

106 Kelly, C.J., Karthikesalingam, A., Suleyman, M., Corrado, G. and King. D. (2019). 'Key challenges for delivering clinical impact with artificial intelligence' *BMC Medicine*. 29 October, 17: 195. [online] Available at: https://bmcmedicine.biomedcentral.com/articles/10.1186/s12916-019-1426-2

Recompensing panellists appropriately:
All participants must be renumerated for their time, but we also recognise the inherent labour of attending these workshops, which may not be adequately covered or reflected by the renumeration offered.

Limitations of resource:
Other organisations hoping to adopt this exercise may be practically constrained by a lack of funding or available expertise. We hope that in future, as participatory methods and processes grow in prominence and the AIA ecosystem develops further, alternate sources of funding and support will be available for organisations wanting to adopt or adapt this framework for their contexts.



4. AIA synthesis

Recommendation

We recommend that after the participatory workshop is completed, the applicant team synthesises its findings with the findings from their original AIA template (completed in the reflexive exercise), building the knowledge produced back into the AIA, in order to ensure the deliberation-based impacts are incorporated and that applicant teams respond to them.

The synthesis step is a critical phase in accountability processes.[107] It serves to summarise and consolidate the information gained throughout the AIA process and ensure they are incorporated into the assessment of impacts, and actionable steps for mitigations of harm.

Implementation detail

1. Throughout the AIA process, thinking and discussion should be captured in the AIA template document, allowing documentation to be revisited after each exercise and as a record for future updates of the AIA.

107 Raji, D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D. and Barnes, P. (2020). 'Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing'. *Conference on Fairness, Accountability, and Transparency*, pp.33–44. Barcelona: ACM. Available at: https://doi.org/10.1145/3351095.3372873

2. Once the synthesis exercise has been completed, the AIA is considered complete. The AIA is then ready to be resubmitted to the NMIP DAC.

Frictions and learnings

Goals of the process:

The synthesis exercise serves two purposes:

1. Documentation provides a stable record of activity throughout the AIA process, for internal and external viewing (by the DAC at re-submission phase, and the public, post-publication).

2. It encourages a critical, reflexive response to the impact-identification process, by asking applicants to revisit their responses in the light of new information and knowledge from the participatory panel in the participatory workshop.

The NHS rapporteur report also incentivises a high-quality synthesis exercise, as it allows the DAC to refer back to a full account of the workshop, which has been reviewed by the patient and public panellists, to come to a final judgement of the applicant team's willingness to incorporate new feedback and ability to be critical of their own processes and assumptions.

Supporting meaningful participation:
Some scholars in the public-participation literature have argued that meaningful participation should be structured around co-design and collaborative decision-making[108]. We designed our participatory process as a feed-in point, for patients and the public to discuss and put forward ideas on how developers of AI systems might address possible benefits and harms.

108 Madaio, M, Stark, L, Wortman Vaughan, J, Wallach, H. (2020). 'Co-designing checklists to understand organisational challenges and opportunities around fairness in AI'. *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp.1-14 [online] Available at: https://dl.acm.org/doi/abs/10.1145/3313831.3376445

The synthesis exercise ensures that these ideas become incorporated into the AIA template document, creating an artefact that the DAC and members of the public can view. Developers can then refer back to the AIA template as required throughout the remainder of the project-development process.

Though undertaking a process of synthesis boosts opportunity for reflexivity and reflection, there is no guarantee that the broader stakeholder discussions that occur in the participatory workshop will lead to tangible changes in design or development practice.

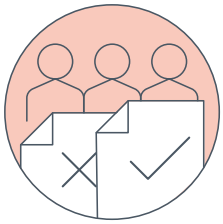
In an ideal scenario, participants would be given opportunity to have direct decision-making power on design decisions. It has been argued that the ideal level of participation in AI contexts amounts to *participatory co-design*, a process that sees people and communities affected by the adoption of AI systems become directly involved in the design decisions that may impact them.[109] In *Participatory data stewardship*, the Ada Lovelace Institute describes a vision for people being enabled and empowered to actively collaborate with designers, developers and deployers of AI systems – a step which goes further than both 'informing' people about data use, via transparency measures and 'consulting' people about AI system design, via surveys or opinion polls.[110]

Similarly, while beyond the scope of this study, it is suggested that participants would ideally be invited for multiple rounds of involvement, such as a second workshop prior to system deployment.[111] This would enable participants to be able to make a permissibility call on whether the system should be deployed, based on the applicant team's monitoring and mitigation plan, and again once the system is in use, so participants can voice concerns or opinions on any impacts that have surfaced *ex post*.

109 Costanza-Chock, S. (2020) Design justice: community-led practices to build the worlds we need. Cambridge: MIT Press

110 Ada Lovelace Institute. (2021). *Participatory data stewardship*. Available at: https://www.adalovelaceinstitute.org/report/participatory-data-stewardship/

111 Madaio, M, Stark, L, Wortman Vaughan, J, Wallach, H. (2020). 'Co-designing checklists to understand organisational challenges and opportunities around fairness in AI'. *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp.1-14 [online] Available at: https://dl.acm.org/doi/abs/10.1145/3313831.3376445



5. Data-access decision

Recommendation

We recommend that the NHS AI Lab uses the NMIP DAC to assess the strength and quality of each AIA, alongside the assessment of other material required as part of the NMIP application.

Implementation detail

1. We recommend the DAC comprises at least 11 members, including academic representatives from social sciences, biomedical sciences, computer science/AI and legal fields and representatives from patient communities (see 'NMIP DAC membership, process and criteria for assessment' below).

2. Once the participatory workshop is complete, and the applicant team has revised their AIA template, providing new evidence, the template is resubmitted to the DAC. In order to come to a data-access decision, the DAC follows the assessment guidelines, reviewing the quality of both the reflexive exercise and the workshop based on the detail in the AIA output template and the strength of engagement in the participatory workshop, as well as the supporting evidence from the NHS rapporteur. If the accounts and evidence have significant divergence, the applicant team may either be instructed to undertake further review and synthesis, or be denied access.

3. The assessment guidelines include questions on whether the DAC agrees with the most-likely, worst-case, and best-case scenarios identified, based on their knowledge of the project team's proposal, and whether the project meets the requirements and expectations of existing NHSX frameworks for digital health technologies.[112] The guidelines also establish normative guidelines for the DAC to ascertain the acceptability of the AIA based on whether the project meets the requirements and expectations of NHSX's 'What good looks like'

112 Such as the *NHS Code of Conduct for Data-driven Health and Care Technology*, available at: https://www.gov.uk/government/publications/code-of-conduct-for-data-driven-health-and-care-technology/initial-code-of-conduct-for-data-driven-health-and-care-technology and NHSX's 'What Good Looks Like' framework, available at: https://www.nhsx.nhs.uk/digitise-connect-transform/what-good-looks-like/what-good-looks-like-publication/

framework, which includes: 'being well led', 'empowering citizens' and 'creating healthy populations' among others. If the process was deemed to have been completed incorrectly or insufficiently, or if the project is deemed to have violated normative or legal red lines, the DAC would be instructed to reject the application.

4. In the NCCID data-access process, if the application is accepted, the applicant team would be required to submit a data-access framework contract and a data-access agreement. We believe the existing documentation from the NCCID, if replicated, would probably require applicant teams to undertake a DPIA, to be submitted with the AIA and other documentation at this stage. (If the Lab decides that not all applicant teams would be required to undertake a DPIA prior to this stage, we recommend the reflexive exercise be amended to include more data privacy considerations – see 'AIA reflexive exercise'). Once these additional documents are completed and signed, access details are granted to the applicant.

NMIP DAC membership, process and criteria for assessment

We recommended to the NHS AI Lab that the NMIP DAC comprises at least 11 members:

1. a chair from an independent institution
2. an independent deputy chair from a patients-rights organisation or civil-society organisation
3. two representatives from the social sciences
4. one representative from the biomedical sciences
5. one representative from the computer science/AI field
6. one representative with legal and data ethics expertise
7. two representatives from patient communities or patients-rights organisations
8. two members of the NHS AI Lab.

For the NCCID, an administrator was required to help manage access requests, which would probably be required in the NMIP context. Similarly, we anticipate that in addition to the core committee, a four-person technical-review team of relevant researchers, data managers and lab managers who can assess data privacy and security questions, may be appointed by the DAC (as per the NCCID terms).

The responsibilities of the DAC in this context are to consider and authorise requests for access to the NMIP, as well deciding whether to continue or disable access. They will base this decision on criteria and protocols for assessment and will assess the completed AIA, including the participatory workshop using the NHS AI Lab rapporteur's account of the exercise (as described previously on p.42) as additional evidence.

For the NCCID project, the DAC assessed applications along the criteria of scientific merit, technical feasibility and reasonable evidence that access to the data can benefit patients and the NHS. This may be emulated in the NMIP, but broader recommendations for application assessment beyond the AIAs are out of scope for this study.

As guidelines to support the DAC to make an assessment about the strength of the AIA we provide two groups of questions to consider: the AIA process and the impacts identified as part of the process.

Questions on the *process* include:

1. Did the project team revise the initial reflexive exercise after the participatory workshop was conducted?

2. Are the answers to the AIA prompts written in an understandable format, reflecting serious and careful consideration to the potential impacts of this system?

3. Did the NHS AI Lab complete a participatory AIA with a panel featuring members of the public?

4. Was that panel properly recruited according to the the NHS AI Lab AIA process guide?

5. Are there any noticeable differences between the impacts/concerns/risks/challenges the the NHS AI Lab rapporteur identified and what the AIA document states? Is there anything unaddressed or missing?

Questions on the *impacts* include:

1. Based on your knowledge of the project team's proposal, do you agree with the most likely, worst-case, and best-case scenarios they have identified?

2. Are there any potential stakeholders who may be more seriously affected by this project? Is that reason well-justified?

3. For negative impacts identified, has the project team provided a satisfactory mitigation plan to address these harms?

    a. If you were to explain these plans to a patient who would be affected by this system, would they agree these are reasonable?

Frictions and learnings

The role of the DAC and accountability:
In an accountability relationship between applicant teams, the NHS AI Lab and members of the public, the DAC is the body that can pose questions and pass judgement, and ultimately is the authoritative body to approve, deny or remove access to the NMIP.

The motivation behind this design choice was the belief that a DAC could contribute to two primary goals of this AIA: accountability, by building an external forum to which the actor must be accountable; and standardisation, whereas applications grow in volume, the DAC will be able to build a case law of common benefits and harms arising from impact assessments and independent scrutiny, which may offer different or novel priorities to the AIA not considered by the applicant team(s).

Recommendations for the composition of the DAC contribute to broadening participation in the process, by bringing different forms of expertise and knowledge into the foreground, particularly those not routinely involved in data-access decision-making such as patient representatives.

The literature review surfaced a strong focus on mandatory forms of assessment and governance in both the healthcare domain and AIA scholarship. In healthcare, many regulatory frameworks and legislation including the MHRA Medical Device Directive, a liability-based regime, ask developers to undertake a risk assessment to provide an indication of the safety and quality of a product and gatekeep entry to the market.

Initiatives like the MHRA Medical Device Directive address questions relating to product safety, but lack robust accountability mechanisms, a transparency or public-access requirement, participation and a broader lens

to impact assessment, as discussed in this report. This AIA was designed to add value for project teams, the NHS and patients in these areas.

Legitimacy without legal instruments:
In the AIA space, recent scholarship from Data & Society argues that AIAs benefit from a 'source of legitimacy' of some kind in order to scaffold accountability and suggest that this might include being adopted under a legal instrument.[113] However, there is not currently a legal requirement for AIAs in the UK, and the timeline for establishing such a legal basis is outside of the scope of this case study, necessitating a divergence from the literature. This will be a recurring challenge for AIAs as people look to trial and evaluate them as a tool at a faster pace than they are being adopted in policy.

This AIA process attempts to address this challenge by considering how alternative sources of legitimacy can be wielded, in lieu of law and regulation. Where top-down governance frameworks like legal regimes may prohibit participation and deliberation in decision-making, this AIA process brings in both internal and external perspectives of harms and benefits of AI systems. We recommended the NHS AI Lab make use of a DAC to prevent organisations building and assessing AIAs independently, as self-assessed AIAs. This may allay some concerns around interpretability and whether the AIA might end up being self-affirming.[114]

Potential weaknesses of the DAC model:
In this study, the DAC has the benefit of giving the AIA process a level of independence and some external perspective. We recognise however that the appointment of a DAC may prove to be an insufficient form of accountability and legitimacy once in place. We recommend the membership of the DAC comprise experts from a variety of fields to ensure diverse forms of knowledge. Out of 11 members, only two are patient representatives, which may disempower the patients and undermine their ability to pass judgement.

113 Madaio, M, Stark, L, Wortman Vaughan, J, Wallach, H. (2020). 'Co-designing checklists to understand organisational challenges and opportunities around fairness in AI'. *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp.1-14 [online] Available at: https://dl.acm.org/doi/abs/10.1145/3313831.3376445

114 Individual interpretation of soft governance frameworks may lead to some organisations picking and choosing which elements to enact, which is known as 'ethics washing'. See: Floridi, L. (2019). 'Translating principles into practices of digital ethics: five risks of being unethical'. *Philosophy & technology*, 32, pp.185-193 [online] Available at: https://link.springer.com/article/10.1007/s13347-019-00354-x

The DAC functions as an accountability mechanism in our context because the committee members are able to pass judgement and scrutinise on the completed AIAs. However, the fairly narrow remit of a DAC may result in an AIA expertise deficit, where the committee may find their new role of understanding and responding to AIAs and adopting a broad lens to impact challenging.

The data-access context means that it is not possible to further specify additional project points where applicant teams might benefit from reflexive analysis of impacts, such as at ideation phase, or at the final moment pre-deployment, that would make the process more iterative.

Additionally, the DAC still sits inside the NHS as a mechanism and is not wholly external: in an ideal scenario, an AIA might be scrutinised and evaluated by an independent third party. This also raises some tensions around whether there might be, in some cases, political pressures on the NMIP DAC to favour certain decisions. The DAC also lacks statutory footing, putting it at the mercy of NHS funding: if funds were to be redirected elsewhere, this could leave the DAC on uncertain ground.

As other AIAs outside this context begin to be piloted, a clearer understanding of what 'good' accountability might look like will emerge, alongside the means to achieve this as an ideal.

6. AIA publication

Recommendation

To build transparency and accountability, we recommend that the NHS AI Lab publishes all completed AIAs, by publishing the final AIA template, alongside the name and contact details of a nominated applicant team member who is willing to field further information and questions on the process from respective interested parties on request.

We also recommend the Lab publishes information on the membership of the DAC, its role and the assessment criteria, so that external viewers can learn how data-access decisions are made.

Implementation detail

1. We recommend that the Lab publishes completed AIAs on a central repository, such as an NMIP website,[115] that allows for easy access by request from the public. Only AIAs that have completed both the reflexive exercise and the participatory workshop will be published. However, there may be value in the DAC periodically publishing high-level observations around the unsuccessful AIAs (as a collective, as opposed to individual AIAs), and we also note that individual applicant teams may want to publish their AIA independently, regardless of the access decision.

2. The designed AIA template is intended to ensure the AIAs are able to be easily published by the Lab without further workload, and the template is an accessible document that follows a standard format. It is likely a nominated NHS AI Lab team member will be needed to publish the AIAs, such as an administrator.

Frictions and learnings

Public access to AIAs:
There is widespread consensus within the AIA and adjacent literature that public access to AIAs and transparent practice are important ideals.[116, 117, 118] Public access to documentation associated with decision-making has been put forward as a way to build transparency and, in turn, public trust in the use of AI systems.[119] This is a particularly significant dimension for a public-sector agency.[120]

115 Such as the website designed for the National Covid Chest Imaging Database (NCCID), see: https://nhsx.github.io/covid-chest-imaging-database/

116 Latonero, M. and Agarwal, A. (2021). *Human rights impact assessments for AI: learning from Facebook's failure in Myanmar*. Carr Center for Human Rights Policy: Harvard Kennedy School. Available at: https://carrcenter.hks.harvard.edu/publications/human-rights-impact-assessments-ai-learning-facebook%E2%80%99s-failure-myanmar

117 Loi, M. in collaboration with Matzener, A., Muller, A. and Spielkamp, M. (2021). *Automated decision-making systems in the public sector. An impact assessment tool for public authorities*. Algorithm Watch. Available at: https://algorithmwatch.org/en/wp-content/uploads/2021/06/ADMS-in-the-Public-Sector-Impact-Assessment-Tool-AlgorithmWatch-June-2021.pdf

118 Selbst, A.D. (2018). 'The intuitive appeal of explainable machines'. *Fordham Law Review* 1085. [online] Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3126971

119 Reisman, D., Schultz, J., Crawford, K. and Whittaker, M. (2018). *Algorithmic impact assessments: a practical framework for public agency accountability*. AI Now Institute. Available at: https://ainowinstitute.org/aiareport2018.pdf

120 Hildebrandt, M. (2012). 'The dawn of a critical transparency right for the profiling era'. *Digital Enlightment Yearbook*, pp.41-56 [online] Available at: https://repository.ubn.ru.nl/handle/2066/94126

Transparency is an important underpinning for accountability, where access to reviewable material helps to structure accountability relationships and improves the strength and efficacy of an impact assessment process.[121] Making AIAs public means they can be scrutinised and evaluated by interested parties, including patients and the public, and also enables deeper understanding and learning from approaches among research communities. Publication in our context also helps standardise applicants' AIAs.

Other impact assessments, such as data protection impact assessments (DPIAs) and human rights impact assessments (HRIAs) have drawn criticism for not demonstrating consistent publication practice,[122] therefore missing opportunities to build accountability and public scrutiny. We also base our recommendation in part on audit processes, where transparent, auditable systems equip developers, auditors and regulators with knowledge and investigatory powers for the benefit of the system itself, but also the wider ecosystem.[123]

Putting transparency into practice:
In this study, we found that translating transparency ideals into practice in this context required some discussion and consensus around establishing the publishable output of the AIA. During our interview process, we surfaced some potential concerns around publishing commercially sensitive information from private companies. The AIA as it appears in the AIA template document does not necessitate commercially sensitive information or detailed technical attributes.

Further transparency mechanisms:
In this context, full transparency is not necessarily achieved by publishing the AIA, and other mechanisms might be needed for more robust transparency. For example, for organisations interested in transparent model reporting, we recommend developers consider completing and publishing a model card template – a template developed by Google

121 Metcalf, J., Moss, E., Watkins, E.A., Ranjit, S. and Elish, M.C. (2021). 'Algorithmic impact assessments and accountability: the co-construction of impacts'. *Conference on Fairness Accountability, and Transparency* [online] Available at: https://dl.acm.org/doi/pdf/10.1145/3442188.3445935

122 Metcalf, J., Moss, E., Watkins, E.A., Ranjit, S. and Elish, M.C. (2021). 'Algorithmic impact assessments and accountability: the co-construction of impacts'. *Conference on Fairness Accountability, and Transparency* [online] Available at: https://dl.acm.org/doi/pdf/10.1145/3442188.3445935

123 Singh, J, Cobbe, J and Norval, C. (2019). 'Decision provenance: harnessing data flow for accountable systems'. *IEEE Access*, 7, pp. 6592-6574 [online]. Available at: https://arxiv.org/abs/1804.05741

researchers to increase machine learning model transparency by providing a standardised record of system attributes.[124] This framework has been adapted to a medical context, based on the original proposal from the team at Google.[125]

7. AIA iteration

Recommendation

We recommend that project teams revisit and update the AIA document at certain trigger points: primarily if there is a significant change to the system or its application.

We also recommend a two-year review point in all cases, because it can be hard to identify what constitutes a 'significant change'. The exercise is designed to be a valuable reflection opportunity for a team, and a chance to introduce new team members who may have joined in the intervening time to the AIA process. The DAC might also make suggestions for an appropriate time period for revision in certain cases, and revision of the AIA could be a requirement of continued access.

Implementation detail

A potential process of iteration might be:

1. After a regular interval of time has elapsed (e.g. two years), project teams should revisit the AIA. For some applicants, this might occur after the proposed AI system has entered into deployment. In this scenario, previously unanticipated impacts may have emerged.

2. Reviewing the AIA output template and updating with new learnings and challenges will help strengthen record-keeping and reflexive practice.
3. All iterations are recorded in the same way to allow stable documentation and comparison over time.

124 More information on model cards, including example model cards, can be found here: https://modelcards.withgoogle.com/about

125 Sendak, M., Gao, M., Brajer, N. and Balu, S. (2020). '"The human body is a black box"': supporting clinical decision-making with deep learning' *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM: New York, NY, USA, pp. 99–109. Available at: https://doi.org/10.1145/3351095.3372827

4. If revision is a condition of continued access, the DAC may see fit to review the revised AIA.

5. The revised AIA is then published alongside the previous AIA, providing important research and development findings to the research community, as with each AIA iteration, new knowledge and evidence may be surfaced.

Frictions and learnings

Benefits of *ex post* assessment:
Although we consider our AIA primarily as a tool for pre-emptive impact assessment, this iterative process provides a means for an AIA to function as both an ex ante and ex post assessment, bridging different impact-assessment methodologies to help build a more holistic picture of benefits and harms. This will capture instances where impacts emerge that have not been adequately anticipated by a pre-emptive AIA.

This would align our AIA with methods like AI assurance,[126] which offer a possible governance framework across the entire AI-system lifecycle, of which impact assessment is one component. There are other similar mechanisms already in place in the healthcare sector, such as the ISO/TR 20416 post-market surveillance standards, which provide users with a way to identify 'undesirable effects' at pace.[127]

Revising the AIA also equips teams with further meaningful opportunity for project reflection[128].

126 Information Commissioner's Office (ICO). (2019). *An overview of the auditing framework for artificial intelligence and its core components*. Available at: https://ico.org.uk/about-the-ico/news-and-events/ai-blog-an-overview-of-the-auditing-framework-for-artificial-intelligence-and-its-core-components/

127 International Standards Organization (ISO). (2021). *New ISO standards for medical devices*. Available at: https://www.iso.org/news/ref2534.html

128 Raji, D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D. and Barnes, P. (2020). 'Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing'. *Conference on Fairness, Accountability, and Transparency*, pp.33–44. Barcelona: ACM. Available at: https://doi.org/10.1145/3351095.3372873

Limitations of the model:
Many impact assessment proposals suggest adopting an incremental, iterative approach to impact identification and evaluation, identifying several different project points for activity across the lifecycle.[129, 130] However, as with other components of AIAs, many do not detail a specific procedure for monitoring and mitigation once the model is deployed.

Trigger points for iteration will probably vary across NMIP use cases owing to the likely breadth and diversity of potential applicants. The process anticipates that many applicants will not have fully embarked on research and development at the time of application, so the AIA is designed primarily as an *ex ante* tool, equipping NMIP applicants with a way to assess risk prior to deployment, while there is still opportunity to make design changes. We consider it as a mechanism that is equipped to *diagnose* possible harms so, accordingly, the AIA may be an insufficient mechanism to *treat* or address harms.

Healthcare and other contexts:
Although we recommend iteration of an AIA, the proposed process does not include an impact mitigation procedure. In the context of AI systems for healthcare, post-deployment monitoring fall under the remit of medical post-market surveillance, known as the medical device vigilance system, and can report any 'adverse incidents' to the MHRA.[131]

The aim of iteration of the AIA is therefore to ensure impacts anticipated by the participatory process are addressed and new potential impacts can be identified. It provides impetus for continual reflection, building good practice for future products, and for ensuring thorough documentation into the future. This context is specific to our study: policymakers and researchers interested in trialling AIAs may find that building an *ex post* monitoring or evaluation framework is appropriate in domains where existing post-deployment monitoring is lacking.

129 Information Commissioner's Office (ICO). *Data protection impact assessments*. Available at: https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/accountability-and-governance/data-protection-impact-assessments/

130 The Equality and Health Inequalities Unit. (2020). *NHS England and NHS Improvement: Equality and Health Inequalities Impact Assessment (EHIA)*. Available at: https://www.england.nhs.uk/wp-content/uploads/2020/11/1840-Equality-Health-Inequalities-Impact-Assessment.pdf

131 Medicines & Healthcare products Regulatory Agency (MHRA). *Guidance: Medical device stand-alone software including apps (including IVDMDs)*. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/999908/Software_flow_chart_Ed_1-08b-IVD.pdf

Applicability of this AIA to other use cases

This case study differs from existing proposals and examples of AIAs in three ways. Those wanting to apply the AIA process will need to consider the specific conditions of other domains or contexts:

1. Healthcare domain

At the time of writing, this is the first detailed proposal for use of an AIA in a healthcare context. Healthcare is a significantly regulated area in the UK, particularly in comparison to other public-sector domains. There is also notable discussion and awareness of ethical issues in the sector, with recognition that many AI applications in healthcare would be considered 'high risk'. In the UK, there are also existing public participation practices in healthcare – typically referred to as 'patient and public involvement and engagement' (PPIE) – and requirements for other forms of impact assessment, such as DPIAs and Equalities Impact Assessments. This means that an AIA designed for this context can rely on existing processes – and will seek to avoid unnecessary duplication of those processes – that AIAs in other domains cannot.

2. Public and private-sector intersection

AIA proposals and implementation have been focused on public-sector uses, with an expectation that those conducting most of the process will be a public-sector agency or team.[132, 133, 134] While AIAs have not yet been applied in the private sector, there has been some application of human rights impact assessments to technology systems,[135] which may surface overlapping concerns through a human-rights lens. There are also similarities with proposals around internal-auditing approaches in the private sector.[136] To date, this case study is unique in looking explicitly

132 Reisman, D., Schultz, J., Crawford, K. and Whittaker, M. (2018). *Algorithmic impact assessments: a practical framework for public agency accountability*. AI Now Institute. Available at: https://ainowinstitute.org/aiareport2018.pdf

133 Ada Lovelace Institute, AI Now Institute, Open Government Partnership. (2021). *Algorithmic accountability for the public sector.* Available at: https://www.opengovpartnership.org/wp-content/uploads/2021/08/algorithmic-accountability-public-sector.pdf

134 Government of Canada. (2020). *Algorithmic impact assessment tool*. Available at: https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html

135 WATERFRONToronto. (2020). *Preliminary Human Rights Impact Assessment for Quayside Project.* Available at: http://blog.waterfrontoronto.ca/nbe/portal/wt/home/blog-home/posts/preliminary+human+rights+impact+assessment+for+quayside+project

136 Raji, D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D. and Barnes, P. (2020). 'Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing'. *Conference on Fairness, Accountability, and Transparency*, pp.33–44. Barcelona: ACM. Available at: https://doi.org/10.1145/3351095.3372873

at the intersection of public and private sector – with applications being developed by a range of mainly private actors for use of data originating in the public sector, with some oversight from a public-sector department (NHS).

3. Data-access context

This AIA is being proposed as part of a data-access process for a public-sector dataset (the NMIP). This is, to our knowledge, unique in AIAs so far. The DAC provides a forum for holding developers accountable where other proposals for AIAs have used legislation or independent assessors – to require the completion of the AIA, to evaluate the AIA and to prevent a project proceeding (or at least, proceeding with NHS data) if the findings are not satisfactory.

These differences, and their implication for the design of this AIA, should be considered by anyone looking to apply parts of this process in another domain or context. We expect elements of this process, such as the AIA template and exercise formats, to prove highly transferrable.

However, the core accountability mechanism – that the AIA is both required and reviewed by the DAC – is not transferrable to many potential AIA use cases outside data access; an alternative mechanism would be needed.

Similarly, the centralisation of both publication and resourcing for the participatory workshops with the NHS AI Lab may not be immediately transferrable – though one could imagine a central transparency register and public-sector resource for participatory workshops providing this role for mandated public-sector AIAs.

Seven operational questions for AIAs

Drawing on findings from this case study, we identify seven operational questions for those considering implementing an AIA process in any context, as well as considerations for how the NMIP AIA process addresses these issues.

1. How to navigate the immaturity of the assessment ecosystem?

AIAs are an emerging method for holding AI systems more accountable to those who are affected by them. There is not yet a mature ecosystem of possible assessors, advisers or independent bodies to contribute to or run all or part of an AIA process. For instance, in environmental and fiscal impact assessment, there is a market of consultants available to act as assessors. There are public bodies and regulators who have the power to require their use in particular contexts under relevant legal statutes, and there are more established norms and standards around how these impact assessments should be conducted. [137]

In contrast, AIAs do not yet have a consistent methodology, lack any statutory footing to require their use, and do not have a market of assessors who are empowered to conduct these exercises. A further complexity is that AI systems can be used in a wide range of different contexts – from healthcare to financial services, from criminal justice to the delivery of public services – making it a challenge to identify the proper scope of questions for different contexts.

This immaturity of the AIA ecosystem poses a challenge to organisations hoping to build and implement AIAs, who may not have

137 Moss, E., Watkins, E.A., Singh, R., Elish, M.C. and Metcalf, J. (2021). *Assembling accountability: algorithmic impact assessment for the public interest*. Data & Society. Available at: https://datasociety.net/library/assembling-accountability-algorithmic-impact-assessment-for-the-public-interest/

the skills or experience in house. It also limits the options for external and independent scrutiny or assessment within the process. Future AIA processes must identify the particular context they are operating in, and scope their questions to meet that context.

In the NMIP case study, this gap is addressed by centring the NMIP DAC as the assessor of NIMP AIAs. They are a pre-existing group already intended to bring together a range of relevant skills and experience with independence from the teams developing AI, as well as with authority to require and review the process.[138]

We focus the NMIP AIA's scope on the specific context of the kinds of impacts that healthcare products and research could raise for patients in the UK, and borrow from existing NHS guidance on the ethical use of data and AI systems to construct our questions. In addition, under this proposal, the NHS AI Lab itself would organise facilitation of the participatory workshops within the AIA.

2. What groundwork is required prior to an AIA?

AIAs are not an end-to-end solution for ethical and accountable use of AI, but part of a wider AI-development and governance process.

AIAs are not singularly equipped to identify and address the full spectrum of possible harms arising from the deployment of an AI system,[139] given that societal harms are unpredictable and some harms are experienced more profoundly by those occupying or holding simultaneous marginalised identities. Accordingly, our AIA should not be understood as a complete solution for governing AI systems.

This AIA process does not replace other standards for quality and technical assurance or risk management already in use in the medical-device sector (see: 'The utility of AIAs in health policy' p.30). Teams hoping to implement AIAs should consider the 'pre' and 'post' AIA work that might be required, particularly given projects may be at different

138 See p.64 for more detail on the data access process, and Annex 2 for a draft terms of reference for the NMIP DAC.

139 Metcalf, J., Moss, E., Watkins, E.A., Ranjit, S. and Elish, M.C. (2021). 'Algorithmic impact assessments and accountability: the co-construction of impacts'. *Conference on Fairness Accountability, and Transparency* [online] Available at: https://dl.acm.org/doi/pdf/10.1145/3442188.3445935

stages, or with different levels of AI governance maturity, at the point that they begin the AIA process.

For example, one proposed stakeholder impact assessment framework sets out certain activities to be taken place at the 'alpha phase'[140] (problem formulation), which includes 'identifying affected stakeholders': applicants may find it helpful to use this as a guide to identify affected individuals and communities early on in the process, and in order to be clear on how different interests might coalesce in this project. This is a useful precursor for completing the impact identification exercises in this AIA.

In the NMIP case study, in recognition of the fact that applicant teams are likely to be in differing stages of project development at the point of application, we make some recommendations for 'pre-AIA' exercises and initiatives that might capture other important project-management processes considered out of the scope of this AIA.

It is also important to have good documentation of the dataset any model or product will be developed on, to inform the identification of impacts. In the case of the NMIP, the AIAs will all relate to the same dataset (or subsets thereof). There is a significant need for documentation around NMIP datasets that sets out key information such as what level of consent the data was collected under, where the data comes from, what form it takes, what kinds of biases it has been tested for, and other essential pieces of information.

We made recommendations to the NHS AI Lab to take the burden of documenting the NMIP dataset using datasheets.[141, 142] For AIAs in different contexts, dataset documentation may also be an essential precondition as it provides an important source of information to consider the potential impacts of uses of that data.

140 Leslie, D. (2019). *Understanding artificial intelligence ethics and safety*. Alan Turing Institute. Available at: https://www.turing.ac.uk/sites/default/files/2019-08/understanding_artificial_intelligence_ethics_and_safety.pdf

141 Boyd, K.L. (2021). 'Datasheets for datasets help ML engineers notice and understand ethical issues in training data'. *Proceedings of the ACM on Human-Computer Interaction*, 5, 438. [online] Available at: http://karenboyd.org/blog/wp-content/uploads/2021/09/Datasheets_Help_CSCW-5.pdf

142 Gebru, T., Mogenstern, J., Vecchione, B., Wortman Vaughan, J., Wallach, H., Daumé III, H. and Crawford, K. (2018). Datasheets for datasets. *ArXiv* [online] Available at: https://arxiv.org/abs/1803.09010

3. Who can conduct the assessment?

Previous studies highlight the importance of an independent 'assessor' in successful impact-assessment models, in other domains such as environmental or fiscal impact assessments.[143] However, most proposals for AIA processes, and the Canadian AIA model in implementation,[144] have instead used self-assessment as the main mechanism.

Part of this difference may be due to whether the focus of an AIA is accountability or reflexivity: accountability prioritises independence of assessment as it creates a relational dynamic between a forum and an actor, whereas reflexivity prioritises self-assessment as a mechanism for learning and improvement on the part of the system developers.

In our NMIP case study, we seek to capture both interests – with the initial exercise allowing a reflexive, in-team process for developers, and the DAC review playing the role of an independent assessor. We acknowledge the significant power this process gives the DAC and the potential limitations of delegating this power to a committee established by the NHS. For example, there may be concerns around the ability of the DAC to make impartial decisions and not those that could serve wider NHS aims. We have included in our recommendations a potential composition of this DAC that includes members of the public, patients or patients-rights advocates, and other independent experts who are external to the NHS.

There is, however, a more immediate and practical constraint for those considering AIAs currently – who can assess. Without the wider ecosystem of assessment mentioned previously, for AIAs proposed in contexts outside a data-access process, or without a centralised body to rely on, it may be a necessary short-term solution for companies to run and assess the AIA and participatory processes themselves. This, however, eliminates much of the possibility for independence, external visibility and scrutiny to improve accountability, so should not be considered a long-term ideal, but rather a response to current practical

143 Moss, E., Watkins, E.A., Singh, R., Elish, M.C. and Metcalf, J. (2021). *Assembling accountability: algorithmic impact assessment for the public interest*. Data & Society. Available at: https://datasociety.net/library/assembling-accountability-algorithmic-impact-assessment-for-the-public-interest/

144 Government of Canada. (2020). *Algorithmic impact assessment tool*. Available at: https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html

constraint. For those building AIA processes in other domains, it will be essential to consider which actors are best equipped to play the role of an independent assessor.

4. How to ensure meaningful participation in defining and identifying impacts?

The literature on AIAs, and other methods of assessing AI systems, makes the case for consultation and participation of multidisciplinary stakeholders,[145] affected communities and the wider public.[146, 147, 148, 149] This can create a more accountable relationship between developers of a technology and those affected by it, by ensuring that impacts are constructed from the perspective of those affected by a system, and not simply those developing a system.

There is, however, differing opinion on the people or groups that should be involved: some proposals are explicit in the requirement to include public perspectives in the impact assessment process, others simply suggest a mix of internal and external stakeholders. Types of participation also vary, and range from simply informing key stakeholders, to consultation, to collaboration for consensus-building.[150]

As with other constitutive components of AIAs, there is currently little procedure for how to engage practically with the public in these

145 It should be noted that public consultation is distinct from public access, which refers to the publication of key documentation and other material from the AIA, as a transparency mechanism. See: Ada Lovelace Institute. (2021). *Participatory data stewardship*, and Moss, E., Watkins, E.A., Singh, R., Elish, M.C. and Metcalf, J. (2021). *Assembling accountability: algorithmic impact assessment for the public interest*. Data & Society. Available at: https://datasociety.net/library/assembling-accountability-algorithmic-impact-assessment-for-the-public-interest/

146 Reisman, D., Schultz, J., Crawford, K. and Whittaker, M. (2018). Reisman, D., Schultz, J., Crawford, K. and Whittaker, M. (2018). *Algorithmic impact assessments: a practical framework for public agency accountability*. AI Now Institute. Available at: https://ainowinstitute.org/aiareport2018.pdf

147 European Commission. (2020). *The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment.* Available at: https://op.europa.eu/en/publication-detail/-/publication/73552fcd-f7c2-11ea-991b-01aa75ed71a1

148 Moss, E., Watkins, E.A., Singh, R., Elish, M.C. and Metcalf, J. (2021). *Assembling accountability: algorithmic impact assessment for the public interest*. Data & Society. Available at: https://datasociety.net/library/assembling-accountability-algorithmic-impact-assessment-for-the-public-interest/

149 Institute for the Future of Work. (2021). *Artificial intelligence in hiring: assessing impacts on equality*. Available at: https://www.ifow.org/publications/artificial-intelligence-in-hiring-assessing-impacts-on-equality

150 For further information on participatory approaches, see: Ada Lovelace Institute. (2021). *Participatory data governance*, and Arnstein, S. (1969). 'A ladder of citizen participation'. *Journal of the American Institute of Planners*, 36, pp.216-224 [online] Available at: https://www.tandfonline.com/doi/full/10.1080/01944363.2018.1559388

processes. Our framework seeks to bridge that gap, drawing from Ada's internal public deliberation/engagement expertise to build a participatory workshop for the NMIP AIA.

A key learning from this process is that there are significant practical challenges to implementing participatory ideals:

- Some participatory exercises may be tokenistic or perfunctory, which means they do nothing to rebalance power between developers and affected communities and may be harmful for participants.[151] Beginning to address this must involve participants being renumerated for their time, given the safety and security to deliberate freely and provide critical feedback, and having assurance that their contributions will be addressed by a developer who will be required to respond to their concerns before the DAC.

- There is a potential challenge in implementing robust, meaningful participatory processes at scale. In our case, the NMIP – as a large dataset comprised of different image formats – has scope to underpin a variety of different algorithms, models and products, so is expected to receive a large number of data-access applications. This means that any framework needs to be flexible and accommodating, and able to be scaled up as required. This may place considerable demand on resources. Pilot studies of our participatory workshop would help us further understand and account for some of these demands, as they arise in practice.

5. What is the artefact of the AIA and where can it be published?

Whether the goal of an AIA process is to encourage greater reflexivity and consideration for harmful impacts from developers or to hold developers of a technology more accountable to those affected by its system, an AIA needs an artefact – a document that comes from the process – to be able to be shared with others, and reviewed and assessed. Most proposals of

151 Sloane, M., Moss, E., Awomolo, O., & Forlano, L. (2020).'Participation is not a design fix for machine learning'. *ArXiv*. [online] Available at: https://arxiv.org/abs/2007.02423

AIAs recommend publication of results or key information,[152] but do not provide a format or template in which to do so.

In public-sector use cases, the Canadian AIA has seen three published AIAs to date, with departments conducting the AIA being responsible for publication of results in an accessible format, and in both official languages – English and French – on the Canadian Open Government portal.[153]

When publishing completed AIAs, an AIA process must account for the following:

- ***What* will be published:** what content, in what format? Our case study provides a template for developers to complete as the first exercise, and update throughout the AIA process, producing the artefact of the AIA that can then be published.

- ***Where* it will be published:** is there a centralised location that the public can use to find relevant AIAs? In our case study, as all AIAs are being performed by applicants looking to use NMIP data, the NMIP can act as a central hub, listing all published AIAs. In public-sector use cases, a public-sector transparency register could be that centralised location.[154]

- **What are the *limitations* and *risks* of publishing:** several of our interview subjects flagged concerns that publishing an AIA may raise intellectual property or commercial sensitivities, and may create a perverse incentive for project teams to write with a mindset for public relations rather than reflexivity. These are very real concerns, but they must be balanced with the wider goal of an AIA to increase accountability over AI systems.

This study seeks to balance this concern in a few ways. In this case study the AIA document would not contain deep detail on the functioning of the system that may raise commercial sensitivities, but rather focus on the potential impacts and a simple explanation of its intended use.

152 Reisman, D., Schultz, J., Crawford, K. and Whittaker, M. (2018). *Algorithmic impact assessments: a practical framework for public agency accountability*. AI Now Institute. Available at: https://ainowinstitute.org/aiareport2018.pdf

153 Government of Canada (2020). *Algorithmic impact assessment tool.* Available at: https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html

154 UK.Gov. (2021). *Algorithmic transparency standard.* Available at: https://www.gov.uk/government/collections/algorithmic-transparency-standard

Study respondents flagged that an AIA within a data-access process may also raise concerns about publishing 'unsuccessful' AIAs – AIAs from applicants to the NMIP who were rejected (which may have been on grounds other than the AIA) – which could raise potential liability issues. Given this constraint, we have chosen to prioritise publication of AIAs that have completed the reflexive exercise and the participatory workshop , and not AIAs that did not proceed past DAC filtering. However, we recognise there could be valuable learnings from AIAs that have been rejected, and would encourage the DAC to share observations and learnings from them, as well as enabling individual teams to voluntarily publish AIAs regardless of data-access outcome.

6. Who will act as a decision-maker on the suitability of the AIA and the acceptability of the impacts it documents?

As well as identifying what standards an AIA should be assessed against, it is necessary to decide who can *assess the assessment*.

There is not yet a standard for assessing the effectiveness of AIAs in a particular context, or a clear benchmark that AIAs can use for what 'good' looks like. This makes it hard to measure both the effectiveness of an individual AIA process in terms of what effects have been achieved or what harms have been prevented, and hard to evaluate different AIA approaches to ascertain which approach is more effective in a particular context.

A potential failure mode of an AIA would be a process that carefully documented a series of likely negative impacts of a system, but then saw the team proceed undeterred with development and deployment.[155] Similarly concerning would be a scenario where an AIA is poorly conducted, surfacing few of the potential impacts, but a team is able to point to a completed AIA as justification for continued development.

An AIA will require a decision to be made about what to do in response to impacts identified – whether that is to take no action (impacts considered

155 Moss, E., Watkins, E.A, Singh, R., Elish, M.C, Metcalf, J. (2021). *Assembling accountability: algorithmic impact assessment for the public interest.* Data & Society. Available at: https://datasociety.net/library/assembling-accountability-algorithmic-impact-assessment-for-the-public-interest/

acceptable), to amend parts of the AI system (impacts require mitigation or action), or to not proceed with the development or use of the system (impacts require complete prevention). This is a high-level decision about the acceptability of the impacts documented in an AIA.

In our contextual example, the applicant team is a voluntary decision-maker (they could propose changes to their system, or choose to end their NMIP application or entire system development as a result of AIA findings). However, the ultimate decision about acceptability of impacts lies with the NMIP DAC who would decide whether data can be made available for the applicant's use case – this is, implicitly, a decision about the acceptability of impacts documented in the AIA (along with other documents) and whether the AIA has been completed to a sufficient standard.

To help the DAC in its decision-making, the proposal includes a draft terms of reference that specifies what a 'good' AIA in this context might look like and what rubric they should review it under. In order to prevent a myopic reading of an AIA, the DAC should comprise of a diverse panel of representatives, including representatives from the NHS AI Lab, the social sciences, biomedical sciences, computer science/AI, legal and data ethics and community representatives. It should also follow standards set for the cadence of committee meetings.

The guidelines instruct the DAC to accept or reject the applicant based on whether the AIA process has been run correctly, with evidence from both the reflexive exercise and the participatory workshop produced as part of the application, reflecting serious and careful consideration of impacts. The impetus behind these approaches is to provide a level of external scrutiny and visibility, which legitimise the process when compared with a wholly self-assessed approach.

In our context, we entrust the NMIP DAC with making the judgement call about the suitability of each AIA, and this then informs the final data-access decision. However, the role of the DAC in the NMIP context is broader than typical, as we are asking members to make an assessment of a variety of potential impacts and harms to people and society, beyond privacy and security of individuals and their data.

Accordingly, AIAs designed for different contexts may require the chosen assessor to fulfil a slightly different role or require additional expertise. Over time, assessors of an AIA will need to arbitrate on the acceptability of

the possible harmful impacts of a system and probably begin to construct clear, normative red lines. Regular and routine AIAs in operation across different domains will lead to clearer benchmarks for evaluation.

7. How will trials be resourced, evaluated and iterated?

Governments, public bodies and developers of AI systems are looking to adopt AIAs to create better understanding of, and accountability for potential harms from AI systems. The evidence for AIAs as a useful tool to achieve this is predominantly theoretical, or based in examples from other sectors or domains. We do not yet know if AIAs achieve these goals in practice.

Anyone looking to adopt or require an AIA, should therefore consider trialling the process, evaluating it and iterating on the process. It cannot be assumed that an AIA is 'ready to go' out of the box.

This project has helped to bridge the gap between AIAs as a proposed governance mechanism, and AIAs in practice. The kinds of expertise, resources and timeframe needed to build and implement an AIA are valuable questions that should be discussed early on in the process.

For trials, we anticipate three key considerations: resourcing, funding and evaluation.

1. To **resource** the design and trialling of an AIA process will require skills from multiple disciplines: we drew on a mix of data ethics, technical, public deliberation, communications and domain expertise (in this case, health and medical imaging).

2. **Funding** is a necessary consideration as our findings suggest the process may prove more costly than other forms of impact assessment, such as a data protection impact assessment (DPIA), due predominantly to the cost of running a participatory process. We argue that such costs should be considered a necessary condition of building an AI system with an application in high-stakes clinical pathways. The cost of running a participatory AIA will bring valuable insight, enabling developers to better understand how their system works in clinical pathways, outside of a research lab environment.

3. A useful trial will require **evaluation**, considering questions such as: is the process effective in increasing the consideration of impacts, does it include those who may be affected by the system in the identification of impacts, does it result in the reduction of negative impacts? This may be done as part of the trial, or through documentation and publication of the process and results for others to review and learn from.

Currently, there are very few examples of AIA practice – just four published AIAs from the Canadian government's AIA process[156] – with few details on the experience of the process or the changes resulting from it. As the ecosystem continues to develop, we hope that clearer routes to funding, trialling and evaluation will emerge, generating new AIAs: though policymakers may be disappointed to find that AIAs are not an 'oven-ready' approach, and that this AIA will need amendments before being directly transferable to other domains, we argue there is real value to be had to in beginning to test AIA approaches within, and across different domains.

156 An example of a publicly-available AIA, from the Canadian Government *Directive on Automated Decision-making* can be found here: https://open.canada.ca/aia-eia-js/?lang=en

Conclusion

This report has set out the assumptions, goals and practical features of a proposed algorithmic impact assessment process for the NHS AI Lab's National Medical Imaging Platform, to contribute to the evidence base for AIAs as an emerging governance mechanism.

It argues that meaningful accountability depends on an external forum being able to pass judgement on an AIA, enabled through standardisation of documentation for public access and scrutiny, and through participation in the AIA, bringing diverse perspectives and relevant lived experience.

By mapping out the existing healthcare ecosystem, detailing a step-by-step process tailored to the NMIP context, including a participatory workshop, and presenting avenues for future research, we demonstrate how a holistic understanding of the use case is necessary to build an AIA that can confront and respond to a broad range of possible impacts arising from a specific use of AI.

As the first detailed proposal for the use of AIAs in a healthcare context, the process we have built was constructed according to the needs of the NMIP: our study adds weight to the argument that AIAs are not 'ready to roll out' across all sectors. However, we have argued that testing, trialling and evaluating AIA approaches will help build a responsive and robust assessment ecosystem, which may in turn generate further AIAs by providing a case law of examples, and demonstrating how certain resources and expertise might be allocated.

This report aligns three key audiences for this work: policymakers interested in AIAs, AIA practitioners and researchers, and developers of AI systems in the healthcare space.

Policymakers should pay attention to how this proposed AIA fits in the existing landscape, and to the findings related to process development that show some challenges, learnings and uncertainties when adopting AIAs.

There is further research to be carried out to develop robust AIA practices. On page 77, we provide researchers with 'Seven operational questions' to consider before adopting and implementing AIAs.

Developers of AI systems that may be required to complete an AIA will want to use the report to learn how it was constructed and how it is implemented, as well as Annex 1 for the 'AIA user guide', which provides step-by-step detail. Building a shared understanding of the value of AIAs, who could adopt them, and what promise they hold for the AI governance landscape, while responding to the nuances of different domain contexts, will be critical for future applications of AIA.

This project has offered a new lens through which to examine and develop AIAs at the intersection of private and public-sector development, and to understand how public-sector activity could shape industry practice in the healthcare space. But this work is only in its infancy.

As this report makes clear, the goals of AIAs – accountability, transparency, reflection, standardisation, independent scrutiny – can only be achieved if there is opportunity for proposals to become practice through new sites of enquiry that test, trial and evaluate AIAs, helping to make sure AI works for people and society.

Methodology

To investigate our research questions and create recommendations for an NMIP-specific AIA process, we adopted three main methods:

- a literature review
- expert interviews
- process development

Our literature review surveyed AIAs in both theory and practice, as well as analogous approaches to improving algorithmic accountability, such as scholarship on algorithm audits and other impact assessments for AI that are frequently adopted in tandem with AIAs. In order to situate discussion on AIAs within the broader context, we reviewed research from across the fields of AI and data ethics, public policy/public administration, political theory and computer science.

We held 20 expert interviews with a range of stakeholders from within the NHS AI Lab, NHSX and outside. These included clinicians and would-be applicants to the National Medical Imaging Platform, such as developers from healthtech companies building on imaging data, to understand how they would engage with an AIA and how it would slot into existing workstreams.

Finally, we undertook documentation analysis of material provided by the NHS AI Lab, NMIP and NCCID teams to help understand their needs, in order to develop a bespoke AIA process. We present the details of this process in 'Annex 1: Proposed process in detail', citing insights from the literature review and interviews to support the design decisions that define the proposed NMIP AIA process.

This partnership falls under NHS AI Lab's broader work programme known as 'Facilitating early-stage exploration of algorithmic risk'.[157]

157 NHS AI Lab. *The AI Ethics Initiative: Embedding ethical approaches to AI in health and care.* Available at: https://www.nhsx.nhs.uk/ai-lab/ai-lab-programmes/ethics/

Acknowledgements

We would like to thank the following colleagues for taking time to review a draft of this paper or offering their expertise and feedback:

- Brhmie Balaram, NHS AI Lab
- Dominic Cushnan, NHS AI Lab
- Emanuel Moss, Data & Society
- Maxine Mackintosh, Genomics England
- Lizzie Barclay, Aidence
- Xiaoxuan Liu, University Hospitals Birmingham NHS Foundation Trust
- Amadeus Stevenson, NHSX
- Mavis Machirori, Ada Lovelace Institute.

This report was lead authored by Lara Groves, with substantive contributions from Jenny Brennan, Inioluwa Deborah Raji, Aidan Peppin and Andrew Strait.

Annex 1: Proposed process in detail

As well as synthesising information about AIAs, this project has developed a first iteration of a process for using an AIA in a public-sector, data-access context. The detail of the process will not be applicable to every set of conditions in which AIAs might be used, but we expect it will provide opportunities to develop further thinking for these contexts.

People and organisations wishing to understand more about, or implement, an AIA process will be interested in the detailed documentation developed for the NMIP and NHS AI Lab:

- NMIP AIA user guide: a step-by-step guide to completing the AIA for applicants to the NMIP.
- AIA reflexive template: the document NMIP applicants will fill in during the AIA and submit to the NMIP with their application.

Annex 2: NMIP Data Access Committee Terms of Reference

Responsibilities

- To consider and authorise requests for access to the National Medical Image Platform (NMIP), a centralised database of medical images collected from NHS trusts.
- To consider and authorise applications for the use of data from the NMIP.
- To consider continuing or disabling access to the NMIP and uses of its data.
- To judge applications using the criteria and protocols outlined in the NMIP's data access documentation request forms, which include but are not limited to:
  - an algorithmic impact assessment (AIA) reflexive template (completed by requesting project teams)
  - an accompanying participatory workshop report (completed by an NHS AI Lab rapporteur on behalf of the patient and public participants for the participatory workshop)
  - a data protection impact assessment (DPIA).
- To judge applications according to the NMIP Data Access Committee (DAC) policy, which includes guidance on the reflexive exercise and participatory workshop requirements. This guidance will be updated regularly and placed on the NMIP website.
- To establish a body of published decisions on NMIP data access requests, as precedents which can inform subsequent requests for NMIP access and use.

- To disseminate policies to applicants and encourage adherence to all guidance and requirements.

Membership

- Membership of the DAC will comprise at least eleven members as outlined below:
  - a chair from an independent institution
  - an independent deputy Chair
  - two academic representatives from the social sciences
  - one academic representatives from the biomedical sciences
  - one academic representative from the computer science/AI field
  - one academic representative with legal and data ethics expertise
  - two non-academic representatives from patient communities
  - two members of the NHS AI Lab.
- In addition to the core DAC, a four-person technical review team will comprise relevant researchers, data managers and lab managers who can assess data privacy and security questions. This team will be appointed by the DAC.
- DAC members will be remunerated for their time according to an hourly wage set by NHS AI Lab.
- An NHS AI Lab participatory workshop rapporteur will attend DAC meetings to provide relevant information when necessary to inform the decisions.
- When reviewing data access requests, the following members from the project team will be in attendance to present their case:
  - the study's principal investigator (PI)

  - a member of the study's technical team.

- When reviewing data access requests, the DAC may request that a representative of the project's funding organisation, members of a technical review team, or representatives reflecting experiential expertise relevant to the project may attend in an *ex officio* capacity to observe and provide information to help inform decisions.

- Members, including the Chair and Deputy Chair, will usually be appointed for three years, with the option to extend for a further three after the first term only. Appointment to the DAC will be staggered in order to ensure continuity of membership. The recruitment process will occur annually, when new appointments are necessary, ahead of the second face-to-face meeting of the year.

- The DAC will co-opt members as and when there is a need for additional expertise. These members will have full voting rights and their term will end on appointment of new members through the annual recruitment process.

Modes of operation

- The DAC will follow the guidance for assessing data access request documentation. Updating this guidance will involve a majority vote of the DAC to approve.

- The DAC will meet virtually to address data access requests once each month. The DAC will meet face to face three times a year to discuss emerging issues in relation to data access and provide information on these to the individual studies and funders. Projects leads will be copied into email correspondence regarding individual applications.

- Quoracy formally requires the attendance of half the full independent members (with at least one independent member with biomedical science expertise and one with social science expertise) and that either the Chair or the Deputy Chair must be present for continuity. For face-to-face meetings, where it is unavoidable, attendance of a member by teleconference will count as being present.

- Comments from the technical review team will be circulated to the DAC along with any applications requesting access to the data.

- Decisions of the DAC on whether to grant access to applications will be based on a majority vote. In the event that either a) a majority decision amongst DAC members is not reached; or b) a project lead has grave concerns that the DAC's decision creates unreasonable risk for the project, the Chair of the DAC will refer the decision to the relevant appeals body.

- Where appropriate, the DAC will take advantage of third-party specialist knowledge, particularly where an applicant seeks to use depletable samples. Where necessary the specialist will be invited to sit on the DAC as a co-opted member.

Reporting

- Decisions of the Committee will be reported on the NMIP website and must be published no more than one month after a decision has been reached. Decisions must be accompanied by relevant documentation from the research.

Annex 3: Participatory AIA process

NHS AI Lab NMIP participatory AIA process outline

Overview

- The recommendation is that NHS AI Lab sets up a paid panel of 25-30 patients and members of the public who reflect the diversity of the population who will be affected by algorithms that interact with NMIP data.
- This panel will form a pool of participants to take part in a small series of activities that form the participatory component of the AIA process.
- When an applicant to NMIP data is running their AIA process, the NHS AI Lab should work with them to set up a workshop with the panel to identify and deliberate on impacts. The applicant then develops responses that address the identified impacts, which the panel members review and give feedback on. The Data Access Committee (DAC) uses the outcomes of this process to support their consideration of the application, alongside the wider AIA.
- The five stages of the participatory component are:

1. recruit panel members
2. induct panel members
3. hold impact identification workshops
4. technology developers (the NMIP applicants) review impacts identified in the workshops and develop plan to address or mitigate them
5. panel review mitigation plans and feedback to NHS AI Lab DAC.

These stages are detailed below, along with an indication of required costs and resources, and additional links for information.

Panel recruitment

The panel forms a pool of people who can be involved in the reflexive impact workshop for each project. This is designed to factor in the panel recruitment and induction burden, by enabling projects to be reviewed in 'batches' – for instance, if the NMIP had quarterly application rounds, a panel may be recruited to be involved in all the reflexive workshops for that round.

Note: the following numbers are estimates based on best practice. Exact numbers may vary depending on expected and actual application numbers.

- 25–30 people who reflect the diversity of the population that might be affected by the algorithm across: age, gender, region, ethnic background, socio-economic background, health condition and access to care. The number 25–30 is designed assuming multiple AIAs are required, to ensure the same people aren't reviewing every algorithm. 25-30 means you could have a different combination of 8–12 participants for each algorithm if there are six or more to review. If the number of AIAs needed is smaller than this, then a smaller panel could be used.

- Recruited either via a social research recruitment agency, or via NHS trusts involved.

- Panel does not need to be statistically representative of the UK public, but instead should reflect the diversity of perspectives and experiences in the populations/communities likely to be affected by the algorithms.[158]

- (Ideally) one or two panel members should sit on the DAC as full members.

- Panel members should be remunerated for their involvement on the panel. The amount should reflect the hours required to participate in all the activities: the induction, the assessment workshops, reviewing materials and feeding back on impact mitigation plans (inc. travel if necessary) (see 'Resourcing and costs').

158 Steel, D., Bolduc, N., Jenei, K. and Burgess, M. (2020). 'Rethinking representation and diversity in deliberative minipublics'. *Journal of Deliberative Democracy*, 16,1, pp.46-57 [online]. Available at: https://delibdemjournal.org/article/id/626/

Panel induction

After being recruited, the NHS AI Lab should run an induction session to inform the panel members about the NMIP, how the application and AIA process works and their role.

Participants:

- **All panel members:** to attend and learn about the NMIP, AIAs – including where this exercise sits in the timeline of the AIA process (i.e. after NMIP applicants have completed internal AIA exercises) and their role.
- **NHS panel coordinator:** to run the session and facilitate discussion.
- **Technology and Society (T&S) professional:** to present to the panel on what algorithms are, what the AIA process is, and some common issues or impacts that may arise.

Structure:

- Two hours, virtual or in-person (for either format, ensure participants have support to access and engage fully).
- Suggested outline agenda:
  - introduction to each other
  - introduction to the NMIP – what it is, what it aims to do
  - introduction to the panel's purpose and aims
  - presentation from T&S professional on what an algorithm is and what an AIA is followed by Q&A
  - interactive exercises and discussion of case studies of specific algorithm use cases, with strawperson examples; mapping how different identities/groups would interact with the algorithm (with a few example patients from different groups).
  - how the panel and participatory AIA process will work
  - what is required of the panel members.

Equipment and tools required:

- Accessible room/venue and or online video-conferencing tool (e.g. Zoom - with provisions for visually or hearing impaired and neurodiverse people as required).
- Slide deck for introductions and presentations (with accessibility provisions).
- Any documentation for further reading (e.g. links to 'about' page of the NMIP, information about AIAs, document outlining participatory process and requirements of participants).

Outputs:

- Participants are equipped with the knowledge they need to be able to be active members of the participatory process.

Participatory workshop

The participatory workshop follows the reflexive exercise and provides the forum for a broad range of people to discuss and deliberate on some impacts of the applicant's proposed system, model or research.

Participants:

- **Panel members (8–12 people):** to participate in the workshop and share their perspectives on the algorithm's potential impacts.
- **Facilitator (one or two people):** to lead the workshop, guide discussion and ensure the participants' views are listened to. Facilitators could be an NHS AI Lab staff member, a user researcher from the applicant organisation or a consultant; either way, they must have facilitation experience and remain impartial to the process. Their role is to ensure the process happens effectively and rigorously, and they should have the skills and position to do so.
- **Rapporteur (one person, may be a facilitator):** to serve the panel in documenting the workshop.

- **Technology developer representative (one or two people):** to represent the technology development team, explain the algorithm, take questions and, crucially, listen to the participants and take notes.

- **(Ideally) 'critical friend' (one person):** a technology and society (T&S) professional to join the workshop, help answer participants' questions, and support participants to fully explore potential impacts. They are not intended to be deeply critical of the algorithm, but to impartially support the participants in their enquiry.

- **(Optional) a clinical 'critical friend' (one person):** a medical professional to play a similar role to the T&S professional.

Structure:

- Three hours, virtual or in-person (for either format, ensure participants have support to access and engage fully).

- Suggested agenda:

  - Introductions to each other and the session, with a reminder of the purpose and agenda (10 mins).

  - Presentation from technology developers about their algorithm, in plain English (20 mins), covering:

    - Who their organisation is, its aims, values and whether it is for or non-profit, if it already works with NHS and how.

    - What their proposed algorithm is: what it aims to do (and what prompted the need for the algorithm), how it works (not in technical detail), what data will be input (both how the algorithm uses NMIP data and the other datasets used to train, if applicable), what outputs the algorithm will generate, how the algorithm will be deployed and used (e.g. in hospitals, via a direct-to-patient app etc.), who it will affect, what benefits it will bring, what impact considerations the team have already considered.

  - Q&A led by the lead facilitator (20 mins).

- A session to identify potential impacts (45–60 mins with a break part way through, and a facilitator taking notes on a [virtual] whiteboard):
  - As one group or in two breakout groups, the participants consider the algorithm and generate ideas for how it could create impacts. With reference to the best, worst and most-likely scenarios that might arise from deployment of the algorithm that applicant teams completed for the reflexive exercise, participants will discuss these answers and provide their thoughts. Technology developer observes but does not participate unless the facilitator brings them in to address a technical or factual point. Critical friend observes and supports as required (guided by facilitator).
  - This task should be guided by the facilitator, asking questions to prompt discussion about the scenarios, such as:
    * What groups or individuals would be affected by this project?
    * What potential risks, biases or harms do you foresee occurring from use/deployment of this algorithm?
    * Who will benefit most from this project and how?
    * Who could be harmed if this system fails?
    * What benefits will this project have for patients and the NHS?
    * Of the impacts identified, what would be potential causes for this impact?
    * What solutions or measures would they like to see adopted to reduce the risks of harm?
- A session to group themes in the impacts into the template and prioritise them (25 mins):
  - As one group or in two breakout groups, the participants consider any common themes in their identified impacts

and group them together. (e.g. multiple impacts might relate to discrimination, or to reduced quality of care.) Technology developer observes but does not participate unless the facilitator brings them in to address a technical or factual point. Critical friend observes and supports as required (guided by facilitator). The facilitator should use a (virtual) whiteboard to fill out the template.

- The participants then prioritise the themes and specific impacts by dot-voting[159] they should be guided by the facilitator, asking questions such as:
    * Of the impacts identified, which are likely to cause high and very high risk of harm?
    * Of the impacts identified, which would you consider to be the most important? How consequential is this harm for the wellbeing of which stakeholders?
    * Of the impacts identified, which are urgent? How immediate would the threat of this impact be?
    * Of the impacts identified, which will be the most difficult to mitigate?
    * Of the impacts identified, which will be the most difficult to detect, given the current design?

— Participants take a break while the technology developer reviews the templates of identified impacts. (10 mins).

— A session with facilitated discussion so the technology developer can ask questions back to the participants, to clarify the impacts identified and further flesh out impacts (and provide overview of next steps: how will the developers be confronting/responding to the impacts identified in the development process, and what that could look like (i.e. model retraining) as well as updates to the AIA (25 mins).

— Wrap up and close (5 mins).

159 Dotmocracy. How to use dot voting effectively. Available at: https://dotmocracy.org/dot-voting/

Equipment and tools required:

- Accessible room/venue and or online video conferencing tool (e.g. Zoom – with provisions for visually or hearing impaired and neurodiverse people as required).
- Slide deck for introductions and presentations (with accessibility provisions). Ideally shared beforehand.
- (Virtual) whiteboard/flipchart and post-its.
- Impacts template prepared and ready to be filled in.

Outputs:

- Filled out template that lists impacts and priority of them (based on dot-votes).
- Technology developers should take notes to deepen their understanding of their algorithms' potential impacts and perspectives of the public.

Applicant teams devise ideas to address or mitigate impacts

Following the workshop the applicant team should consider mitigations, solutions or plans to address the impacts identified during the workshop, and update the first iteration of the AIA in light of the participants' deliberations.

This analysis should be worked back into the template as part of the synthesis exercise.

Panel reviews mitigation plans

- The applicant team's plans to address impacts are shared with the panel participants, who review them and share any feedback or reflections. This can be done asynchronously via email, over a period of two-to-four weeks. Assuming all participants are supported to engage: accessible materials, access to the web, etc.

- Panel can make a judgement call on permissibility of the algorithm based on the developer's updated proposals, for the DAC to consider.

- The panels' comments are used by the NHS AI Lab NMIP DAC to support their assessment of the overall AIA.

Resourcing and costs

Staff resources:

- **Panel coordinator:** a member of NHS AI Lab staff to coordinate and run the panel process, and to ensure it is genuinely and fully embedded in the wider AIA and considered by the DAC. This individual should have experience and knowledge of: public engagement, public and stakeholder management, workshop design and working with those with complex health conditions. This could be a stand-alone role, or a part of another person's role, as long as they are given sufficient time and capacity to run the process well.

- **Facilitators:** additional NHS staff, partner organisation staff or freelancers to support workshop facilitation as required.

- **Technology developers and critical friends** who participate in the impact identification workshops should be able to do this as part of their professional roles, so would not typically require remuneration.

Panel participant cost estimates:

Recruitment: there are two approaches to recruiting participants:

1. Panel coordinator works with NHS trusts and community networks to directly recruit panel members (e.g. by sending email invitations). The coordinator would need to ensure they reach a wide population, so that the final panel is sufficiently diverse. This option has no additional cost, but is significantly time-intensive, and would require the co-ordinator to have sufficient capacity, support and skills to do so.

2. Commission a research participant recruitment agency to source panel members and manage communication and remuneration.

Costs:

- Recruitment cost estimated at £100 per person for 30 people: £3,000.
- Administrative cost estimate for communications and remunerating participants: £2,000 – £4,000.
- Remuneration: participants should be remunerated at industry best practice rates of £20–£30 per hour of activity.
- Assuming 30 participants who each participates in the induction (two hours) and a single 'batch' of NMIP applications, for example five workshops (15 hours) and reviews five mitigation plans (six hours), estimated remuneration costs would be: £13,800 - £20,700.

Miscellaneous costs to consider:

- If hosting workshops virtually: cost for any software and accessibility support such as interactive whiteboards, video-conferencing software, live captioning, etc.
- If hosting workshops in-person: venue hire, catering, travel etc.
- Materials: printing, design of templates, information packs etc. as required.

Bibliography

Ada Lovelace Institute and DataKind UK. (2020). *Examining the black box: tools for assessing AI systems*. Ada Lovelace Institute. Available at: https://www.adalovelaceinstitute.org/report/examining-the-black-box-tools-for-assessing-algorithmic-systems/

Ada Lovelace Institute. (2021). *Participatory data stewardship*. Available at: https://www.adalovelaceinstitute.org/report/participatory-data-stewardship/

Ada Lovelace Institute. (2021). *The data divide*. Available at: https://www.adalovelaceinstitute.org/wp-content/uploads/2021/03/The-data-divide_25March_final-1.pdf

Ada Lovelace Institute, AI Now Institute, Open Government Partnership. (2021). *Algorithmic accountability for the public sector*. Open Government Partnership. Available at: https://www.opengovpartnership.org/documents/algorithmic-accountability-public-sector/

Adelle, C. and Weiland, S. (2012). 'Policy assessment: the state of the art'. *Impact Assessment and Project Appraisal* 30.1, pp. 25-33 Available at: https://www.tandfonline.com/doi/full/10.1080/14615517.2012.663256

Amsterdam Algorithm Register Beta. *What is the algorithm register?* Available at: https://algoritmeregister.amsterdam.nl/

Angwin, J., Larson, J., Mattu, S. and Kirchnir, L. (2016). 'Machine bias'. *ProPublica*. Available at: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Antun, V., Renna, F., Poon, C., Adcock, B., Hansen, A. C. (2020). 'On instabilities of deep learning in image reconstruction and the potential costs of AI'. *Proceedings of the National Academy of Sciences of the United States of America*, p. 117, 48 [online] Available at: https://www.pnas.org/content/117/48/30088

Article One. *Challenge: From 2017 to 2018, Microsoft partnered with Article One to conduct the first-ever Human Rights Impact Assessment (HRIA) of the human rights risks and opportunities related to artificial intelligence (AI)*. Available at: https://www.articleoneadvisors.com/case-studies-microsoft

Balayn, A and Gürses, S. (2021). *Beyond debiasing: regulating AI and its inequalities*. European Digital Rights. Available at: https://edri.org/wp-content/uploads/2021/09/EDRi_Beyond-Debiasing-Report_Online.pdf

Banerje, I et al. (2021). 'Reading race: AI recognises patient's racial identity in medical images'. *Computer Vision and Pattern Recognition*. Available at: https://arxiv.org/abs/2107.10356

Barocas, S. and Selbst, A.D. (2016). 'Big data's disparate impact'. *California Law Review*, 104, pp. 671- 732. [online] Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2477899

Beede, E., Elliott Baylor, E., Hersch, F., Iurchenko, A., Wilcox, L., Ruamviboonsuk, P. and Vardoulakis, L. (2020). 'A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy'. In: *CHI Conference on Human Factors in Computing Systems* (CHI '20), April 25-30, 2020, Honolulu, HI, USA. ACM, New York, NY, USA. Available at: https://dl.acm.org/doi/fullHtml/10.1145/3313831.3376718

Binns, R. (2018). 'Algorithmic accountability and public reason'. *Philosophy & Technology*, 31, pp.543-556. [online] Available at: https://link.springer.com/article/10.1007/s13347-017-0263-5

Bohr, A. and Memarzadeh. K. (2020). 'The rise of artificial intelligence in healthcare applications'. *Artificial Intelligence in Healthcare*, pp.25-60. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7325854/

Bovens, M. (2006). Analysing and assessing public accountability. A conceptual framework. *European Governance Papers* (EUROGOV) No. C-06-01. Available at: https://www.ihs.ac.at/publications/lib/ep7.pdf

Boyd, K.L. (2021). 'Datasheets for datasets help ML engineers notice and understand ethical issues in training data'. *Proceedings of the ACM on Human-Computer Interaction*, 5, 438. [online] Available at: http://karenboyd.org/blog/wp-content/uploads/2021/09/Datasheets_Help_CSCW-5.pdf

Buolamwini, J. and Gebru, T. (2018). 'Gender shades: intersectional accuracy disparities in commercial gender classification'. *Conference on Fairness, Accountability and Transparency*, pp.1-15.[online] Available at: https://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf

Chakradhar, S. (2019). 'Widely used algorithm in hospitals is biased, study finds'. *STAT*. Available at: https://www.statnews.com/2019/10/24/widely-used-algorithm-hospitals-racial-bias/

Cheah, P.Y. and Piasecki, J. (2020). Data access committees. *BMC Medical Ethics*, 21, 12 [online] Available at: https://link.springer.com/article/10.1186/s12910-020-0453-z

City of Helsinki AI register. *What is the AI register?* Available at: https://ai.hel.fi/

Congress.Gov. (2019). *H.R.2231 – Algorithmic Accountability Act of 2019*. Available at: https://www.congress.gov/bill/116th-congress/house-bill/2231#:~:text=Introduced%20in%20House%20(04%2F10%2F2019)&text=This%20bill%20requires%20specified%20commercial,artificial%20intelligence%20or%20machine%20learning

Data Smart Schools. (2021). *Deb Raji on what 'algorithmic bias' is (...and what it is not)*. Available at: https://data-smart-schools.net/2021/04/02/deb-raji-on-what-algorithmic-bias-is-and-what-it-is-not/

Davenport, T. and Kalakota, R. (2019). 'The potential for artificial intelligence in healthcare'. *Future Healthcare Journal*, 6,2, pp.94-98. [online] Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6616181/

Department of Health and Social Care. (2018). *The future of healthcare: our vision for digital, data and technology in health and care*. UK Government. Available at: https://www.gov.uk/government/publications/the-future-of-healthcare-our-vision-for-digital-data-and-technology-in-health-and-care/the-future-of-healthcare-our-vision-for-digital-data-and-technology-in-health-and-care

Department of Health and Social Care. (2020). *A guide to good practice for digital and data-driven health technologies. UK Government*. Available at: https://www.gov.uk/government/publications/code-of-conduct-for-data-driven-health-and-care-technology/initial-code-of-conduct-for-data-driven-health-and-care-technology

Esteva, A., Chou, K., Yeung, S., Naik, N., Madani, A., Mottaghi, A., Liu, Y., Topol, E., Dean, J., and Socher, R. (2021). 'Deep learning-enabled medical computer vision'. *npj Digital Medicine*, pp.1-9 [online]. Available at: https://www.nature.com/articles/s41746-020-00376-2

European Commission. (2020). *Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment*. Available at: https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment

Floridi, L. (2019). 'Translating principles into practices of digital ethics: five risks of being unethical'. *Philosophy & technology*, 32, pp.185-193 [online] Available at: https://link.springer.com/article/10.1007/s13347-019-00354-x

Frakt, A. (2020) 'Bad medicine: the harm that comes from racism'. *The New York Times*. [online] Available at: https://www.nytimes.com/2020/01/13/upshot/bad-medicine-the-harm-that-comes-from-racism.html

Freeman, K., Geppert, J., Stinton, C., Todkill, D., Johnson, S., Clarke, A. and Taylor-Phillips, S. (2021). 'Use of artificial intelligence for image analysis in breast cancer screening programmes: systematic review of test accuracy'. *British Medical Journal 2021*, 374 [online] Available at: https://pubmed.ncbi.nlm.nih.gov/34470740/

Gastil, J. (ed.) (2005). *The deliberative democracy handbook: strategies for effective civic engagement in the twenty-first century*. 1. ed., 1. impr. Hoboken, N.J: Wiley.

Gebru, T., Mogenstern, J., Vecchione, B., Wortman Vaughan, J., Wallach, H., Daumé III, H. and Crawford, K. (2018). Datasheets for datasets. *ArXiv* [online] Available at: https://arxiv.org/abs/1803.09010

Gichoya, J.W. et al (2021). 'Reading race: AI recognises patient's racial identity in medical images'. *arXiv*. Available at: https://arxiv.org/abs/2107.10356

Government of Canada. (2020). *Algorithmic impact assessment tool*. Available at: https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html

Government of Canada. (2020). *Directive on Automated Decision-Making*. Available at: https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592

Government of Canada. (2021). *Algorithmic Impact Assessment – ArriveCAN Proof of Vaccination Recognition*. Available at: https://open.canada.ca/data/en/dataset/afc17416-3781-422d-a4a9-cc55e3a053c8

Government of Canada. (2021). *Algorithmic Impact Assessment – ATIP Online Request Service*. Available at: https://open.canada.ca/data/en/dataset/cea9985f-5e0f-425e-9b7e-e1d122272c56

Government of Canada. (2021). *Algorithmic Impact Assessment – Spouse or Common-Law Partner in Canada Advanced Analytics Pilot*. Available at: https://open.canada.ca/data/en/dataset/d41f9ec2-bf01-4b2a-bd8d-1b3a8424f534

Hildebrandt, M. (2012). 'The dawn of a critical transparency right for the profiling era'. *Digital Enlightment Yearbook*, pp.41-56 [online] Available at: https://repository.ubn.ru.nl/handle/2066/94126

Information Commissioner's Office (ICO). (2019). *An overview of the auditing framework for artificial intelligence and its core components*. Available at: https://ico.org.uk/about-the-ico/news-and-events/ai-blog-an-overview-of-the-auditing-framework-for-artificial-intelligence-and-its-core-components/

Information Commissioner's Office (ICO). (2020). *Data protection impact assessments*. Available at: https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/accountability-and-governance/data-protection-impact-assessments/

ISO. (2021). *New ISO standards for medical devices*. Available at: https://www.iso.org/news/ref2534.html

International Organization for Standardization (ISO). *14971:2019 Medical devices – application of risk management to medical devices*. Available at: https://www.iso.org/obp/ui/#iso:std:iso:14971:ed-3:v1:en

Johnson, K. (2021). 'The movement to hold AI accountable gains more steam'. *Ars Technica*. Available at: https://arstechnica.com/tech-policy/2021/12/the-movement-to-hold-ai-accountable-gains-more-steam/3/

Kaminski, M. (2020). 'Understanding transparency in algorithmic accountability'. *Cambridge Handbook of the Law of Algorithms*, e.d. Woodrow Barfield. Cambridge: Cambridge University Press [online] Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3622657

Kaminski, M.E. and Malgieri, G. (2020). 'Algorithmic impact assessments under the GDPR: producing multi-layered explanations'. *International Data Privacy Law*, 11,2, pp.125-144. Available at: https://doi.org/10.1993/idpl/ipaa020

Karlin, M. and Corriveau, N. (2018). 'The Government of Canada's Algorithmic Impact Assessment: Take Two'. *Supergovernance*. Available at: https://medium.com/@supergovernance/the-government-of-canadas-algorithmic-impact-assessment-take-two-8a22a87acf6f

Katell, M., Young, M., Dailey, D., Herman, B., Guetler, V., Tam, A., Bintz, C., Raz, D. and Krafft, P. M. (2020). 'Toward situated interventions for algorithmic equity: lessons from the field'. *Conference on Fairness, Accountability, and Transparency* pp.44-45 [online] ACM: Barcelona. Available at: https://dl.acm.org/doi/abs/10.1145/3351095.3372874

Kelly, C.J., Karthikesalingam, A., Suleyman, M., Corrado, G. and King. D. (2019). 'Key challenges for delivering clinical impact with artificial intelligence' *BMC Medicine*. 29 October, 17: 195. Available at: https://bmcmedicine.biomedcentral.com/articles/10.1186/s12916-019-1426-2

Knowles, B. and Richards, J. (2021). 'The sanction of authority: promoting public trust in AI'. *Computers and Society*. Available at: https://arxiv.org/abs/2102.04221

Latonero, M. and Agarwal, A. (2021). *Human rights impact assessments for AI: learning from Facebook's failure in Myanmar*. CARR Center for Human Rights Policy Harvard Kennedy School. Available at: https://carrcenter.hks.harvard.edu/files/cchr/files/210318-facebook-failure-in-myanmar.pdf

Legislation.gov.uk. (2018). *Data Protection Act 2018*. Available at: https://www.legislation.gov.uk/ukpga/2018/12/contents/enacted

Leslie, D. (2019). *Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector*. The Alan Turing Institute. Available at: www.turing.ac.uk/sites/default/files/2019-06/understanding_artificial_intelligence_ethics_and_safety.pdf

Loi, M. in collaboration with Matzener, A., Muller, A. and Spielkamp, M. (2021). *Automated decision-making systems in the public sector. An impact assessment tool for public authorities*. Algorithm Watch. Available at: https://algorithmwatch.org/en/wp-content/uploads/2021/06/ADMS-in-the-Public-Sector-Impact-Assessment-Tool-AlgorithmWatch-June-2021.pdf

Madaio, M., Stark, L., Wortman Vaughan, J., Wallach, H. (2020). 'Co-designing checklists to understand organisational challenges and opportunities around fairness in AI'. *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp.1-14 [online] Available at: https://dl.acm.org/doi/abs/10.1145/3313831.3376445

Margetts, H. and Dorobantu, C. (2019). 'Rethink government with AI'. *Nature*. Available at: https://www.nature.com/articles/d41586-019-01099-5

Medicines and Healthcare products Regulatory Agency. (2020). *Medical devices: conformity assessment and the UKCA mark. UK Government*. Available at: https://www.gov.uk/guidance/medical-devices-conformity-assessment-and-the-ukca-mark

Medicines and Healthcare products Regulatory Agency (2020). *Regulating medical devices in the UK*. UK Government. Available at: https://www.gov.uk/guidance/regulating-medical-devices-in-the-uk

Metcalf, J., Moss, E., Watkins, E.A., Ranjit, S. and Elish, M.C. (2021). 'Algorithmic impact assessments and accountability: the co-construction of impacts'. *Conference on Fairness Accountability, and Transparency* [online] Available at: https://dl.acm.org/doi/pdf/10.1145/3442188.3445935

Miller, C. (2015). 'When algorithms discriminate'. *The New York Times*. Available at: https://www.nytimes.com/2015/07/10/upshot/when-algorithms-discriminate.html

Moss, E., Watkins, E.A., Singh, R., Elish, M.C. and Metcalf, J. (2021). *Assembling accountability: algorithmic impact assessment for the public interest*. Data & Society. Available at: https://datasociety.net/library/assembling-accountability-algorithmic-impact-assessment-for-the-public-interest/

NHS. (2019). *The NHS Long Term Plan*. Available at: https://www.longtermplan.nhs.uk/wp-content/uploads/2019/08/nhs-long-term-plan-version-1.2.pdf

NHS Digital. *How NHS Digital makes decisions about data access.* Available at: https://digital.nhs.uk/services/data-access-request-service-dars/how-nhs-digital-makes-decisions-about-data-access

NHS Health Research Authority. *What is public involvement in research?* Available at: https://www.hra.nhs.uk/planning-and-improving-research/best-practice/public-involvement/

NHSX. (2019). *Artificial Intelligence: how to get it right*. Available at: https://www.nhsx.nhs.uk/media/documents/NHSX_AI_report.pdf

NHS AI Lab. *AI in imaging.* Available at: https://www.nhsx.nhs.uk/ai-lab/ai-lab-programmes/ai-in-imaging/

NHSX. *How NHS and care data is protected*. Available at: https://www.nhsx.nhs.uk/key-tools-and-info/data-saves-lives/how-nhs-and-care-data-is-protected

NHS AI Lab. *National Medical Imaging Platform (NMIP)*. Available at: https://www.nhsx.nhs.uk/ai-lab/ai-lab-programmes/ai-in-imaging/national-medical-imaging-platform-nmip/

NHS AI Lab. *The AI Ethics Initiative: Embedding ethical approaches to AI in health and care*. Available at: https://www.nhsx.nhs.uk/ai-lab/ai-lab-programmes/ethics/

NHS AI Lab. *The NHS AI Lab: accelerating the safe adoption of AI in health and care*. Available at: https://www.nhsx.nhs.uk/ai-lab/

NHSX. (2021). *What Good Looks Like framework*. Available at: https://www.nhsx.nhs.uk/digitise-connect-transform/what-good-looks-like/what-good-looks-like-publication/

NICE. *Evidence standards framework for digital health technologies.* Available at: https://www.nice.org.uk/about/what-we-do/our-programmes/evidence-standards-framework-for-digital-health-technologies

Noble, S.U. (2018). *Algorithms of oppression: how search engines reinforce racism*. NYU Press

Raji, D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D. and Barnes, P. (2020). 'Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing'. *Conference on Fairness, Accountability, and Transparency*, pp.33–44. Barcelona: ACM. Available at: https://doi.org/10.1145/3351095.3372873

Reform UK. *Data-driven healthcare: Regulation & regulators*. Available at: https://reform.uk/research/data-driven-healthcare-regulation-regulators

Reisman, D., Schultz, J., Crawford, K. and Whittaker, M. (2018). *Algorithmic impact assessments: a practical framework for public agency accountability.* AI Now Institute. Available at: https://ainowinstitute.org/aiareport2018.pdf

Royal College of Radiologists. *Policy priorities: Artificial Intelligence*. Available at: https://www.rcr.ac.uk/press-and-policy/policy-priorities/artificial-intelligence

Scassa, T. (2020). *Administrative law and the governance of automated decision-making: a critical look at Canada's Directive on Automated Decision-Making*. Forthcoming, University of British Columbia Law Review. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3722192

Selbst, A.D. (2017). 'Disparate impact in big data policing'. *52 Georgia Law Review* 109, pp.109-195. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2819182

Selbst, A.D. (2018). 'The intuitive appeal of explainable machines'. Fordham Law Review 1085. [online] Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3126971

Selbst, A.D. (2021). 'An institutional view of algorithmic impact assessments', *Harvard Journal of Law & Technology* (forthcoming). Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3867634

Sendak, M., Gao, M., Brajer, N. and Balu, S. (2020). 'Presenting machine learning model information to clinical end users with model facts labels'. *npj Digital Medicine*, 3,41, p1-4. [online] Available at: https://www.nature.com/articles/s41746-020-0253-3

Sendak, M. et al. (2020). '"The human body is a black box": supporting clinical decision-making with deep learning'. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*. Association for Computing Machinery, New York, NY, USA, pp. 99–109. Available at: https://doi.org/10.1145/3351095.3372827

Seyyad-Kalantari, L., Zhang, H., McDermott, M., Chen, I. Y., Ghassemi, M. (2021). 'Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in underserved patient populations'. *Nature Medicine*, 27, pp. 2176-2182. Available at: https://www.nature.com/articles/s41591-021-01595-0

Singh, J, Cobbe, J and Norval, C. (2019). 'Decision provenance: harnessing data flow for accountable systems'. *IEEE Access*, 7, pp. 6592-6574 [online]. Available at: https://arxiv.org/abs/1804.05741

Sloane, M., Moss, E., Awomolo, O., & Forlano, L. (2020). 'Participation is not a design fix for machine learning'. *arXiv*. [online] Available at: https://arxiv.org/abs/2007.02423

Steel, D., Bolduc, N., Jenei, K. and Burgess, M. (2020). 'Rethinking representation and diversity in deliberative minipublics'. *Journal of Deliberative Democracy*, 16(1), pp.46-57. Available at: https://delibdemjournal.org/article/id/626/

Thorogood A., and Knoppers, B.M. (2017). 'Can research ethics committees enable clinical trial data sharing?'. *Ethics, Medicine and Public Health*, 3,1, pp.56-63.[online] Available at: https://www.sciencedirect.com/science/article/abs/pii/S2352552517300129

Topol, E. (2019). 'High performance medicine: the convergence of human and artificial intelligence'. *Nature Medicine*, 25, pp.45-56. [online] Available at: https://www.nature.com/articles/s41591-018-0300-7

UCL. (2020). *UCLH Covid-19 data access committee set up*. Available at: https://www.ucl.ac.uk/joint-research-office/news/2020/jun/uclh-covid-19-data-access-committee-set

University Hospital Southampton. *Involving patients and the public*. Available at: https://www.uhs.nhs.uk/ClinicalResearchinSouthampton/For-researchers/PPI-in-your-research.aspx

Wen, D., Khan, S., Ji Xu, A., Ibrahim, H., Smith, L., Caballero, J., Zepeda, L., de Blas Perez, C., Denniston, A., Lui, X. and Martin, R. (2021). 'Characteristics of publicly available skin cancer image datasets: a systematic review'. *The Lancet: Digital Health* [online]. Available at: https://www.thelancet.com/journals/landig/article/PIIS2589-7500(21)00252-1/fulltext

Wieringa, M. (2020). 'What to account for when accounting for algorithms: a systematic literature review on algorithmic accountability'. *Conference on Fairness, Accountability, and Transparency*, pp.1-18 [online] Barcelona: ACM. Available at: https://dl.acm.org/doi/10.1145/3351095.3372833

About the Ada Lovelace Institute

The Ada Lovelace Institute (Ada) is an independent research institute with a mission to make data and AI work for people and society.

We are working to create a shared vision of a world where AI and data are mobilised for good, to ensure that technology improves people's lives. We take a sociotechnical, evidence-based approach and use deliberative methods to convene and centre diverse voices. We do this to identify the ways that data and AI reorder power in society, and to highlight tensions between emerging technologies and societal benefit.

Ada was established by the Nuffield Foundation in early 2018, in collaboration with the Alan Turing Institute, the Royal Society, the British Academy, the Royal Statistical Society, the Wellcome Trust, Luminate, techUK and the Nuffield Council on Bioethics.

We are funded by the Nuffield Foundation, an independent charitable trust with a mission to advance social well-being. The Foundation funds research that informs social policy, primarily in education, welfare and justice. It also provides opportunities for young people to develop skills and confidence in STEM and research. In addition to the Ada Lovelace Institute, the Foundation is also the founder and co-funder of the Nuffield Council on Bioethics and the Nuffield Family Justice Observatory.

Find out more:

Website: Adalovelaceinstitute.org
Twitter: @AdaLovelaceInst
Email: hello@adalovelaceinstitute.org

designbysoapbox.com

Preferred citation: Ada Lovelace Institute. (2022). *Algorithmic impact assessment: a case study in healthcare*. Available at: https://www.adalovelaceinstitute.org/report/algorithmic-impact-asssessment-case-study-healthcare

ISBN: 978-1-8382567-9-1