Protocol for developing best practice guidelines in healthcare organisation artificial intelligence deployment: CHecklist for Actioning Responsible MLOps (CHARM)

J Zhang*[1,2], J Morley*[3], J Gallifant[2,4], W William[5], R M Carrillo-Larco[6], Prof. J T Teo[7], Prof. L A Celi[4,8], Prof. H Ashrafian[1,9], and the CHARM expert working group[±]

1 Institute of Global Health Innovation, Imperial College London, UK

2 Guy's and St. Thomas' Hospital, London, UK

3 Bennett Institute for Applied Data Science, University of Oxford, UK

4 Laboratory for Computational Physiology, Massachusetts Institute of Technology, Cambridge, MA, USA

5 Department of Biomedical Sciences and Engineering, Mbarara University, Uganda

6 Hubert Department of Global Health, Emory University, Atlanta, GA, USA

7 AI Centre for Value Based Healthcare, King's College London, UK

8 Department of Biostatistics, Harvard T.H, Chan School of Public Health, Boston, MA, USA

9 University of Leeds Business School, Leeds, UK

***Corresponding authors: JZ (joe.zhang@imperial.ac.uk); JM (jessica.morley@phc.ox.ac.uk)**

[±]Collaborators

Full list of expert collaborators will be published on the CHARM website: https://charmforhealth.com

Introduction

A decade of rapid growth in artificial intelligence (AI) research and funding has pushed clinical AI towards an inflection point[1]. Powerful yet accessible algorithms can maximise the potential of complex clinical datasets and can be readily applied to a myriad of use-cases across diagnosis, risk prediction, and planning. However, it is arguable that algorithmic progress has outstripped other necessary components of successful clinical AI deployment[2]. These include physical qualities such as data availability, data quality, and data/deployment infrastructure, but also include human resources, and pathway considerations such as clinical healthcare and workflow impacts and financial viability[3–5]. In the on-going rush to scale AI deployments, there are risks arising from a lack of data and pathway components, including to patient safety, and through the reinforcement of existing healthcare disparities[6,7].

Medical device regulators are critical for securing post-deployment safety, effectiveness, and fairness. However, rapid growth has left organisations such as the U.S. Food and Drug Administration (FDA) in a position of 'playing catch-up'[8]. Areas of risk arising from insufficient regulatory oversight are frequently described, and include: (1) overwhelming numbers of approvals based solely on retrospective dataset evaluation rather than real-world performance data (including 126 out of 130 FDA approved devices in a 2021 review by Wu et al[9]); (2) a lack of requirement to submit subpopulation-specific calibration data, risking unobserved biases in

deployment[10]; (3) absence of reporting transparency for data and models, removing ability for peer review and responsible end-user interpretation[11]; (4) continued delays in implementing approaches to dynamic model evolution (termed "change control") in response to changing local contexts and data; and (5) difficulties in establishing clear boundaries between software as medical devices (SaMD) and clinical decision support[12]. These factors have led product vendors to create levels of self-regulation in anticipation of future regulatory updates[13]. Rapid developments driven by Large Language Models (LLMs) further complicate this landscape, as the regulatory problem statement is markedly different from diagnostic or predictive SaMD, where objective metrics are more easily reviewed for narrow task descriptions. These lead to a problem of scalability. The volume of regulatory data, review time, and expertise required for pre- and post-market lifecycle evaluation will only increase as current recommendations translate into regulation, and as producers of SaMD continue to multiply exponentially[8,14]. Ultimately, regulatory bodies may find it increasingly difficult to maintain balance between permitting innovation, and ensuring quality and safety standards.

Key, overlooked actors are the healthcare organisations that deploy AI on clinical pathways. Whether organisations procure regulated SaMD, or whether they deploy unregulated algorithmic decision support, their practices currently fall outside of regulatory scope. Despite this, they hold practical oversight over product systems, including control over elements such as data generation, translation of AI outputs into interventions, and

integration with clinical pathways and staff[15]. Advantages to introducing healthcare organisation oversight of AI have been discussed[10,12], while an exemplar of whole-organisation AI implementation (*Mayo Clinic Platform*, Rochester MN, USA) has been previously described in case-study[3]. In summary, future patients will likely encounter multiple predictive models on a healthcare journey that is particular to the local context. There are therefore advantages to considering (1) shared deployment practices that suit local infrastructure, expertise, and patient need; (2) shared post-market surveillance through unified impact and safety reporting; and (3) shared model maintenance practices for ensuring calibration to local populations and monitoring for drift.

Anticipating a future where clinical AI is adopted at scale, it is likely that healthcare organisations will become central nodes in operationalisation. We suggest that levels of responsibility for AI deployment will redistribute between manufacturers and healthcare organisations. Resultantly, there is a clear need for organisation-centric guidance in maintaining AI deployments. Outside of healthcare, AI scaling within organisations is more mature, with widespread acceptance of a Machine Learning Operations, or MLOps, paradigm[16]. As an extension of software engineering concepts such as Development Operations (DevOps), MLOps refers to the processes that maintain the continuous lifecycle of a deployed model - including infrastructure, production, governance, security, observation, and revalidation. Key is the understanding that robust MLOps enables safe and effective deployment of models across many use-

cases. While we can take important lessons from non-healthcare industries, MLOps for healthcare presents unique challenges. In particular, risk management must contend with adverse and potentially catastrophic impacts on patients (as opposed to company operations or revenue)[17,18]. While a major component of MLOps addresses data/model drift, the implications of biased predictions are arguably more significant in healthcare terms[7]. Additionally, MLOps must consider ubiquitous challenges in healthcare data, that include multimodality, the inconsistent application of standards, risk of bias, substantial quality variation, and susceptibility to drift in response to coding practices[19].

Robust guidelines for clinical MLOps within healthcare organisations are a necessity. Several well-validated checklists exist to ensure transparent reporting of AI models in a research and development context (STARD-AI, TRIPOD-AI, PROBAST-AI, DECIDE-AI)[20–22], while there is growing consensus around the concepts that are most important to responsible usage of AI within healthcare[6,23]. However, there is less consideration given to translating these into implementation and deployment processes that are both practical and generalisable for organisations. In this paper, we lay out a protocol for developing CHARM (CHecklist for Actioning Responsible MLOps), using an expert working group and international Delphi consensus methodology.

Aims

We aim for CHARM to set comprehensive guidelines that healthcare organisations can follow to implement and maintain MLOps across their digital ecosystem. We specifically consider deployment of clinical diagnostic or predictive models that are trained using machine-learning algorithms. We define healthcare organisation as any that oversees patient pathways, which may include any public or private provider, provider networks or collaborations, or payers (for example in population health use-cases). The definition may be extended to companies that are contracted to operate clinical pathway functions. CHARM will retain relevance for continuous deployment of a single model, or where organisations scale to many simultaneous deployments.

CHARM places emphasis on two areas: Actionability, and Responsibility. The former refers to guidelines that are practical for organisations to incorporate into local processes. We aim to strike a balance between process necessity, and allowing adaptation for local context and digital maturity. This recognises that organisations may need to achieve a certain level of data and digital maturity to scale AI successfully, but that best practices can always be introduced to safeguard deployments regardless of digital circumstance.

'Responsible AI' in healthcare describes development and usage of AI in a bio-medically ethical manner[7,23–25]. This considers impact on safety (nonmaleficence), effectiveness (beneficence), fairness (justice), and preserving privacy and patient-centred care (autonomy). This provides an

essential starting point for adapting MLOps to clinical use-cases, and we aim to incorporate (at a minimum) a set of key topics that have existing consensus on their importance within a responsible AI framework. These include: (1) interrogation and mitigation of biases within data; (2) calibration of models across subpopulations; (3) model interpretability; (4) interrogation and mitigation of biases in predictions and intervention allocation; (5) processes for detecting and addressing drift; (6) processes for monitoring both intended and unintended consequences; and (7) transparency and auditability across product lifecycle.

Method

While CHARM is not intended for research reporting, our method is adapted from the framework presented by the EQUATOR network[26], including stages of literature review, Delphi consensus panel, and final consensus meeting. We make use of industry expertise and non-academic literature in non-healthcare fields where the routine operationalisation of AI is more mature. This protocol therefore outlines four stages: (1) identify accepted stages and processes of MLOps, including specific considerations for healthcare, through literature review and expert input; (2) achieve consensus through Delphi on important clinical MLOps process items; (3) approve a final checklist that is actionable for healthcare organisations; (4) test implementation of processes within a real-world AI platform in the

National Health Service (NHS). Study stages and timeline are shown in *Figure 1*.

The work is organised across study investigators who are responsible for protocol design, literature review, and study coordination, and the CHARM expert working group who are providing expert guidance in checklist and MLOps design. An international consensus panel will be recruited to participate in a modified Delphi study.

Expert working group

The expert working group consists of key stakeholders with experience of active or evolving AI implementations, or with expertise in governance or responsible AI, who can bring domain expertise and real-world experience into checklist design. As such, the team includes clinical or academic scientists involved in deploying models to clinical pathways; industry experts involved in technical design, governance, or implementation of AI pathways; health service directors who procure or manage AI models across their organisation; experts in policy, ethics, or responsible AI; regulatory representatives; and end-users of clinical AI products (including patient representatives).

We include experts who work outside of traditional academic and publishing contexts, to take advantage of substantial experience derived from real-world projects. Public and healthcare provider AI expertise are presently represented by collaborators from across six continents, including those from low to low-middle income countries. The group also

includes experts from non-healthcare settings with substantial experience of active MLOps and responsible AI governance implementations, as well as healthcare AI companies with active clinical deployments in high and low resource areas. Implementations run by group members represent multiple data modalities, including in clinical decision support, radiomics, biomarkers, and natural language processing. The expert group and affiliations will be named collaborators on any published materials.

This group will contribute to stages 1 and 3 as described, including guiding synthesis of proposed MLOps guidelines prior to the Delphi stage, and approval of the final checklist. A subset of the working group and study investigators with appropriate organisation and data governance approvals will undertake implementation as described in stage 4.

Delphi panellists & recruitment

We aim for the Delphi panel to fulfil two broad consensus functions in stage 2: actionability of each guideline item for organisations, and importance of each item for maintaining responsible AI. International panellists will therefore be invited from two broad groups. The first are organisation digital or data leaders, who would hold potential responsibility for directing or overseeing any organisation-wide technology introduction. Organisation leadership may not have individual experience with AI, but would have access to, or would be responsible for managing, individual experts. The second group are wider cross-sector clinical AI stakeholders, including scientists, clinicians, or industry employees involved in AI development or

deployment, regulators involved in creating guidance or assessing AI products, academics or practitioners of AI policy, law, and ethics, and end-users of AI tools including clinicians and payers.

The panel will be recruited through professional networks of the investigators and the expert working group, and via further snowball recruitment if minimum numbers are not achieved. We aim to recruit at least 20 healthcare organisation leaders, and more than 100 wider stakeholders over a recruitment window of 2 months.

Panellists will be invited through email with direction to an online consent form, and a website with documentation and interactive visual explanation of study rationale and the Delphi process. Participants are given the option of contacting study investigators by email to obtain clarifications, or attending an online webinar to further discuss the study. The modified Delphi is conducted through an electronic survey platform with data stored in the UK, and with General Data Protection Regulation compliance. Data entry is made anonymously, and participants can withdraw consent at any time during the study. Individual panellist participation will be acknowledged as part of any published materials.

Patient involvement

Substantial preceding work in the form of focus groups and literature review has been performed to map opinions and public concerns in AI development, deployment, and usage. These findings will be directly input into the checklist design process. Patient representatives in the expert

working group have substantial experience of advocacy in healthcare data, and understanding of digital pathways, and will continue to play a key role in checklist development throughout the study. The CHARM website (https://charmforhealth.com) is designed to facilitate public engagement and education, with lay and visual explanations of MLOps, responsible AI, and the study process. This protocol is submitted for open peer review to allow maximal and transparent engagement.

Defining MLOps

Scoping literature reviews are on-going to synthesise descriptions of MLOps processes in current practice. A systematic search of peer-reviewed healthcare publications was first performed using PubMed, EMBASE, and pre-print databases to discover existing definitions of healthcare MLOps, frameworks for AI implementation, or case studies of real-world implementation. Due to the paucity of detailed healthcare MLOps literature, we are conducting a parallel review through 'gray' literature and multimedia sources including textbooks, white papers, industry articles, recorded webinars, and peer-reviewed and preprint non-healthcare publications. General searches for written materials are conducted through Google Scholar and Google Books. Targeted searches were also performed for materials published by authoritative information sources, including Gartner, and top cloud technology supplier white papers. The review process is being conducted by JM, JG, JZ, and HA.

We first synthesise key processes that have existing consensus around inclusion in any general (non-healthcare) MLOps framework. This is conducted with assistance from technical practitioners and governance experts in the working group. The expert working group will then modify and add process items to improve clinical relevance, to establish a set of baseline items for consideration in a Delphi study. Facilitated by the study investigators, this process will be conducted through asynchronous online collaboration using an online 'whiteboard', and through focussed small group meetings.

To maximise transparency outside of academic publications, we plan to make public a scoping document that summarises review findings, as well as summary outputs of expert group discussions. These documents will be released on the website.

Modified Delphi study

We adopt a modified Delphi methodology[27] through electronic surveys, allowing easy dissemination, response collection, and a lower barrier to participation for globally situated panellists. This methodology saw increased use during the COVID-19 pandemic and has been termed an 'e-Delphi'. We will conduct a maximum of three Delphi rounds, facilitated by JZ, JM, and JG.

Following closure of the recruitment window, a first-round survey will be made available online, and links will be emailed to each panellist. The

survey will be kept open for 4 weeks, and a reminder will be sent at the end of week 3. This same timeline will follow for subsequent rounds.

In each survey round, panellists are asked to score each candidate item along 5-point Likert scales (Strongly Disagree to Strongly Agree) across two statements: "*this process is actionable by healthcare organisations*" and "*this process is important for responsible AI deployment*". Items are presented alongside guidance notes that outline any accompanying rationale. We ask panellists to record their professional role group at the start of each survey (as organisation leader, or clinical AI stakeholder - no other identifiable information is collected). We define consensus threshold as >=75% with a modification: in addition to overall consensus for each item, actionability scores must reach consensus in the organisation leader group, and importance scores must reach consensus in the clinical AI stakeholder group. Whitespace is provided to comment on process guidelines, or suggest additional guideline items.

After Round 1, items reaching consensus are reserved and removed from the Delphi process. Items that do not meet consensus and do not demonstrate benefit from revision are removed from the study. Delphi round 2 therefore contains: (1) modified items that did not meet consensus in round 1, and have received revisions proposed from panellist comments; (2) new items suggested by panellists with agreement from expert working group members. A third Delphi round will only be conducted if the working

group deems consensus items to be insufficient for purposes of creating a final checklist.

Consensus meeting

A summary of Delphi rounds and resulting consensus items are published on the CHARM website and distributed to the expert working group and panellists who participated in all Delphi rounds.

Considering feedback received during protocol design, the final consensus and approval stage prioritises flexibility and participation opportunity. This considers the needs of members across global and low resource regions, and those with limited time that can be dedicated to prolonged multi-day discussion and voting sessions across detailed guidelines.

Participants firstly include members of the expert working group. Additional invitations will also be extended to Delphi panellists who represent key organisation leaders and stakeholders who will have responsibility for actioning any checklist.

The final consensus group will be given three weeks to review Delphi results in a collaborative whiteboard document, where asynchronous comments and discussion are welcomed. Group members will also have opportunity to forward comments for individual items separately and in confidence to the study investigators, who will note these comments anonymously against items. One week prior to a consensus meeting, all items will be put to an asynchronous virtual vote, with items retained if

>=75% support amongst participants is reached. Following this virtual vote, a consensus meeting will be conducted through videoconferencing with an agenda set in advance. Items that did not reach >=75% consensus will be individually discussed, modifications proposed, and voted on. Any items reaching >=75% consensus in this final vote will be retained. Depending on participant availability, multiple consensus meetings may take place to discuss different item sets.

This methodology has two distinct advantages. First, an extended, asynchronous, collaborative stage allows maximal participation across a globally diverse and cross-sector group. Second, a first stage of asynchronous voting removes the necessity to discuss individual items with high group consensus in a face-to-face session. Finally, more focus can be given to process items with split votes that may benefit from detailed synchronous discussion and modification.

Publication & dissemination

CHARM will be made publicly available on a study website. We aim for process items to be individually explained and accompanied by detailed guidance for execution and governance. Development will be reported in peer-reviewed literature. We aim to present CHARM in academic conferences, and in healthcare provider and industry events.

Implementation

The final checklist will be implemented within a real-world transformation project for developing an enterprise clinical AI solution for the NHS London region (investigators and expert working group members: GC, PB, JZ, JTT). Data and operational infrastructure are owned by the North-East London commissioning service. Processes and teams will be actioned to deploy a population health AI model developed and validated in a separate process, initially on a regional population of 2.2 million patients. The outcome of this implementation exercise will be reported and published separately.

Ethics

Ethics approval is obtained through the Central University Research Ethics Committee (CUREC) at the University of Oxford, and assigned study number: [].

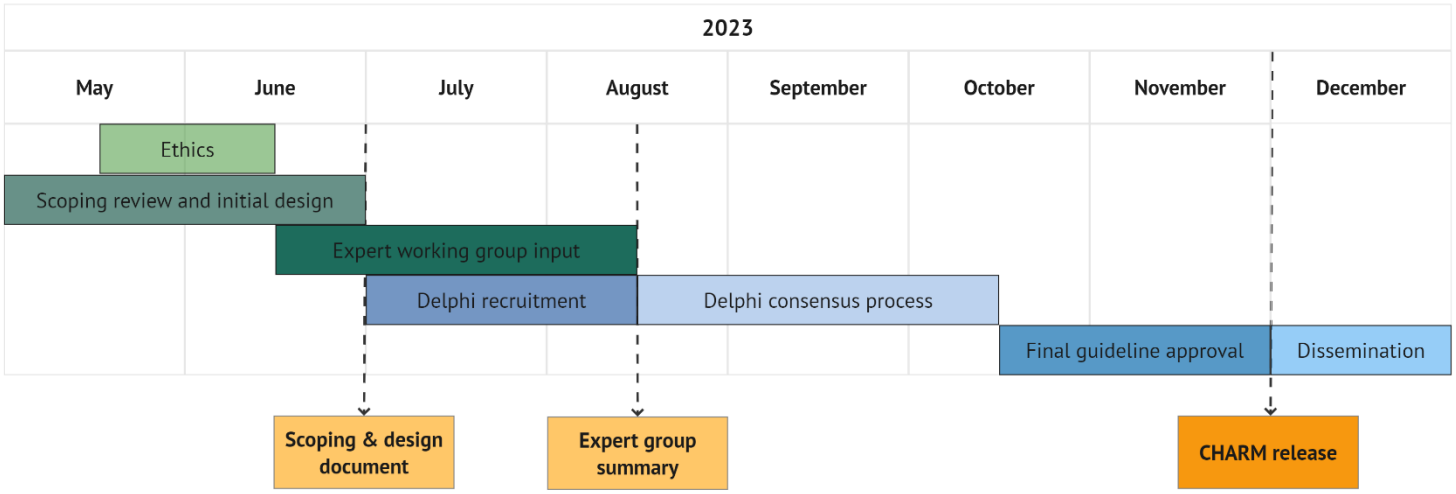

Figure 1 *– planned timeline for CHARM, indicating anticipated release timing for reports and guidelines.*

References

1. Zhang, J. *et al.* An interactive dashboard to track themes, development maturity, and global equity in clinical artificial intelligence research. *Lancet Digit. Health* **4**, e212–e213 (2022).
2. Madai, V. I. & Higgins, D. C. Artificial Intelligence in Healthcare: Lost In Translation? *ArXiv210713454 Cs* (2021).
3. Zhang, J. *et al.* Moving towards vertically integrated artificial intelligence development. *Npj Digit. Med.* **5**, 143 (2022).
4. AIX-COVNET *et al.* Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nat. Mach. Intell.* **3**, 199–217 (2021).
5. Reddy, S. *et al.* Evaluation framework to guide implementation of AI systems into healthcare settings. *BMJ Health Care Inform.* **28**, e100444 (2021).
6. Badal, K., Lee, C. M. & Esserman, L. J. Guiding principles for the responsible development of artificial intelligence tools for healthcare. *Commun. Med.* **3**, 47 (2023).
7. Chen, I. Y. *et al.* Ethical Machine Learning in Healthcare. *Annu. Rev. Biomed. Data Sci.* **4**, 123–144 (2021).
8. Panch, T. *et al.* A distributed approach to the regulation of clinical AI. *PLOS Digit. Health* **1**, e0000040 (2022).
9. Wu, E. *et al.* How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals. *Nat. Med.* **27**, 582–584 (2021).

10. Gallifant, J., Nakayama, L. F., Gichoya, J. W., Pierce, R. & Celi, L. A. Equity should be fundamental to the emergence of innovation. *PLOS Digit. Health* **2**, e0000224 (2023).
11. Benjamens, S., Dhunnoo, P. & Meskó, B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. *Npj Digit. Med.* **3**, 118 (2020).
12. Zhang, J. *et al.* Addressing the "elephant in the room" of AI clinical decision support through organisation-level regulation. *PLOS Digit. Health* **1**, e0000111 (2022).
13. Mark McCarthy. AI developers should build robust change control protocols despite absence of FDA guidance. (2022).
14. Heinz-Uwe Dettling. How the challenge of regulating AI in healthcare is escalating. (2021).
15. Gerke, S., Babic, B., Evgeniou, T. & Cohen, I. G. The need for a system view to regulate artificial intelligence/machine learning-based software as medical device. *Npj Digit. Med.* **3**, 53 (2020).
16. Kreuzberger, D., Kühl, N. & Hirschl, S. Machine Learning Operations (MLOps): Overview, Definition, and Architecture. *IEEE Access* **11**, 31866–31879 (2023).
17. Amodei, D. *et al.* Concrete Problems in AI Safety. *ArXiv160606565 Cs* (2016).
18. Lyell, D., Wang, Y., Coiera, E. & Magrabi, F. More than algorithms: an analysis of safety events involving ML-enabled medical devices

reported to the FDA. *J. Am. Med. Inform. Assoc.* ocad065 (2023) doi:10.1093/jamia/ocad065.

19. Sauer, C. M. *et al.* Leveraging electronic health records for data science: common pitfalls and how to avoid them. *Lancet Digit. Health* **4**, e893–e898 (2022).
20. Sounderajah, V. *et al.* Developing a reporting guideline for artificial intelligence-centred diagnostic test accuracy studies: the STARD-AI protocol. *BMJ Open* **11**, e047709 (2021).
21. Collins, G. S. *et al.* Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. *BMJ Open* **11**, e048008 (2021).
22. Vasey, B. *et al.* Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *BMJ* e070904 (2022) doi:10.1136/bmj-2022-070904.
23. Wiens, J. *et al.* Do no harm: a roadmap for responsible machine learning for health care. *Nat. Med.* **25**, 1337–1340 (2019).
24. Morley, J. *et al.* Ethics as a Service: A Pragmatic Operationalisation of AI Ethics. *Minds Mach.* **31**, 239–256 (2021).
25. Morley, J. *et al.* Operationalising AI ethics: barriers, enablers and next steps. *AI Soc.* **38**, 411–423 (2023).
26. Moher, D., Schulz, K. F., Simera, I. & Altman, D. G. Guidance for Developers of Health Research Reporting Guidelines. *PLoS Med.* **7**, e1000217 (2010).

27. Niederberger, M. & Spranger, J. Delphi Technique in Health Sciences: A Map. *Front. Public Health* **8**, 457 (2020).