EXPLAINABILITY IN DATA-DRIVEN HEALTH AND CARE TECHNOLOGY

Guidance for Principle 7 of the NHS Code of Conduct for data-driven health and care technology
Future Advocacy is a think tank and consultancy working on some of the greatest challenges that humanity faces in the 21st century. We advocate for smart, forward-thinking policies that will allow us to capitalise on the opportunities and mitigate the risks presented by artificial intelligence.

futureadvocacy.org

Acknowledgements
Written and researched by Olly Buston, Nika Strukelj, and Dr Matthew Fenech.

© Future Advocacy 2019. All rights reserved.
CONTENTS

Introduction 4-6

General processes 7-11
  i) Stakeholder analysis
  ii) Openness

Specific processes 12-19
  i) Assess data issues and identify algorithm(s)
  ii) Prove algorithm(s) is/are effective
  iii) Consider interaction with wider healthcare system
  iv) Comply with ‘right to explanation’
  v) Explain how acceptable use of algorithm determined

Appendices 20-32
  i) Glossary of terms
  ii) Overlap with information required by other frameworks
  iii) List of contributors
  iv) Methodology by which patient/public contributors were recruited
INTRODUCTION

The Code of Conduct for data-driven health and care technologies is intended to go one step beyond high-level ‘ethical principles’ to set out “the behaviours [expected] from those developing, deploying and using data-driven technologies”. Nevertheless, the healthtech community are now at the stage where we need to move from aspirations for ‘good’ behaviour, to practical and practicable guidance on how to operationalise the incorporation of these principles in their everyday work and in the tools that they are developing.

This guidance is especially needed in the context of an increasingly complex governance environment to ensure that the process for demonstrating ‘good practice’ is as simple and clear as possible and responsive to changes in technology and regulation. This latter point will become increasingly important as we move from ‘soft’ governance (principles, behaviours, guidance) to ‘hard’ governance (regulation and legal requirements).

Whilst this process might seem slow, or unnecessarily arduous, the point has been made that, given their far-reaching socioeconomic impacts and their capacity to be applied rapidly and at scale, the impact of data-driven technologies on healthcare “is more akin to that of automobiles or personal computers” on society, than of medicines and medical devices.² Traditional models of assessment of safety and effectiveness of medical products are insufficient to cope with the challenges posed by these technologies, particularly with respect to their potential to disrupt whole healthcare systems, rather than impact the health of individuals. For this reason any guidance to developers on adhering to the various principles and frameworks must support the appreciation of how the use of these data-driven tools will impact the wider healthcare system, which is not something that is always clear in legislative settings.

In view of these issues, NHSX is producing (in collaboration with the healthtech ecosystem) a toolkit for ‘Responsibly Applied AI in Health and Care.’ This will be made up of various resources to enable developers to demonstrate best practice as outlined across the whole Code of Conduct. This toolkit will be available on GitHub and on the NHSX website, and will comprise (as a minimum):

- Landscape mapping, horizon scanning, and case study collation, being carried out in collaboration with various organisations including the Academic Health Sciences Networks (AHSN) Network, the Royal Society of Arts, Manufactures and Commerce (RSA), and the Association of Medical Research Charities (AMRC)
- Results of engagement exercises to understand stakeholder wants and concerns around AI in health
- Maps of existing health-related datasets
- A pathway for ‘regulation as a service’

Moreover, NHSX will accept submissions from developers for their accounts and results of undertaking the processes described in the guidance below, for publication on the NHSX website (see Section B (ii): Openness).

---

¹ See the Glossary for a definition of ‘developers’ for the purposes of this document.

This particular document aims to provide those developing, deploying, and using data-driven technologies with resources to support a structured approach to demonstrating adherence with Principle 7 of the Code of Conduct. We anticipate that those charged with procuring, deploying, and overseeing the use of these tools, as well as the end-users themselves, will find this guidance helpful in allowing them to assess the merits of a particular data-driven technology.

**Principle 7: Show what type of algorithm is being developed or deployed, the ethical examination of how the data is used, how its performance will be validated and how it will be integrated into health and care provision**

Consider how the introduction of AI will change relationships in health and care provision, and the implications of these changes for responsibility and liability. Use current best practice on how to explain algorithms to those taking actions based on their outputs.

When building an algorithm, be it a stand-alone product or integrated within a system, show it clearly and be transparent of the learning methodology (if any) that the algorithm is using. Undertake ethical examination of data use specific to this use-case. Achieving transparency of algorithms that have a higher potential for harm or unintended decision-making, can ensure the rights of the data subject as set out in the Data Protection Act 2018 are met, to build trust in users and enable better adoption and uptake.

Work collaboratively with partners, specify the context for the algorithm, specify potential alternative contexts and be transparent on whether the model is based on active, supervised or unsupervised learning. Show in a clear and transparent specification:

- the functionality of the algorithm
- the strengths and limitations of the algorithm (as far as they are known)
- its learning methodology
- whether it is ready for deployment or still in training
- how the decision has been made on the acceptable use of the algorithm in the context it is being used (for example, is there a committee, evidence or equivalent that has contributed to this decision?)
- the potential resource implications

This specification and transparency in development will build trust in incorporating machine-led decision-making into clinical care.

**Box 1: Principle 7 from the Code of Conduct for data-driven health and care technology (as at 19th February 2019)**

**Format**

This guidance takes the form of a set of processes that we encourage developers to undertake. The processes are divided into two: recommendations for general processes that apply across all aspects of Principle 7, and recommendations for specific processes that apply to certain subsections of Principle 7 (Figure 1). Many of these processes require information that developers may well have gathered as part of other compliance and regulatory submission processes. Appendix D, Section (iii) provides a cross-referencing tool to highlight this and to help developers with avoiding duplication of work.

The overall process is summarised in Figure 1.
Figure 1: A schematic outlining the different components of this guidance.

A note about proportionality

Various other guidelines and frameworks have adopted a risk stratification-based approach. A cumulative set of requirements is usually outlined, with a higher risk rating (indicating a greater potential of harm to users) requiring more processes or more robust evidence.\(^3\)\(^4\) We believe that the processes outlined in this document are ones that developers should be undertaking regardless of the type of digital health tool they are creating, and they should therefore seek to address all the requirements. It should be noted, however, that it is entirely reasonable that not all tools will require the same level of detail. We anticipate that the comprehensive assessment of more complex tools with higher potential risk will lead to the provision of more information than that of simpler tools with lower risk. Moreover, this document is not intended to be definitive and final as the field is developing and the full suite of benefits and risks may not be fully understood. We will, therefore, keep this guidance (and the other elements of the toolkit) under regular review, with any updates and changes communicated to the relevant stakeholders, and we encourage regular feedback.

---

\(^3\) The NICE Evidence Standards Framework for Digital Health Technologies, for example, provides a ‘Functional Classification’ of the tools, which is used to guide readers as to the appropriate type of evidence required by the framework. See https://www.nice.org.uk/Media/Default/About/what-we-do/our-programmes/evidence-standards-framework/digital-evidence-standards-framework.pdf

GENERAL PROCESSES

i) Stakeholder analysis

- Undertaking a robust and inclusive process of stakeholder analysis will help highlight and preserve relationships of importance in healthcare, ensuring that the various players in the diverse relationships making up a healthcare system are identified and involved in the development process.\(^5,^6,^7\)
- This approach fits with the other principles in the Code of Conduct, which emphasise the importance of being guided by the needs of the user.
- The results of the stakeholder analysis are relevant to multiple subsections of Principle 7, and will be drawn upon when undertaking work described by various specific requirements in Section C of the guidance.

For these reasons, we ask that:

- Developers **undertake a process of stakeholder analysis.** This should go beyond simply identifying direct and indirect stakeholders, but should provide a foundation for (i) the process by which these stakeholders are meaningfully considered in the development process, and (ii) a deeper understanding of the wider cultural context (be it in the healthcare system or in wider society) **in which the data-driven tool will be embedded** (relevant to Section C (iii) below).\(^8\)
- Once stakeholders are identified, **their requirements and concerns (that is, both positively-valued and negatively-valued beliefs) be considered through the use of value and consequence matrices** (Table 1). Essentially, we ask that developers take each of these domains in turn and try to understand (a) what each stakeholder might want or be concerned about in the context of this domain [the value matrix], and (b) the potential impact the data-driven tool may have on these requirements and concerns [the consequence matrix]. Such a process allows developers to think about all types of best- and worst-case scenarios at an early stage in development.
- The process of stakeholder analysis be **repeated at regular intervals**, scheduled in advance to cover.

Relevant resources


NHS Improvement, “Online library of Quality, Service Improvement and Redesign tools: Stakeholder analysis”, available at
  ○ https://improvement.nhs.uk/documents/2169/stakeholder-analysis.pdf (contains useful guidance on identifying and prioritising stakeholders in the healthcare context)


a) Value matrix

<table>
<thead>
<tr>
<th>Direct and Indirect Stakeholders</th>
<th>Respect the dignity of individuals as persons</th>
<th>Care for the wellbeing of each and all</th>
<th>Connect with each other sensitively, openly, and inclusively</th>
<th>Protect the priorities of justice, social values and the public interest</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stakeholder 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stakeholder 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stakeholder 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stakeholder 4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

b) Consequence matrix

<table>
<thead>
<tr>
<th>Direct and Indirect Stakeholders</th>
<th>Respect the dignity of individuals as persons</th>
<th>Care for the wellbeing of each and all</th>
<th>Connect with each other sensitively, openly, and inclusively</th>
<th>Protect the priorities of justice, social values and the public interest</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stakeholder 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stakeholder 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Guidance for Principle 7 of the NHS Code of Conduct for data-driven health and care technology

<table>
<thead>
<tr>
<th>Stakeholder 3</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Stakeholder 4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Value and consequence matrices. a) Once stakeholders are identified by developers, their concerns, wishes, values, and other views can be considered in the context of the SUM principles developed by the Alan Turing Institute highlighted above and detailed below b) Once these views and concerns are understood, the developer should determine how deploying their proposed data-driven technology could impact these, with a judgement applied as to whether this is a negative or positive impact. The use of colour coding (e.g. traffic light system) could then provide an at-a-glance view of the areas of greatest potential benefit and greatest concern.
<table>
<thead>
<tr>
<th>SUM Value</th>
<th>Details</th>
</tr>
</thead>
</table>
| **Respect** the dignity of individuals as persons | ▶ Ensure abilities of individuals to make free and informed decisions about their own lives  
▶ Safeguard their autonomy, their power to express themselves, and their right to be heard  
▶ Secure their capacities to make informed and well-considered life choices  
▶ Support their abilities to fully develop themselves and to pursue their passions and talents according to their own freely determined life plans |
| [This is underpinned by the values of: (1) autonomy and authority of persons; (2) self-realisation and flourishing of persons] |                                                                                                                                                                                                     |
| **Care** for the wellbeing of each and all | ▶ Design and deploy data-driven technology to cultivate the welfare of all stakeholders whose interests are affected by their use  
▶ Do no harm with these technologies and minimise the risks of their misuse or abuse  
▶ Prioritise the safety and the mental and physical integrity of people when scanning horizons of technological possibility, convening of, and deploying AI applications |
| [This is underpinned by the values of: (1) beneficence, safety and non-harm; (2) stewardship of the biosphere (the healthcare ecosystem)] |                                                                                                                                                                                                     |
| **Connect** with each other sincerely, openly, and inclusively | ▶ Safeguard the integrity of interpersonal dialogue and connection  
▶ Protect the human interaction as a key for trust and empathy  
▶ Use technology to foster this capacity to connect so as to reinforce responsibility and mutual understanding |
| [This is underpinned by the values of: (1) integrity of the interpersonal relationship; (2) participation-based innovation and stakeholder inclusion] |                                                                                                                                                                                                     |
| **Protect** the priorities of justice, social values and the public interest | ▶ Treat all individuals equally and protect social equality  
▶ Use digital technologies as a bulwark for the protection of fair and equal treatment under the law  
▶ Prioritise social welfare, public interest, and the consideration of the social and ethical impacts of innovation in determining the legitimacy and desirability of AI technologies  
▶ Use AI to empower and to advance the interests and well-being of as many individuals as possible |
| [This is underpinned by the values of: (1) Justice; (2) Prioritisation of the public interest] |                                                                                                                                                                                                     |
and the common good]

Table 2: SUM values developed by the Alan Turing Institute to ensure that algorithmic systems are fair, accountable, sustainable and transparent

ii) Openness

Openness will:

- Provide developers with additional opportunities to influence the development of regulation of data-driven technologies in healthcare, making sure that regulation becomes a more collaborative process between industry and regulators. This is crucial to increasing the acceptability of regulatory frameworks in data-driven technologies.\(^9\,10,11\)
- Create an environment where developers who are open and transparent about the results of their use of processes such as these are seen as undertaking best practice, whereas those who don’t will rightly have questions asked of them by commissioners, users. If the requirements of undertaking processes outlined in the guidance are not too onerous, then this will create a level playing field amongst developers, be they large incumbents or small start-ups.

For these reasons, we ask that:

- Developers embark on a process of complying with the Code of Conduct for data-driven health and care technologies, and **make public and freely-available the results of the processes undertaken to do so**, for example those suggested by this guidance, alongside their own internal code of conduct or ethical framework for context. These results should be regularly reviewed and updated and submitted to NHSX for review and publication. Developers may also wish to make this information available on their website.
- A framework for conducting this assessment is provided separately in the policy toolkit.

---


SPECIFIC PROCESSES

Specific Processes: what can developers do to meaningfully comply with the Code of Conduct’s requirement that they:

i) Report on the kind of algorithm that is being developed/deployed and how it was trained, and demonstrate that adequate care has been given to ethical considerations relating to the selection of, obtaining of, and use of data for the development of the algorithm

- Identifying and publishing, in a freely-available manner, how personal data is collected, used, and shared, as well as the algorithms that underpin the data-driven technology, will increase trust in these data-driven technologies, and thus support their public acceptability.  

- Identifying and publishing, in a freely-available manner, the algorithms that underpin the tool being developed will help regulators better understand the data-driven technology landscape—that is, which types of technologies are being used and for which applications.

For these reasons, we ask that:

- Developers undertake a process of reflection on their proposed means of collecting, storing, using and sharing data, and on the proposed way that their algorithm(s) will work. We recommend the ‘Datasheets for Datasets’ approach advocated by Gebru et al (of Microsoft Research, in collaboration with the Georgia Institute of Technology, Cornell University, University of Maryland, and the AI Now Institute, New York).  
Alternatively, developers may wish to adapt existing resources such as the Open Data Institute’s ‘Data Ethics Canvas’. Whichever approach is selected, we strongly encourage developers to use (or adapt) frameworks for reflection that are publicly available and are already in use, rather than developing their own bespoke frameworks. This will allow comparison of responses across different developers and for different tools, allowing technology users, purchasers, and regulators to get a clearer picture of ‘what good looks like’.

- Developers clearly identify what type(s) of algorithm(s) constitute their data-driven technology, and go on to answer specific questions associated with that type of algorithm. A non-exclusive list of algorithm types that could be used in data-driven technologies in healthcare is provided in Table 3, along with information that should be provided by developers for each algorithm type. For machine learning models,

---


developers could adopt the approach advocated by Mitchell et al (of Google Research and the University of Toronto), namely the completion of a 'model card'.

Relevant resources

- ‘Model Cards for Model Reporting’, available at https://arxiv.org/pdf/1810.03993.pdf. This paper provides a framework to construct a ‘model card’ for a particular machine learning model. Once again, two case studies showing what a model card could look like are provided.

---

<table>
<thead>
<tr>
<th>Type of algorithm</th>
<th>Required information for this algorithm type</th>
<th>Rationale</th>
</tr>
</thead>
</table>
| Symbolic AI (e.g. expert system, production system) | ● Explanation of process by which rules determined and coded  
                  ● Publication of rules  
                  ● List of experts consulted (with declaration of interest (Doll))                                          | > Doll identifies whether any experts/developers have any reason to recommend one course of action over another                         |
| Machine learning: supervised learning        | ● What data was used to train the algorithm?  
                  ● Who labelled it?  
                  ● What a priori criteria for successful regression/classification were used?                                | > Identifies issues with data set (e.g. representativeness, coverage)  
                                                                            > Identifies expertise of labellers  
                                                                            > Avoids post hoc claims of success ('moving the goalposts') |
| Machine learning: unsupervised learning      | ● What data was used to train the algorithm?  
                  ● Why were these inputs chosen?  
                  ● Why was this project undertaken?                                                                | > As above                                                                                                                                 |
|                                                                                     |                                                                                                           | > Unsupervised learning may identify previously unknown associations/correlations. It is important to place these in some sort of theoretical framework, otherwise, we can end up focusing on meaningless associations |
| Machine learning: reinforcement learning     | ● What data was used to train the algorithm?  
                  ● What targets/reward functions were used?                                                            | > As above                                                                                                                                 |
|                                                                                     |                                                                                                           | > Explicitly identifies what goals of algorithm are                                                                                        |
| Machine learning: deep                     | ● The above may all apply (as DL can be)                                                                     |                                                                                                                                            |
learning/neural network | used in supervised/unsupervised/reinforcement methodologies | Topology of network E.g. number of nodes in input layer, number of hidden layers, representations in hidden layers (if possible), number of outputs, connections | > The topology of a neural network has implications for how the network functions and its learning methodology.

Table 3: A specification of different types of AI algorithms, with the specific information that developers should provide for each type, in addition to any general information about the tool such as that provided in a ‘datasheet’ for the model or dataset (see ‘Recommendations’ above).

ii) Prove the degree to which the algorithm is effective

► Any decision on the use of a medical intervention is based on a clear understanding of the risks and benefits. The use of data-driven technology should be no different. Studies to objectively and rigorously quantify the benefit of using the data-driven technology will provide this information in a way that is understandable to healthcare professionals, commissioners, and other stakeholders. If conducted properly, they will also gather data on potential risks (including those unforeseen at the time of development), allowing mitigating strategies to be drawn up. 18,19,20,21

► A number of studies have shown that the perception of benefit is a major factor that drives patients’ and the public’s readiness to engage with data-driven technology.22,23,24 Clearly communicating benefits in terms of positive outcomes to individuals and communities is likely to increase public acceptability, and is only possible with studies aimed at quantifying these benefits.

For these reasons, we ask that:

---

18 Interviews with Jeni Tennison (19th February 2019) and Naomi Lee (1st March 2019)
19 Interview with Luke Oakden Rayner (12th February 2019)
Guidance for Principle 7 of the NHS Code of Conduct for data-driven health and care technology

- Developers consider submission of their data-driven tools for external validation by assessment of their performance against standardised, validated datasets (as and when these become available e.g. For example, the World Health Organisation (WHO) and the International Telecommunications Union (ITU)—both UN specialised agencies—are developing a repository of validated and benchmarked secret datasets across a number of different clinical domains, the first of which should be available later this year.25 The Medicines and Healthcare products Regulatory Agency (MHRA) is also working with NHS Digital and the Care Quality Commission (CQC) on creating simulated datasets to be used for similar purposes.26)

- Given that NICE Evidence Standards Framework explicitly states that “it is not designed for use with [data-driven technologies] that incorporate artificial intelligence using adaptive algorithms”, developers who are creating tools containing this class of algorithms need to engage with NHSX at the earliest stage of development, in order to communicate:
  - The proposed method of continuous audit
  - The expected inputs and outputs against which performance will be continuously audited
  - How these inputs and outputs were determined
  - How these inputs and outputs are likely to impact the different stakeholders identified in the stakeholder analysis (Section B(i)).

- When reporting the results of clinical studies performed to determine the effectiveness of data-driven technologies, developers should use standard reporting frameworks, such as those being developed by the Equator Network.27

iii) Demonstrate that due consideration has been given to how the algorithm will fit into the wider healthcare system, and report on potential wider resource implications of deployment of the algorithm

- In order to fully understand the impacts of data-driven technologies, a systems-wide approach needs to be taken.28,29,30,31,32

- These technologies may have impacts akin to public health interventions, rather than drugs or traditional devices. Therefore, the traditional models of assessment of safety and effectiveness of medical products are insufficient to cope with their potential to disrupt whole healthcare systems, rather than simply impact the health of individuals.

---

25 Interview with Naomi Lee (1st March 2019)
26 Interview with David Grainger (1st March 2019)
27 Interview with Naomi Lee (1st March 2019)
31 Interview with Jonathan Pearson-Stuttard (5th March 2019)
For these reasons, we ask that:

▸ Developers clearly identify:
  ◦ The stakeholders (as per the exercise of stakeholder analysis outlined in Section B(i))
  ◦ The need/use case for the data-driven technology, and the existing care pathway(s) impacted by the tool
  ◦ The associated care pathways that interact with the target care pathway. For example, a tool designed for patients with diabetes may well have impacts on cardiovascular disease care pathways, and renal disease care pathways, as patients with diabetes are frequently seen on these pathways.
  ◦ The potential impacts on these target and associated care pathways of the tool

▸ Developers undertake the economic impact assessment outlined in Section B of the NICE Evidence Standards Framework.33

iv) Explain the algorithm to those taking actions based on its outputs, and to those on the receiving end of such decision-making processes

What do we mean by explain?

There are multiple reasons why full algorithmic explicability is difficult to achieve. Much has been made of the inherent impenetrability of certain types of machine learning algorithms, such as deep neural networks, and the same argument can reasonably be applied to any algorithm relying on huge biological datasets for training—the millions of variables that may be considered per sample places computational constraints on the ability to provide a full explanation of the workings of the algorithm.34 Experts in fact recognise that human explanations of their own behaviour (including those provided by healthcare practitioners) are largely based on a post hoc rationalisation of the decision taken, rather than an exhaustive understanding of our brain’s complex decision-making process. Different stakeholders may require different types of explanation, so it may be difficult to provide a single explanation that satisfies all interested parties. At the very least, we should distinguish between ‘model-centric explanations’ (where we try to understand exactly how the model works), and ‘subject-centric explanations’ (where we ‘just’ try to understand those aspects of model functioning that are relevant to the subject [the patient, in the healthcare setting]).35

It has also been (reasonably) suggested that ‘black box’ algorithms may be acceptable in the healthcare setting, provided that the trade-off is that the algorithm is clinically efficacious and improves patient outcomes, and that a meaningful explanation is given (as opposed to a full one). Therefore, rather than trying to achieve ‘full transparency’, it may be better to understand what a ‘good enough’ explanation would be for each stakeholder group identified in Section B

---


of this guidance. Such an explanation should give due regard to the type of application for the tool—that is, where and how it is intended to be used, by whom, and for whom (all of which should be identified in the stakeholder analysis), and its potential risk profile (as judged using the NICE Evidence Standards Framework, for example). It is expected that algorithms with higher harm potential affecting more vulnerable groups will require deeper, more complete explanations for function than ‘lower risk’ algorithms with less vulnerable users, for example.

The GDPR’s right to explanation can be viewed as a promising mechanism for developers to achieve accountability and transparency in algorithms, but Sandra Wachter and colleagues have argued that the GDPR requires data controllers (those handling personal data) to inform an individual in a meaningful way of how an algorithm is being used, and its envisaged consequence, but without explaining the steps taken by the algorithm. Furthermore, this article of the GDPR is only relevant if a complaint is made and if it can be proven that a data subject has suffered a negative consequence as a result of an algorithmic decision, and it may be difficult to enforce. All in all, this leads us to suggest that GDPR provides a necessary, but insufficient, framework to guide thinking around algorithmic explicable in healthcare.

Thus we argue that:

- Undertaking this process of providing transparency and accountability will increase trust in these data-driven technologies, and thus support their public acceptability.
- Part of the model validation process should include some understanding of how the outputs were arrived at, and whether these ‘make sense’ in the context of what is known about how human diseases originate and progress. Therefore, ensuring model veracity requires some element of algorithmic explicability.
- Although the GDPR does not provide a legally binding article on the right to explanation, it is possible that regulation may go in this direction in future. By being one step ahead, developers will be able to input on any future regulation that may put ‘the right to an explanation’ on a firmer statutory footing.

For these reasons, we ask that:

- Developers clarify the extent to which a decision based on an algorithmic tool is automated and the extent that a human has been involved in the process—that is, full transparency on the use of an algorithm.
- Developers use the stakeholder analysis exercise outlined in Section B (i) to clarify what is meant by the term ‘meaningful explanation’ for each stakeholder group.
- Developers coordinate with patient representative groups and other stakeholders to help develop ‘meaningful’ language as part of the explanation that will be understood by patients and other stakeholders.


39 Interview with Iain Forbes (25th February 2019)
Guidance for Principle 7 of the NHS Code of Conduct for data-driven health and care technology

- Where explanations remain too complex for lay comprehension, developers should support third parties that are trusted by patients (e.g. disease-specific charities) in acting as advocates for their patient groups.

v) Explain how the decision has been made on the acceptable use of the algorithm in the context it is being used (i.e. is there a committee, evidence or equivalent that has contributed to this decision)

Undertaking a process of transparently discussing and determining the acceptable use of the algorithm will increase trust in these data-driven technologies, and thus support their public acceptability.40,41 42 43 44

For this reason, we ask that:

- Developers **undertake a process of stakeholder analysis** as outlined in section B(i)
- Developers **utilise explicit activities** (such as user research, talking to patient groups and representatives, citizen juries, etc) to assess their thinking on the acceptable use of an algorithm. For example, nurses and clinicians should participate in the development of an algorithm that determines staff rotas.
- Developers **openly document their justification for and planning of these activities.**
- Developers **monitor user reactions** to the use of the data-driven technology, and **gauge levels of its acceptance on a continuous basis.** This could occur by means of regular monitoring exercises, scheduled in advance.

---


41 Interview with Aditya Nori (18th February 2019)


43 Interviews with Maxine Mackintosh (18th February 2019) and Jeni Tennison (19th February 2019)

APPENDICES

i) Glossary of Terms

Algorithm: A set or sequence of step-by-step operations that need to be carried out to perform a calculation, to process a set of data, or to test a logical statement.

Analytics: The scientific process of discovering and communicating meaningful patterns in data; and the process of applying those patterns towards effective decision making.

Artificial intelligence (AI, or machine intelligence): A field of study that combines computer science, engineering and related disciplines to build machines capable of behaviour that would be said to require intelligence were it to be observed in humans. Such behaviour includes solving problems.

Automation: The use of automatic processes and equipment in manufacturing or other settings.

Big data: large structured or unstructured datasets that are so complex that traditional data processing application software is inadequate to deal with them. The term has also been applied to the discipline of data analytics that has emerged to extract value from and identify patterns in these data.

Bioethics: A field of study dealing with the ethical implications of biological research and applications especially in medicine.

Black box: A device, system or object which can be viewed in terms of its inputs and outputs, without any knowledge of its internal workings.

Care pathway: A multidisciplinary outline for the mutual decision-making and organisation of care processes for a patient, placed in a well-defined period.

Deep learning: A branch of machine learning that involves algorithms that analyse data through multiple layers of complex processing. Each layer’s output becomes the input to the next layer to carry out pattern analysis and classification and to establish hierarchical relationships for both supervised and unsupervised learning.

Deep neural networks: A kind of deep-learning architecture based on artificial neural networks that uses multiple layers of processing units that loosely mimic human brain structure and can model complex nonlinear relationships.

Developer: Throughout this document, the term ‘developer’ is used to refer to individuals or organisations that are developing data-driven technologies for applications in health and care—that is, taking a tool from the idea stage, through the coding, training, and validation stages, and up to the point that it is ready for deployment. It should be stressed that these individuals or organisations could be from any sector, including the commercial, public, or academic sectors. Therefore, individuals working entirely within the NHS or in universities should see themselves as the audience for these guidelines just as much as developers working in start-ups, SMEs, or large corporates.

Machine learning: A type of artificial intelligence that has risen to recent prominence. It refers to the ability of computers to learn without being explicitly programmed. Algorithms use complex statistical methods to recognize patterns in data, learn from these patterns, and subsequently make predictions based on these data. Various techniques allow the algorithm to continuously improve its pattern-finding and predictive abilities.

Matrix: A grid that helps to evaluate and prioritise sets of values and information.
Online machine learning: A method of machine learning in which data becomes available and is used to update the model continuously, such that the algorithm changes as it is used. It is used in situations where it is necessary for the algorithm to dynamically adapt to new patterns in the data, or when the data itself is generated as a function of time.

Stakeholders: All those that research develop, design, deploy or use AI, as well as those that are (directly or indirectly) affected by AI— including but not limited to companies, organisations, researchers, public services, institutions, governments, regulators, social partners, individuals.

Triage: The assignment of degrees of urgency to wounds or illnesses to decide the order of treatment.
ii) Overlap with information required by other frameworks

It is important to note that demonstrating best practice outlined in Principle 7 does not exempt developers from seeking regulatory approval. Nevertheless, when seeking regulatory approval for example from bodies including the Medicines and Healthcare products Regulatory Agency (MHRA), developers will need to undertake work and data collection that will address a number of the processes outlined in this guidance. The table below indicates where some overlap may occur. This should be taken as a guide, not an exhaustive list, and we anticipate that it will change as regulation and associated frameworks evolve.

<table>
<thead>
<tr>
<th>Process</th>
<th>You may have already done this as part of:</th>
</tr>
</thead>
<tbody>
<tr>
<td>i) Stakeholder analysis, including value/consequence matrices</td>
<td>Medical Devices Directive Annex II:</td>
</tr>
<tr>
<td></td>
<td>1.1. The Technical Documentation should include (a) a general description of the device including its intended users; (c) the intended patient population</td>
</tr>
<tr>
<td></td>
<td>Medical Devices Directive, Article 10:</td>
</tr>
<tr>
<td></td>
<td>Manufacturers shall establish, document, implement and maintain a system for risk management as described in Section 3 of Annex I. The quality management system shall address at least the following aspects:</td>
</tr>
<tr>
<td></td>
<td>(c) responsibility of the management;</td>
</tr>
<tr>
<td></td>
<td>(j) handling communication with competent authorities, notified bodies, other economic operators, customers and/or other stakeholders;</td>
</tr>
<tr>
<td></td>
<td>Medical Devices Directive, Article 32:</td>
</tr>
<tr>
<td></td>
<td>For class III devices, the manufacturer shall draw up a summary of safety and clinical performance, which shall include target populations; possible diagnostic or therapeutic alternatives</td>
</tr>
<tr>
<td></td>
<td>MHRA Medical device essential requirements - general. Manufacturers must consider risks &amp; benefits; and consider the user’s technical knowledge etc</td>
</tr>
<tr>
<td>Section C. Specific Processes</td>
<td></td>
</tr>
<tr>
<td>-------------------------------</td>
<td></td>
</tr>
<tr>
<td><strong>i) Assess data issues and identify algorithm(s):</strong></td>
<td>Data Protection Impact Assessment (DPIA), required by the General Data Protection Regulation (GDPR), Articles 35 &amp; 36.</td>
</tr>
<tr>
<td>Reflect on proposed means of collecting, storing, using and sharing data</td>
<td>Article 35(1): Where a type of processing using new technologies, and considering the nature, scope, context and purposes of the processing, is likely to result in a high risk to the rights and freedoms of natural persons, the controller shall, prior to the processing, carry out an assessment of the impact of the envisaged processing operations on the protection of personal data. A single assessment may address a set of similar processing operations that present similar high risks.</td>
</tr>
<tr>
<td></td>
<td>Article 35(3): A data protection impact assessment referred to in paragraph 1 shall be required in the case of: (a) a systematic and extensive evaluation of personal aspects relating to natural persons which is based on automated processing, including profiling, and on which decisions are based that produce legal effects concerning the natural person or similarly significantly affect the natural person.</td>
</tr>
<tr>
<td></td>
<td>See ICO guidance</td>
</tr>
<tr>
<td></td>
<td>Application to the Confidentiality Advisory Group (CAG) if intending to access confidential patient information without consent in England and Wales</td>
</tr>
<tr>
<td></td>
<td>Application to an HRA Research Ethics Committee (REC) in England</td>
</tr>
<tr>
<td><strong>i) Assess data issues and identify algorithm(s):</strong></td>
<td>Medical Devices Directive Annex I: 15.1 Diagnostic devices and devices with a measuring function, shall be designed and manufactured in such a way as to provide sufficient accuracy, precision and stability for their intended purpose, based on appropriate scientific and technical methods. The limits of</td>
</tr>
</tbody>
</table>
| i) Prove algorithm(s) is/are effective | accuracy shall be indicated by the manufacturer.  

Medical Devices Directive Annex II:  
1.1. The Technical Documentation should include (j) a general description of the key functional elements, e.g. its parts/components (including software if appropriate)  

NHS Digital Clinical Safety team’s Standard DCB0129:  
3.5.1 The Health Organisation MUST produce a Clinical Safety Case Report to support each lifecycle phase (i.e. deployment, use, maintenance and decommissioning) of the Health IT System. |
| i) Prove algorithm(s) is/are effective | Medical Devices Directive Annex II:  
3. The Technical Documentation should include (b) complete information and specifications, including the manufacturing processes and their validation, their adjuvants, the continuous monitoring and the final product testing.  

NICE Evidence Standards Framework |
| ii) Prove algorithm(s) is/are effective: Engage with NHSX at the earliest stage of development, in order to communicate the proposed method of continuous audit, and the inputs & outputs | Medical Devices Directive, Article 10:  
Manufacturers shall establish, document, implement and maintain a system for risk management as described in Section 3 of Annex I.  

The quality management system shall address at least the following aspects: (e) risk management as set out in Section 3 of Annex I; (m) processes for monitoring and measurement of output, data analysis and product improvement.  

Medical Devices Directive Annex II:  
4. The documentation shall contain information for the demonstration of conformity with the general safety and performance requirements set out in Annex I |
that are applicable to the device taking into account its intended purpose, and shall include a justification, validation and verification of the solutions adopted to meet those requirements.

Medical Devices Directive Annex IX:
“The manufacturer shall establish, document and implement a quality management system as described in Article 10(9) and maintain its effectiveness throughout the life cycle of the devices concerned.”

2.1. “The manufacturer shall lodge an application for assessment of its quality management system with a notified body.”

Article 32: “...all manufacturers should have a quality management system and a post-market surveillance system in place which should be proportionate to the risk class and the type of the device in question”

ISO 14971:2012 Medical Devices: Application of Risk Management to Medical Devices

NHS Digital Clinical Safety team’s Standard DCB0129:

2.7.1 The Health Organisation MUST formally review its clinical risk management process at planned, regular intervals.

3.1.1 The Health Organisation MUST establish at the start of a project a Clinical Risk Management File for the Health IT System.

3.2.1 The Health Organisation MUST produce at the start of a project a Clinical Risk Management Plan, which will include risk acceptability criteria, covering the deployment of a new Health IT System

| iii) Consider the algorithm's interaction with the wider healthcare system (identify stakeholders; need; impact on care pathways) | NHS Digital Clinical Safety team’s Standard DCB0129: |
| | 4.2.3 The Health Organisation MUST define the operational environment and users of the |
| Health IT System which is to be deployed.  
7.2.1 The Health Organisation MUST establish, document and maintain a process to collect and review reported safety concerns and safety incidents for the Health IT System following its deployment. [This monitoring needs to extend beyond the Health IT System itself to include the impact on users, related healthcare processes and any change in intended use.] |
|---|
| NICE Evidence Standards Framework  
Evidence to show successful implementation of the DHT in the UK health and social care system. |
| iv) Comply with ‘right to an explanation’ |
| GDPR Recital 71  
The data subject should have the right not to be subject to a decision, which may include a measure, evaluating personal aspects relating to him or her which is based solely on automated processing and which produces legal effects concerning him or her or similarly significantly affects him or her, such as automatic refusal of an online credit application or e-recruiting practices without any human intervention.  
...  
In any case, such processing should be subject to suitable safeguards, which should include specific information to the data subject and the right to obtain human intervention, to express his or her point of view, to obtain an explanation of the decision reached after such assessment and to challenge the decision.  
Medical Devices Directive, Article 32:  
1. For implantable devices and for class III devices, other than custom-made or investigational devices, the manufacturer shall draw up a summary of safety and clinical performance. The summary of safety and clinical performance shall be written in a way that is clear to the intended user and, if relevant, to the patient and shall be made available to the public via Eudamed. |
| v) Explain how acceptable use of the algorithm is determined: Justify and use explicit activities (e.g. user research) | NICE Evidence Standards Framework  
Published or publicly available evidence to show that representatives from intended user groups were involved in the design, development or testing of the DHT and to show that users are satisfied with the DHT. |
| v) Explain how acceptable use of the algorithm is determined: Monitor user reactions on a continuous basis | Medical Devices Directive, Article 84:  
The post-market surveillance plan shall be part of the technical documentation specified in Annex II  

NHS Digital Clinical Safety team’s Standard DCB0129:  
7.2.1 The Health Organisation MUST establish, document and maintain a process to collect and review reported safety concerns and safety incidents for the Health IT System following its deployment. [This monitoring needs to extend beyond the Health IT System itself to include the impact on users, related healthcare processes and any change in intended use.]  

NICE Evidence Standards Framework  
Evidence that data on outcomes or user satisfaction is being collected in line with the minimum standard and can be made available to relevant decision-makers. |
iv) List of Contributors

a) Interviewees

► **Jacob Beswick**, Policy Analyst at the Office for Artificial Intelligence at Department for Business, Energy and Industrial Strategy (BEIS)

► **Dr Steven Bishop**, Strategic Analytics and AI Manager at CMR Surgical

► **Professor Rafael Calvo**, Professor at the Dyson School of Design Engineering (Imperial College London) and Professor and Director of the Wellbeing Technology Lab at the University of Sydney

► **Yo-en Chin**, Deloitte Digital

► **Professor Virginia Dignum**, Professor of Social and Ethical AI at University of Umeå in Sweden and Associate Professor at the Faculty of Technology, Policy and Management, Delft University of Technology

► **Dr Orla Doyle**, Lead Data Scientist at IQVIA

► **Iain Forbes**, Head of the Centre for Connected and Autonomous Vehicles at the Department for Transport/Department of Business, Innovation and Skills

► **David Grainger**, Devices Software & Apps Manager at the Medicines and Healthcare products Regulatory Agency (MHRA)

► **Ed Greig**, Chief Disruptor at Deloitte Digital

► **Ben Hawes**, Freelance Consultant on AI Policy and Strategy

► **Professor Chris Holmes**, Programme Director for Health and Medical Sciences at The Alan Turing Institute, and Professor of Biostatistics at the University of Oxford

► **Dr Matthew Howard**, Director of AI at Deloitte

► **Stephanie Kelley**, PhD Candidate in Management Analytics at the Smith School of Business at Queen's University Canada

► **Dr Naomi Lee**, Executive Editor - Digital at The Lancet

► **Maxine Mackintosh**, PhD candidate at University College London’s Farr Institute of Health Informatics and Co-founder of One HealthTech

► **Jess Morley**, Technology Adviser at NHSX

► **Dr Parashkev Nachev**, Senior Clinical Research Associate at the Institute of Neurology, University College London
Dr Aditya Nori, Principal Researcher and Healthcare ML Team Lead at Microsoft Research Cambridge

Professor Jim Norton, Chair of the Royal Academy of Engineering Community of Practice in Digital Systems Engineering, and Pro-Chancellor of Coventry University

Dr Luke Oakden-Rayner, Director of Medical Imaging Research at Royal Adelaide Hospital, and PhD candidate at the University of Adelaide

Dr Jonathan Pearson-Stuttard, Wellcome Trust Clinical Research Fellow at Imperial College London, and Public Health Registrar at Imperial College NHS Trust

Professor Daniel Ray, Director of Data Science at NHS Digital

Professor Geraint Rees, Dean of the UCL Faculty of Life Sciences and Professor of Cognitive Neurology and Wellcome Trust Senior Clinical Fellow at University College London

Tim Sheppard, SVP & General Manager, Northern Europe at IQVIA

Jeni Tennison, CEO of the Open Data Institute

Robert Turpin, Healthcare Market Development Manager at the British Standards Institution (BSI)

Aashish Upmanyu, IoT Studio, Innovation Senior Consultant at Deloitte Consulting

Dr Adrian Weller, Programme Director for AI at the Alan Turing Institute, and Senior Researcher in the Machine Learning Group at the University of Cambridge

Dr Sebastian Vollmer, Programme Co-Director for Health and Director of the Turing Data Study Groups, and Associate Professor at the Departments of Mathematics and Statistics at the University of Warwick

b) List of attendees at expert round tables

Dr Saif Ahmad, Clinical Lecturer in Clinical Oncology at the University of Cambridge

Dr Hutan Ashrafian, Clinical Lecturer in Surgery at Imperial College London

Dr Steven Bishop, Strategic Analytics and AI Manager at CMR Surgical

Dr Christopher Burr, Postdoctoral Researcher at the Oxford Internet Institute

Emma Doyle, Head of Data Policy at NHS England

Dr Keith Grimes, Clinical Innovation Director at Babylon Health

Matthew Honeyman, Policy Researcher at The King’s Fund
Dr Alan Karrhikesalingam, Senior Clinician Scientist at DeepMind Health

Dr Pearse Keane, Consultant Ophthalmologist at Moorfields Eye Hospital, and NIHR Clinician Scientist at the University College London Institute of Ophthalmology

Dr Naomi Lee, Executive Editor of the Lancet

Sinead MacManus, Senior Programme Manager for Digital Health at Nesta

Dr Jagtar Nijjar, NIHR Clinical Lecturer in Rheumatology at the University of Cambridge

Professor Andrea Rockall, Clinical Chair of Radiology at Imperial College London, and Hon Consultant Radiologist at Imperial College Healthcare NHS Trust

Mark Salmon, Programme Director of Information Resources at the National Institute for Health and Care Excellence (NICE)

Hetan Shah, Executive Director of the Royal Statistical Society

Dr Harpreet Sood, Associate Chief Clinical Information Officer for NHS England

Dr Adam Steventon, Director of Data Analytics at the Health Foundation

Dr David Wong, Lecturer in Health Informatics at Leeds University

Dr Louise Wood, Director of Science, Research & Evidence at the Department of Health and Social Care (DHSC)

c) Contributors in writing

Professor Michael D. Abràmoff, The Robert C. Watzke, MD, Professor of Ophthalmology and Visual Sciences, Electrical and Computer Engineering, and Biomedical Engineering, Department of Ophthalmology and Visual Sciences, University of Iowa Hospital and Clinics, and CEO and Founder of IDx.

Professor Grigoris Antoniou, Professor of Semantic & Knowledge Technologies at the University of Huddersfield

Dr Peter Bannister, VP of Clinical Innovation at Mirada Medical, and Chair of the IET Healthcare Panel

Haitham Baomar, AI & Aerospace Researcher at University College London

Professor David Barber, Professor of Machine Learning and Director of the UCL Centre for Artificial Intelligence at University College London

Priya Basker, Data Scientist at NHS Digital
Dr Tom Foley, Senior Clinical Lead for Data at NHS Digital

Adrian Jonas, Associate Director of Data and Analytics at the National Institute for Health and Care Excellence (NICE)

Hasan Jouni, Business Development Manager at Siemens Digital Healthcare Services

Dr Dominic King, Clinical Lead at DeepMind Health and Honorary Clinical Lecturer in Surgery at Imperial College London

Dr Nicole Mather, Non-Executive Director of the Health Research Authority, and Director of Life Sciences and Healthcare at Deloitte Strategy

Dr Homer Pien, Senior VP and Chief Scientific Officer at Philips Healthcare

Mark Salmon, Programme Director of Information Resources at the National Institute for Health and Care Excellence (NICE)

d) Patients and members of the public who contributed to this report

Richard Ballerand

Ann Cawley

Jeremy Dearling

Eric Deeson

Martin Lewis-Jones

John Marsh

Deirdre Mcellan

Stephen Pattison

Alan Quarterman

Ursula Van Mann

v) Methodology by which patient/public contributors were recruited

Patients and members of the public that participated in our round table on the 8th April 2019 were recruited via one of two methods. Firstly, we placed an advert on the ‘People in Research’ website, which is run by the National Institute of Health Research (NIHR)’s INVOLVE programme, which aims to support active public involvement in the NHS, public health and social care research. Secondly, we recruited from ongoing projects and bodies with established
groups of patients and members of the public that are involved in their research. These included the Royal College of Physicians’ Patient and Carer Network, and the British Heart Foundation’s Patient Data Panel. Administrators for these bodies / research kindly circulated the notice regarding our roundtable through their networks.
Guidance for Principle 7 of the NHS Code of Conduct for data-driven health and care technology