Identifying Ethical Considerations for Machine Learning Healthcare Applications

Authors: Danton S. Char, Michael D. Abràmoff & Chris Feudtner

Year of publication: 26th Oct 2020

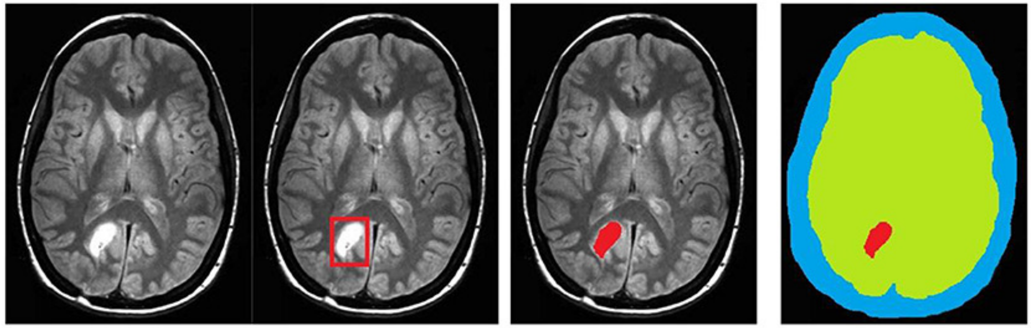
Overview

- ABSTRACT
- CHALLENGES TO IDENTIFYING ETHICAL CONSIDERATIONS
- PIPELINE FRAMEWORK TO IDENTIFY ETHICAL CONSIDERATIONS
- USING THE PIPELINE FRAMEWORK
- CONCLUSION

ABSTRACT

Along with potential benefits to healthcare delivery, machine learning healthcare applications(ML-HCAs) raise a number of ethical concerns.

Example: Machine Learning algorithms can be used in Medical imaging (such as X-Rays/ MRI scans) using Pattern Recognition to look for patterns that indicate a particular disease.



Classification Detection Instance Segmentation Semantic Segmentation

Ethical evaluations of ML-HCAs will need to structure the overall problem of evaluating these technologies, especially for a diverse group of stakeholders. IMPLEMENTATION OF ML-HCAS, AND THE PARALLEL
This paper outlines a systematic approach to identifying ML-HCA ethical concerns, starting with a conceptual model of the pipeline of the CONCEPTION, DEVELOPMENT, PIPELINE OF EVALUATION AND OVERSIGHT TASKS A TEACH STAGE.

Over this model we layer key questions that raise value-based issues, along with ethical considerations identified in large part by a literature review, but also identifying some ethical considerations that have yet to receive attention.

This pipeline model framework will be useful for systematic ethical appraisals of ML-HCA from development through implementation, and for interdisciplinary collaboration of diverse stakeholders that will be required to understand and subsequently manage the ethical implications of ML-HCAs.

CHALLENGES TO IDENTIFYING ETHICAL CONSIDERATIONS

Before laying out the pipeline model, we need to clarify five significant challenges to identifying ethical considerations arising from ML-HCAs design, implementation, and evaluations, as any approach to the identification task should be designed to meet these challenges.

1. UNCERTAIN IMPACT OF EMERGING TECHNOLOGIES

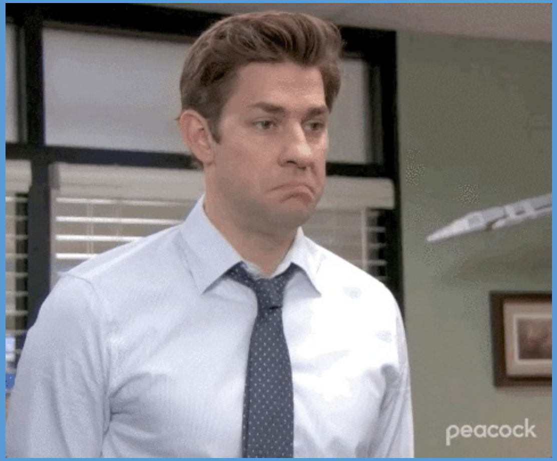
2. ML & AI EXCEPTIONALISM

3. BREADTH OF APPLICATIONS

4. ALLURE OF HIGHLY RESTRICTED FOCUS

5. DIVERSE STAKEHOLDERS

UNCERTAIN IMPACT OF EMERGING TECHNOLOGIES



ML-HCAs, like all new technologies, present uncertainty regarding their future impact. Ethical frame-works that focus on articulating guiding principles without first systematically identifying potential problems do not specifically address this uncertainty.

While various conceptual frameworks have been proposed to guide anticipatory ethical analyses of emerging technologies or to ascertain the values inherent in design approaches, a common general feature of these methods is the importance of having a systematic guided by an underlying evaluative framework to identify key considerations across as full a range as is possible of potential impacts

ML and AI Exceptionalism

As advanced as ML-HCAs are, built with cutting edge technology, no sound reason as yet exists to believe that the health applications powered by ML are, in and of themselves, exceptional. The clinical applications all seek to perform, in novel and hopefully better ways, standard healthcare tasks, such as diagnosis, generating a prognosis, or assisting with treatment decision-making. These tasks each have already identified ethical considerations that likely apply to ML-HCAs. The technology itself is also built from essentially standard clinical information, such as patient demographic or clinical information, such as laboratory values or diagnostic images, and while this information is being analyzed in remarkable ways, standard ethical considerations about these data also likely apply to ML-HCAs. Accordingly, a framework to guide identifying ethical considerations does not need to be focused on exceptions, even as it should leave space for exceptional considerations to be identified

Breadth of Applications

The breadth of emerging ML-HCAs, regarding what they aim to do, how they are constructed, and where they are being applied, is remarkably broad.

ML-HCAs range from fully autonomous artificial intelligence diagnosis of diabetic retinopathy in primary care settings to non-autonomous mortality predictions to guide insurance and allocation of healthcare resources.

The analytic framework guiding the identification of ethical considerations should therefore ideally be sufficiently generic to be useful across a wide variety of ML-HCAs.

For the ethical appraisal of any given ML-HCA, detailed content and context-specific knowledge will always be needed to provide more thorough and precise ethical evaluation, and this will require cross-disciplinary collaborations.

A framework for identification of ethical considerations, one that can accommodate a broad range of ML-HCAs, would help such collaboration.

Allure of Highly Restricted Focus

Many ML-HCA computer scientists have already turned away from ethical analysis as unworkable or not adequately responsive to ongoing ML-HCA development, have instead focused exclusively on the ethical consideration of fairness and emerging concerns regarding bias, and have begun to pursue an ideal of "algorithmic fairness," or the ability to computationally demonstrate a lack of between-group bias with an ML application.

They reason that if latent biases can be identified, ML approaches might be used to correct for them or improve "fairness" Highly focused approaches such as this assume an a priori comprehensive understanding of where and why such biases are occurring; if this assumption is wrong, these approaches risk introducing a complex set of unintended biases in attempts to correct the initial bias

More generally, a highly restrictive focus and limited framework may be applicable for ultimately addressing a specific ethical consideration and set of concerns, but will not suffice to manage the uncertainty regarding other potential ethical considerations.

Diverse Stakeholders

Finally, ML-HCAs are likely to have a broad range of stakeholders, from patients and health care practitioners, to computer scientists, engineers, and entrepreneurial developers, to healthcare organizations and payers, to oversight bodies charged with regulating medical practice. Any framework to help identify ethical considerations should provide for potential perspectives and concerns of each of these diverse stakeholders, commensurate with their expertise

PIPELINE FRAMEWORK TO IDENTIFY ETHICAL CONSIDERATIONS

1. Conception: Auditability, Transparency Standards, and Conflicts of Interest

2. Development: Perpetuation of Bias within Training Data, Risk of Harm Due to Group Membership, and Obtaining Training Data

3. Calibration: Accuracy, Trading off Test Characteristics, and Calibrated Risk of Harm

4. Implementation, Evaluation, and Oversight: Adverse Events, Ongoing Assessment of Accuracy and Usage

Conception: Auditability, Transparency Standards, and Conflicts of Interest

- When designers and implementers of a ML-HCA clearly declare the intentions, indications for use, and goals for an application, clinicians, patients, regulators, and other stakeholders are better enabled to exercise their own evaluative and decisional autonomy. Without transparency about intentions or specific goals, stakeholders will not be able to decide for themselves whether they want to support these intentions, or whether they believe that the ML-HCA will advance these intentions and the stated goals
- Stakeholders do not need to understand in detail the inner working of an ML-HCA in order to achieve "auditability."
- To support evaluative autonomy, transparency will require "auditability": ML systems in medicine must have an explainable architecture, designed to align with human cognitive decision-making processes familiar to physicians, and directly tied to clinical evidence.
- A simple but key aspect of determining the safety of any healthcare application depends upon the ability to inspect the application—to literally disassemble and examine a physical application to determine how the parts work together, to see the mechanisms at work, and thus better understand how the application might fail.

- The process is similar for software applications and, by analogy, to the components and physiologic mechanisms of medications or mechanical devices. ML-HCAs, however, can present a “black box” problem, with workings that are not inspectable by evaluators, clinicians, and patients.
- Transparency standards should also clarify whether a ML-HCA is “locked” or “continuously learning.” Continuous learning ML-HCAs automatically update using inputs during use, as opposed to locked ML-HCAs, which are deterministic.
- Some have argued that continuous ML learning in healthcare contexts may be harmful.
- With continuous learning, “distributional shift” can occur, if target training data does not match ongoing patient data. Leading an ML-HCA to begin to draw inaccurate conclusions.
- Transparency standards should also specify whether a ML-HCA is assistive or autonomous.
- Last but not least, with growing understanding that mores and values can intentionally or unintentionally become embedded in the design of engineered systems transparency will be required regarding any potential conflicts of interest.

Development: Perpetuation of Bias within Training Data, Risk of Harm Due to Group Membership, and Obtaining Training Data

- An important and acknowledged concern in the development of ML-HCAs relates to the possibility of bias, particularly whether latent biases in training data may be perpetuated or even amplified.
- Examples already exist of predictive scores failing both because of poorly composed training data and because, when expanded to broader populations, racially discriminatory outcomes occurred.
- For example, ML programs designed to aid judges in sentencing by predicting an offender’s risk for recidivism have shown a disturbing propensity for racial discrimination.
- Furthermore, any perpetuated biases incorporated into a ML-HCA may subsequently impact clinical decisions and support self-fulfilling prophesies.
- For example, if clinicians currently routinely de-escalate or withhold interventions in patients with specific severe injuries or progressive conditions, ML systems may classify such clinical scenarios as nearly always fatal, and any ML-HCA built on such a classification would likely result in an even higher likelihood of de-escalation or withholding, thereby reducing the opportunity to improve outcomes for such conditions
- Training of ML-HCAs against real world data, rather than high-quality research-grade data, may simply perpetuate sub-optimal clinical practices that are not aligned with the best scientific evidence. Conversely, an algorithm’s over-reliance on research-grade data alone may miss important clinically relevant sources of knowledge, lowering the quality of care delivered.

- A related concern is obtaining needed training data, and questions of data ownership, pricing and protecting privacy.
- Machine learning requires large amounts of training data. The aggregation and curation of these large datasets raises not only issues regarding specifying the standards that high-quality reference standard data must achieve, but also issues regarding data privacy and data ownership.
- ML-HCAs may be based on data from non-clinical sources (such as personal devices, social media, financial, or legal sources),which may contain potentially controversial data elements or have been collected via novel means that we cannot foresee.
- There has also been ongoing patient activism for inclusion in recognition for specimen contribution to scientific advances

Calibration: Accuracy, Trading off Test Characteristics, and Calibrated Risk of Harm

- In order for a ML-HCA to maximize clinical benefits and minimize harm, the application must perform in accordance with the cardinal design features of safety(to prevent injuries and hazards), efficiency (that the application effectively solves the problem it was designed for and does so at a reasonable cost, in particular regarding the costs of incorrect classifications, such as false negative or false positive diagnoses), and equity (that the advantages of the application are shared fairly by all).
- In concrete terms, this means at a minimum that the application will need to provide accurate diagnostic or predictive information on the vast majority of patients for whom the ML-HCA is intended to be used, irrespective of subgroup such as age or race.
- Determining the accuracy of a ML-HCA is, how-ever, not straightforward. Unlike ML designed for other contexts, such as to play games of skill (e.g. chess, go), many medical decisions and diagnoses can-not be perfectly labeled as correct or incorrect and down-stream outcomes cannot always be anticipated. This is a known challenge with reference “gold standards” in healthcare
- While ML accuracy can be higher than that of individual experts in interpretation of clinical images such as radiologic scans, pathology slides, and photo-graphs of skin lesions the estimated accuracy of a ML-HCA is dependent on the clinical context in which the application is being assessed
- Validation studies therefore need to be done not only in the context of rigorously managed research trials, but also in general populations of patients.

- An equitable ML-HCA will provide equivalent levels of accuracy within the intended-use population across multiple patient subgroups or characteristics, and also achieve equivalent levels of "determinability," or the ability of the ML-HCA to provide a clinically relevant output based on the clinically available inputs (and not simply declare that the inputted information is not sufficient).
- The notion of accuracy in an ML-HCA, inherently involves tradeoffs between test characteristics, guided by designer value judgments with consequent ethical implications.
- Even if a specific ML-HCA is found to be superior to an established clinical practice with regard to all test characteristics, that specific ML-HCA will have calibrated not only greater accuracy, but also specific forms of inaccuracy: the design will predictably generate false positives and false negatives, or indeterminate results, as must be the case with any method of classification, whether based on human judgment or machine learning. The key ethical consideration would be whether these inaccuracies (and any consequent harms) are outweighed by potential benefits and dis-tributed among patients in an equitable manner.

Implementation, Evaluation, and Oversight: Adverse Events, Ongoing Assessment of Accuracy and Usage

- During development, when ML systems may be vali-dated on idealized data, their accuracy may be measured to be “perfect” (in other words, not statistically different from a perfect algorithm or observer who always outputs the true state of disease). But in real-world settings—where there is the potential for human operator error, data inputs of lower quality and nearly infinite variance, and additional potentially relevant data captured in a modality not accessible to the ML-HCA—the true accuracy is typically lower, even when the underlying ML-HCA has been locked and unchanged.
- As the measured sensitivity, specificity, and determinability change, so too will the potential benefits and potential harms, and the resulting benefit-to-harm ratio.
- For example, earlier computer-aided diagnostic tools such as EKG interpretation and mammography appeared in preliminary studies to offer value-adding diagnostic accuracy, yet in subsequent evaluations of their actual intended use (specifically, to assist front-line clinicians in making medical decisions) have failed to demonstrate benefit and raised the possibility of some degree of harm.

- Unintended uses of a ML-HCA, with new potential harms as well as any hoped-for benefits, will also need to be monitored. Some potential unintended uses maybe predictable before implementation (such as a ML system for mortality prediction being co-opted to limit hospital mortality statistics or costs). Assuring that a ML system is not being inadvertently yet inappropriately re-purposed will also require ongoing monitoring. For example, a system intended for diagnosis of diabetic retinopathy might be co-opted (or unintentionally interpreted by patients or health providers) as an ophthalmic screening exam for broader conditions than just diabetic retinopathy.
- Lastly, based on experiences with the implementation of electronic medical record platforms, monitoring will also be warranted to assess the equity of access to ML-HCA, which may be more readily available in larger or better financed health systems than in small systems or practices, which in turn could result in poorer outcomes in these smaller sites.

USING THE PIPELINE FRAMEWORK

Now that we have laid out the framework of a pipeline model of ML-HCAs, let us outline how the framework can be used for the purpose of ethical analysis.

As the model makes clear, there are many potential points in the ML-HCA pipeline where an individual or a group might want to identify and think through ethical considerations that arise specifically at that point in the overall pipeline. The questions posed in the framework for a given stage of the pipeline may help in identifying other, novel considerations.

The framework also should be used, even when focused on a particular point in the pipeline, to identify and examine ethical considerations in previous steps. ML-HCA developers and users poised at a particular point in the pipeline inherit the ethical operating characteristics that arise from previous decisions about how the ML-HCA has been constructed.

Identification of potential future consequences can aid ethical evaluation and decisions regarding design, development, implementation, and evaluation.

As mentioned above, these activities can be done by individuals or groups, in particular multi-stakeholder groups.

The pipeline framework also offers groups of diverse stakeholders a “bigger picture” of ML-HCAs that can, with dialogue, help to forge a shared mental model of the range of relevant questions and ethical considerations that should guide design and evaluation decisions. The broadness of the framework will help combat any tendency to focus narrowly on one ethical consideration while potentially neglecting other relevant considerations and thus sidestepping grappling with tradeoffs.

Lastly, the common basic elements of the pipeline —an application is conceived of, developed, calibrated, implemented, and evaluated, with various forms of oversight—allows for ready comparison of the ML-HCA pipeline to the pipelines of other medical technologies, and to see that while ML-HCAs do raise some novel issues, they also raise many issues common to existing diagnostic or therapeutic technologies. This can put a check on unwarranted ML-HCA exceptionalism in our thinking about the ethics of this emerging technology.

CONCLUSION

Machine learning in healthcare has arrived. Along with many potential benefits to healthcare delivery, ML-HCA is likely to raise complex and as yet only partially considered ethical considerations with implementation. The pipeline framework, starting with a map of the conception, development, implementation, and the parallel evaluation and oversight tasks of ML-HCAs, and then layering over this map key questions, value-based issues, and ethical considerations, is an approach for systematically identifying these ethical considerations and for facilitating inter-disciplinary dialogue and collaboration to better understand and subsequently manage the ethical implications of ML-HCAs.

THANK YOU

PLEASE FEEL FREE TO ASK ANY QUESTIONS.