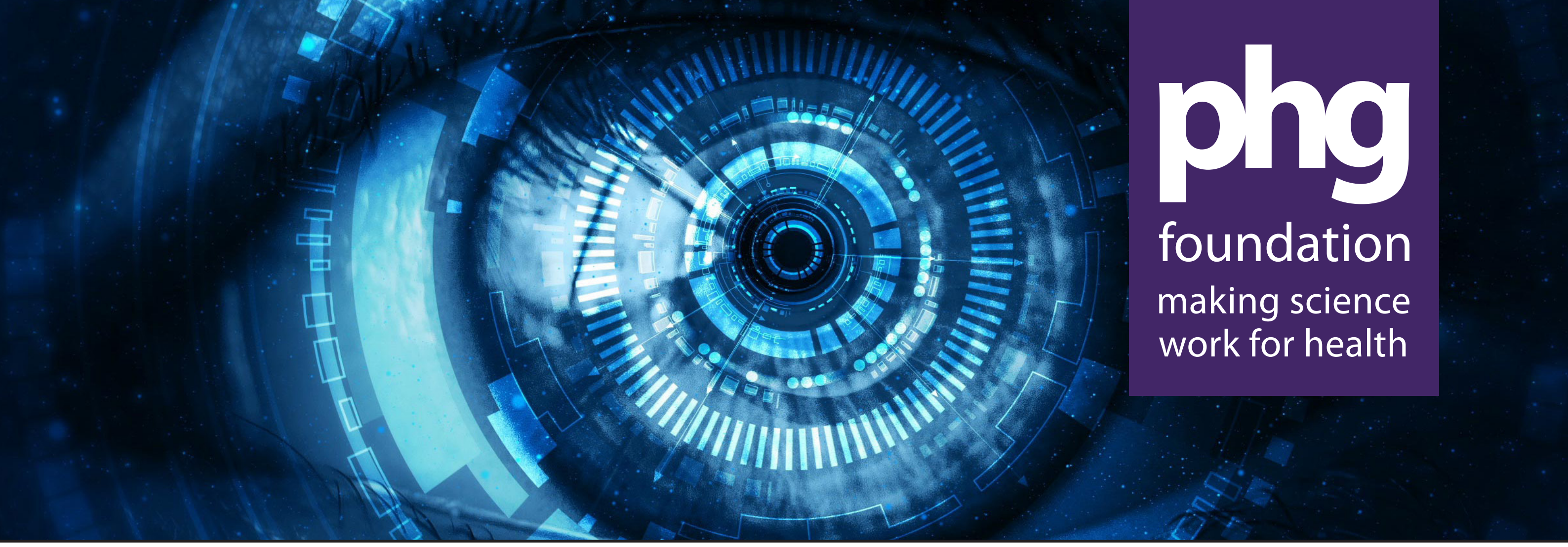

phg foundation
making science work for health

May 2020

Black box medicine and transparency

Artificial intelligence (AI) offers great potential for health in both medical research and care, but the computerised analysis of data by opaque (black box) machine learning processes poses significant practical, ethical and regulatory challenges.

Medical applications of machine learning (a form of AI) offer exciting opportunities for automated data analysis to provide new insights and more accurate predictions than even expert professionals can achieve. This could offer a wealth of benefits, from more efficient triaging and scheduling of appointments and rotas through to earlier detection of disease, more precise diagnosis and treatment, and more effective public health surveillance. However, the logic involved and the processes by which these machine learning models reach conclusions is not obvious - a potential hurdle for effective regulatory oversight, and for securing user and public trust.

Summary

- AI has many useful applications for medical research and healthcare. The context guides the regulatory framework that applies, but the boundary between these two areas is often blurred
- Machine learning models for many medical applications will need to be interpretable to variable degrees depending on their nature and purpose
- The context will influence whether to optimise their usability potentially at the cost of being less interpretable, or to use approaches that demonstrate compliance with ethical and legal requirements
- The nature and risk profile of the intended medical application should also guide the intepretability of a prospective machine learning model
- A new framework for developers of machine learning tools for healthcare can help ensure consideration of all relevant factors to determine the optimal transparency required

UNIVERSITY OF CAMBRIDGE

Machine learning and interpretability

Unlike traditional computer programs, machine learning models are trained with data relevant to the task they are undertaking, the better to automate this task. This can improve performance, but also means that it is less easy to understand how they work, even for technical experts, and even where there are no restrictions on the disclosure of such information, for example commercial sensitivities.

The interpretabilty of machine learning models describes the ability to explain their working, including meaningful information about the logic involved, to a human. Good explanations should be appropriate to the audience, and are likely to vary depending on context.

Why interpretability matters for medicine

The ability to explain how machine learning models work is important to:

- Evidence safety and effectiveness
- Facilitate human-computer interaction
- Assist in scientific or causal understanding
- Underpin control by the data subject or controller accountability

It may be particularly important to secure an adequate explanation where decisions result in high-stakes outcomes, which is often the case in medical research or healthcare. Ensuring that machine learning tools can be explained also supports transparency, a prerequiste for trustworthiness of healthcare tools or processes. Developers of machine learning models can assess how interpretable they are likely to be, using different methods depending on the stage of development, intended application and risk profile - for example, a model that could directly influence significant treatment decisions may have a higher risk profile than one that predicts which patients are likely to miss their next appointment.

Transparency and regulation

Some medical applications of machine learning are subject to a complex range of legal controls, including the General Data Protection Regulation (GDPR) in the UK and EU. This could apply to machine learning in healthcare in multiple ways, where any form of personal data is used (both in and outside Europe), or any type of treatment decision or recommendation generated.

Principles of data processing, data subject rights and special conditions for automated processing place separate but linked obligations on data controllers and processors. In combination, these obligations may mean that explanations of the model as a whole (global interpretability) and of individual decisions (local interpretability) are necessary.

Resources for policy-makers and developers

The *Black box medicine and transparency* project has produced a detailed set of resources for legal, regulatory and health policy audiences, examining the issues in detail. In addition, a dedicated Interpretability by Design framework for developers of machine learning models for healthcare sets out a simple process to review and optimise interpretability. These resources are all freely available from the PHG Foundation website.

The *Black box medicine and transparency* project was generously funded by the Wellcome Trust

PHG Foundation is a policy think tank with a special focus on how genomics and other emerging health technologies can provide more effective, personalised healthcare

Contact: intelligence@phgfoundation.org

phg foundation
making science
work for health