
Position: The Role of Patient and Public Involvement in Building Trustworthy and Safe Healthcare Generative AI

Soumya Banerjee
University of Cambridge
Cambridge, UK

Jad Sbai
University of Cambridge
Cambridge, UK

Anna Moore
University of Cambridge
Cambridge, UK

Alisa Anokhina
University of Cambridge
Cambridge, UK

Abstract

Artificial intelligence (AI) is helping researchers and clinicians make great strides in healthcare. However, there is a trust deficit in AI, especially when applied to critical areas like healthcare. Hence, communicating the beneficial medical applications of AI and engaging the public with healthcare AI research is critical. One way to achieve this is by getting the community involved in co-designing better AI systems in healthcare projects, especially in those related to mental health. We believe that integrating patient and public involvement (PPI) in AI projects may help in adoption and acceptance of these technologies.

We contend that for AI algorithms in healthcare to be effective and reliable, they must be developed collaboratively with input from patients and healthcare professionals. We suggest a roadmap for this collaborative approach in AI model building. This will involve actively including patients with lived experience of a disease, as well as creating a research advisory group to walk patients through the process of AI model building.

We suggest formulating and scoping a problem, and then generating a hypothesis that patients and scientists agree on. Data scientists should then start by building simple models and explain them to patients and clinicians.

We outline a case study from a research project using generative AI in mental health. We involved patients in this research project and they gave suggestions that gave us insights into our research questions.

In our work on mental health, we built simple models that can be explained very easily. This will lay the groundwork for progressively building more complex models and explaining each step of the model building process to patients.

We have difficulty trusting what we do not understand. We may easily trust what we create (and hence understand). The road to building trustworthy and safe AI systems may become easier if all stakeholders are involved in co-creating AI models.

1 Introduction

Artificial intelligence (AI) is helping researchers and clinicians make great strides in healthcare. Some people, however, believe that AI presents more drawbacks than benefits. Hence communicating its beneficial medical applications and engaging the public with healthcare research is critical.

One way to achieve this is by getting the community involved in co-designing better AI systems in healthcare projects. We believe that integrating patient and public involvement (PPI) in AI projects may help in adoption and acceptance of these technologies.

2 Building better AI systems

Explainable AI is a set of methods that allow users to comprehend and trust the results and outputs created by AI algorithms. What explanations are useful will be different for computer scientists, clinicians and patients alike. One way to identify which explanation is most useful is to let people co-develop models, play with these models and generate their own explanations.

We argue that AI algorithms for healthcare should be co-designed with patients and healthcare workers, so that they are useful and trustworthy.

In one of our projects, a patient with a severe mental illness approached us. The patient had bipolar disorder and was taking a medication called lithium: a very effective medication, but which can damage the kidney if taken for many decades. The patient asked us whether we could use AI to understand whether stopping lithium may reverse kidney damage. We examined hospital records to identify patients who had discontinued lithium after prolonged use, and to determine whether this cessation had led to a reversal in kidney damage. This study was approved by an Institutional Review Board (IRB) at our university.

A patient coming to us with a well-defined hypothesis to guide the research agenda was profound. This patient-driven method ensures that study participants are not just treated as data points and avoids patients being marginalised in the research process. We built simple models to start the process. Initially, we built simple regression models and explained them to the patient. Subsequently, we added layers of complexity to these models and in monthly meetings, took steps to explain and validate them with everyone involved.

We suggest a roadmap for this collaborative approach in AI model building. This will involve actively including patients with lived experience of a disease, as well as creating a research advisory group (RAG) to walk patients through the process of AI model building.

We suggest formulating and scoping a problem, and then generating a hypothesis that patients and scientists agree on. Data scientists should then start by building simple models and explain them to patients and clinicians. In our work, we built simple regression models which can be explained very easily. This will lay the groundwork for progressively building more complex models and explaining each step of the model building process to patients.

3 The benefits of building together

Perhaps the most important benefit of this methodology is that co-developing models transparently, and marrying the technical with the human, may help people build trust in AI systems. It enables patients to not only help shape the models, but also develop a relationship of trust with the team of researchers behind them, because they listen to people and try to understand their needs and allay their concerns.

Opening AI models to investigation not only enables rich interaction, but also pinpoints errors and areas of improvement. It can also highlight unconscious biases.

4 Case study

In this section, we outline a project (as a case study) on how we used generative AI in a mental health project and also involved patients.

In clinical data analysis, missing values in digital health records present significant challenges that compromise the accuracy and effectiveness of machine learning algorithms. Current imputation methods often lack interpretability and fail under high levels of missingness.

To address these limitations, we created an imputation method that leverages subpopulation dynamics within the dataset and incorporates domain knowledge. Our approach uses Large Language Models (LLMs) to serve as an interface between the imputation process, clinicians, and patients.

Our research focuses on the National Community Child Health (NCCH) Dataset, which includes birth records across different health services in Wales (recording variables such as maternal age and birth weight). This dataset, used to predict mental health disorders in children Crowley et al. [2024], shows up to 80% missingness in several key risk factors, such as whether the mother smoked during pregnancy.

The NCCH dataset contains records of all children born, resident, or treated in Wales, born after 1987. In total, the dataset comprises **233,632** records.

For brevity, we will not describe all variables of the dataset here; a detailed description is available in Jones et al. [2024]. Instead, we focus on the variables identified as risk factors for mental health disorders.

Key risk factors such as Maternal Smoking and Breastfeeding have more than 60% and 80% missingness, respectively. This significant level of missingness is the main motivation for imputing these variables in this dataset. Without imputing these missing values, it is challenging to build accurate predictive models Shadbahr et al. [2023].

An important point to address is how we create the truth values to evaluate our methods. We use a complementary dataset that is part of the SAIL databank, the Maternity Indicator Dataset (MIDS), which also records these variables.

Using the complementary dataset, we have truth values for **maternal smoking**, **breastfeeding at birth**, and **birth weight** (including its associated categorical variable, birth weight category), on which we can evaluate our approach.

We specifically investigated patterns in maternal smoking and breastfeeding data. We observed that if a person does not disclose breastfeeding at birth, they are likely to withhold smoking information as well.

We show an example output where the imputation was 1 for maternal smoking, but the LLM decided to impute 0 (shown in Table 1). We used the Microsoft Phi3 model, which has 3 billion parameters, offering a practical alternative to larger and close-source models such as GPT-4.

4.0.1 PPI Insights in a mental health prediction project

To understand potential reasons for missing data, we conducted a Patient and Public Involvement (PPI) survey via Microsoft Teams (5 responses). The patients gave insights as to why certain values may be missing in the first place in electronic healthcare records.

To recruit participants, we posted messages on our Research Advisory Group channels (we have one for parents/carers and one for young people). The Research Advisory Group consisted of 14 young people and 10 parents/ caregivers.

Members of the Research Advisory Group had completed relevant informed consent procedures to be part of the group (and have taken part in extensive upskilling), on the understanding that they would take part in various PPI activities - including this one. They additionally gave consent to be added to the Teams channel for their group (young people and parents have separate ones), and we made it clear that responding to the discussion question was voluntary.

Here are some excerpts from what the members of the Research Advisory Group said:

"In the specific examples of maternal smoking and breastfeeding, there is an element of shame surrounding those who DO smoke maternally and those who do not/cannot breastfeed as well as concerns around being judged for breastfeeding 'too late'."

"A lot of information is collected based on a formula. For example as medical students we are given a proforma for certain histories that we take. For example, in a gynae history you'd ask about STIs but you wouldn't ask that if you were on a cardiovascular ward for example."

Table 1: LLM Sample Output (Anonymised)

<p>Based on the provided information and indicators, here is the analysis:</p> <ul style="list-style-type: none"> - Indicator 1: The patient's maternal age is [AGE] which is below [YOUNG_AGE]. This increases likelihood of smoking. - Indicator 2: The patient's baby's birth weight is [BIRTH_WEIGHT] grams, which is more than [NORMAL_WEIGHT] grams. This does not particularly indicate smoking. - Indicator 3: The patient's gestational age is [GESTATIONAL_AGE] weeks, which is more than [NORMAL_GESTATIONAL_AGE] weeks. The birth seems healthy which doesn't point towards smoking. - Indicator 4: In the SMOKE_NC_imputed_full scenario, there are more smokers in the closest records than amongst the farthest records. This might indicate potential smoking. However, in both the SMOKE_NC_imputed_full_selected scenario and SMOKE_NC_imputed_cluster scenario, the number of smokers amongst the closest records is less than amongst the farthest records. <p>Considering these insights, only 1 indicator supports that the person smokes. Therefore, the patient is likely not a smoker.</p>

These findings suggest that the high percentage of missing values for smoking and breastfeeding may be informative in a way not explicitly recorded in the dataset. It also suggests that missing data could be influenced by factors such as social stigma.

Overall, our approach allowed us to discover interesting mechanisms of missingness and find a way to include them in the imputation approach when the missingness itself is informative. With better PPI insights, we could potentially include them in the imputation process and leverage LLMs to use them.

Patients gave us insights as to why certain values may be missing in the first place. By engaging patients in the research process, we obtained valuable information that would have been difficult to discover otherwise.

5 Patient perspective

Lived experience of a disease is important in healthcare research Borg et al. [2011]. Patients and carers have important research ideas about how best to improve quality of life, manage symptoms, offer existing treatments, or develop new interventions. Often these ideas differ from those prioritised by academia or the pharmaceutical industry.

We had a long and deep engagement with patients in all steps of AI research, from hypothesis generation to model building and understanding. In this way, patients felt they were involved in this project. This also gave a sense of agency and voice to patients.

6 Recommendations

We specifically suggest: 1) including patients with lived experience of the disease and carers, 2) creating a Research Advisory Group (RAG) and using these group meetings to involve patients and carers in all stages of the scientific process (starting from hypothesis generation). We also recommend explaining the process of AI model building, starting with simple (e.g. linear) models. We suggest using freely available AI models that run in the browser (such as the Teachable Machine from Google)

to explain the basics of AI to patients. These meetings should be repeated to elicit feedback from the stakeholders, explain model predictions and get guidance on model modifications.

In RAG meetings, we built trust and solicited comments on how patient data could and should be analysed. We showed patients how we took precautions to preserve privacy and allayed other concerns. We sought to reduce the “hype” around AI and demystify these techniques using simple examples.

We explained how AI will be used on clinical data and how the expected outcomes might benefit patients. In turn, we learnt from patients and carers about important features of the data, and about the concerns that must be addressed to implement AI models in practice - including the potential for inadvertent discrimination by AI Health [2019] McCradden et al. [2020] Leufer [2020].

To build truly trustworthy AI systems, it is also essential to engage not only people who have lived experiences relevant to the research question but also individuals from diverse backgrounds. This diversity should encompass a wide range of factors beyond ethnicity, including age, sex, comorbid conditions, and neurodivergence. Engaging with people who have different experiences, attitudes, and needs allows us to better understand the “real-world” context in which the data exists. This broader perspective is crucial to ensure that AI models are developed in a way that is both inclusive and reflective of the varied ways in which people interact with and experience healthcare systems.

We propose a methodology for fostering patient trust in AI within healthcare. This methodology can be tailored to meet the specific needs and budgetary limitations of individual projects.

7 Caveats and limitations

Unfortunately, such a participatory approach does not happen often because it can be difficult. It is much easier to get access to data and write papers than overcome logistical challenges like recruiting patients and finding time for everyone to meet in a Research Advisory Group. There are also few incentives for researchers to work in this way as we typically gain tenure for publishing impactful papers rapidly, rather than experimenting with new engagement techniques.

8 Discussion

To build trustworthy AI systems, it is crucial to adopt a multidisciplinary approach that brings together not only patients but also data scientists and clinical teams. Working collaboratively with patients ensures that AI models are relevant and address the needs of those who will be directly impacted by them. Moreover, involving data scientists ensures that the technical elements of AI development are strong. Additionally, collaboration with clinicians bridges the gap between technical innovation and practical healthcare application. By integrating the expertise of all these stakeholders, we can develop responsible AI systems that are not only technically sound but also ethically grounded.

A collaborative process ideally necessitates the creation of a research advisory group to involve all stakeholders in the process of building AI models. We believe this is how all AI research in healthcare and potentially other critical domains should be conducted: where everyone, including end users (patients) and domain experts (like clinicians and data scientists) can work together effectively with a shared goal, and collectively benefit from the results.

An understanding of what AI can and cannot do will build trust and allay fears. A realistic appraisal of risks and benefits may help in adoption and democratise access to AI for healthcare, which is in everyone’s interests.

We have difficulty trusting what we do not understand. We may easily trust what we create (and hence understand). The road to building trustworthy AI systems may become easier if all stakeholders are involved in co-creating AI models.

The term AI has become linked to frightening concepts such as machines taking over the planet. A way to shift this narrative is by empowering individuals to explore the potential of AI and allowing them to form their own understanding of these technologies. Engaging people in every facet of AI creation and deployment can help make AI more human-centric and safe. For AI to thrive in the future within society, it is essential to involve people in both understanding and constructing AI.

References

- R. Crowley et al. Machine learning for prediction of childhood mental health problems in social care. *medRxiv*, 2024. URL <https://doi.org/10.1101/2024.2005.2003.24306756>. Preprint.
- Hope Jones, Mike J Seaborne, Natasha L Kennedy, Michaela James, Sam Dredge, Amrita Bandyopadhyay, Adele Battaglia, Sarah Davies, and Sinead Brophy. Cohort profile: Born in wales—a birth cohort with maternity, parental and child data linkage for life course research in wales, uk. *BMJ open*, 14(1):e076711, 2024.
- Tolou Shadbahr, Michael Roberts, Jan Stanczuk, Julian Gilbey, Philip Teare, Sören Dittmer, Matthew Thorpe, Ramon Viñas Torné, Evis Sala, Pietro Lió, et al. The impact of imputation quality on machine learning classifiers for datasets with missing values. *Communications Medicine*, 3(1):139, 2023.
- Marit Borg, Bengt Karlsson, Ann Mari Lofthus, and Larry Davidson. "hitting the wall": Lived experiences of mental health crises. *International Journal of Qualitative Studies on Health and Well-being*, 6, 2011. ISSN 17482623. doi: 10.3402/qhw.v6i4.7197.
- The Lancet Digital Health. There is no such thing as race in health-care algorithms. *The Lancet Digital Health*, 1:e375, 12 2019. ISSN 25897500. doi: 10.1016/S2589-7500(19)30201-8.
- Melissa D. McCradden, Shalmali Joshi, Mjaye Mazwi, and James A. Anderson. Ethical limitations of algorithmic fairness solutions in health care machine learning, 5 2020. ISSN 25897500.
- Daniel Leufer. Why we need to bust some myths about ai. *Patterns*, 1, 10 2020. ISSN 26663899. doi: 10.1016/j.patter.2020.100124.