Ethical Considerations for Artificial Intelligence in Medical Imaging: Data Collection, Development, and Evaluation

Jonathan Herington1, Melissa D. McCradden2, Kathleen Creel3, Ronald Boellaard4, Elizabeth C. Jones5, Abhinav K. Jha6, Arman Rahmim7, Peter J.H. Scott8, John J. Sunderland9, Richard L. Wahl10, Sven Zuehlsdorff11, and Babak Saboury2

1Department of Health Humanities and Bioethics and Department of Philosophy, University of Rochester, Rochester, New York; 2Department of Bioethics, Hospital for Sick Children, Toronto and Dana Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada; 3Department of Philosophy and Religion and Khoury College of Computer Sciences, Northeastern University, Boston, Massachusetts; 4Department of Radiology and Nuclear Medicine, Cancer Centre Amsterdam, Amsterdam University Medical Centres, Amsterdam, The Netherlands; 5Department of Radiology and Imaging Sciences, Clinical Center, National Institutes of Health, Bethesda, Maryland; 6Department of Biomedical Engineering and Mallinckrodt Institute of Radiology, Washington University in St. Louis, St. Louis, Missouri; 7Departments of Radiology and Physics, University of British Columbia, Vancouver, British Columbia, Canada; 8Department of Radiology, University of Michigan Medical School, Ann Arbor, Michigan; 9Departments of Radiology and Physics, University of Iowa, Iowa City, Iowa; 10Mallinckrodt Institute of Radiology, Washington University in St. Louis, St. Louis, Missouri; and 11Siemens Medical Solutions USA, Inc., Hoffman Estates, Illinois

The development of artificial intelligence (AI) within nuclear imaging involves several ethically fraught components at different stages of the machine learning pipeline, including during data collection, model training and validation, and clinical use. Drawing on the traditional principles of medical and research ethics, and highlighting the need to ensure health justice, the AI task force of the Society of Nuclear Medicine and Molecular Imaging has identified 4 major ethical risks: privacy of data subjects, data quality and model efficacy, fairness toward marginalized populations, and transparency of clinical performance. We provide preliminary recommendations to developers of AI-driven medical devices for mitigating the impact of these risks on patients and populations.

Key Words: AI ethics; AI as medical device; software as medical device; health disparity; socioeconomic determinants of health

J Nucl Med 2023; 00:1–7
DOI: 10.2967/jnumed.123.266080

Artificial intelligence (AI) and machine learning systems are an exciting area of research in nuclear medicine and molecular imaging. The development of artificially intelligent medical devices (AIMDs) promises to improve the accuracy of diagnoses, expedite treatments, drive down costs, and improve patient outcomes. However, the development of these AIMDs raises several ethical challenges, including the privacy of data subjects, the risk of unintended harm to patients, and fairness to marginalized populations (1,2). Grappling with these ethical issues is essential for researchers and developers of AIMDs before their widespread adoption in nuclear medicine and molecular imaging.

The AI task force of the Society of Nuclear Medicine and Molecular Imaging has set out to make clear the assignments of responsibility between developers, physicians, and regulators by distinguishing between the development and deployment of AIMDs. This article focuses on the ethics of developing AIMDs for applications in nuclear medicine and is structured around 3 different phases in the AIMD pipeline (Table 1). In section 1, we discuss the duties of researchers during the data collection phase, including duties owed to data subjects, as well as data quality and bias issues. In section 2, we discuss the ethical considerations during the training and validation of the AI-based tool, including designing for safety, efficacy, interpretability, and fairness. Finally, we discuss the ethical imperative for appropriate and transparent evaluation of the AIMD. In a companion piece (3), we discuss the ethical obligations of clinicians and regulators during the deployment of an AIMD.

Researchers in nuclear medicine will be familiar with human subject research protections and their underlying principles of respect for persons (i.e., autonomy), beneficence, and justice (4–6). Others have built on this foundation to develop AIMD ethics frameworks that focus on the core ethical principles of autonomy, beneficence, nonmaleficence, justice, explicability, and transparency (2,7). Although these frameworks are well suited to governing individual doctor–patient relations, we argue that they are less useful for governing the development of AIMDs, where risks to different stakeholders (i.e., research participants, patients, and clinicians) can be embedded into a system at multiple places in the development pipeline (8). Instead, it may be more productive to view the development of AIMDs as a social contract between researchers, clinicians, patients, and the general public. The goal of such a contract is justice: to ensure that the benefits of AIMDs can be realized while respecting one another’s rights and fairly distributing the benefits and burdens of producing AIMDs (9). Thus, whereas traditional principles of clinical and research ethics may be useful as starting points, special emphasis should be placed on the promotion of health justice. Throughout this paper, we consider the ethical development of AIMDs across 3 different domains of ethical value: individual welfare (nonmaleficence and beneficence), individual autonomy, and social justice.
DATA COLLECTION

The foundation of AIMDs is the data that are used to train them. Ordinarily, the ethics of data collection focuses on the rights of the data subjects, including questions about consent and privacy. Importantly, however, attempts to respect the rights of data subjects must be balanced against the imperative that data be of high quality, tailored to different deployment contexts, and representative of diverse populations.

Obligations to Data Subjects

Researchers should, at minimum, be aware of their legal obligations with respect to data collected directly from subjects (10). This includes data collected directly from patients, from participants in a standalone research study, and from data annotators outside the study team (e.g., clinical experts or mechanical turkers). If directly collecting data from identifiable subjects, researchers must seek the approval of a research subject review board, which may require consent from each subject after informing them of the nature of the research. If in doubt, researchers should seek advice, since a failure to treat subjects in accordance with human subject research protections both is unethical and may require the destruction of datasets and AIMDs.

Table 1 shows the ethical dimensions of AIMDs according to data collection, development, and evaluation. Some entries denote legally or ethically required conduct; use of promote or avoid denotes ethically ideal conduct.

<table>
<thead>
<tr>
<th>Category</th>
<th>Data Collection</th>
<th>Development</th>
<th>Evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Welfare</td>
<td>Promote projects with high expected social benefit; avoid model designs and data collection with risk of reidentification</td>
<td>Promote generalizability of performance across clinical contexts</td>
<td>Ensure that clinical utility, not just in silico performance, is evaluated</td>
</tr>
<tr>
<td>Autonomy</td>
<td>Ensure consent to direct data collection; consider notification for secondary reuse of data</td>
<td>Avoid hard-encoding developers’ judgments about risk tradeoffs; ensure that algorithmic uncertainty is interpretable by clinicians</td>
<td>Ensure that performance and limitations are clearly declared to clinicians</td>
</tr>
<tr>
<td>Justice</td>
<td>Assess representativeness of data for task (especially by race, sex, etc.); promote projects that lead to broadly shared benefits</td>
<td>Carefully justify use of sensitive attributes (race, sex, etc.) as input features; avoid performance disparities by race, sex, etc.</td>
<td>Ensure that performance is evaluated for population subgroups (by race, sex, etc.)</td>
</tr>
</tbody>
</table>

Use of ensure denotes legally or ethically required conduct; use of promote or avoid denotes ethically ideal conduct.

Many datasets used in the development of an AIDM involve secondary analysis of deidentified data that were initially collected for other purposes. Clinical imaging datasets and electronic health records are commonly repurposed long after the original data were collected. Such datasets are exempted from most human subject research protections when personally identifiable information is removed (5,11). The permissiveness of current regulatory frameworks does not, however, diminish the ethical obligation to foster trust with respect to the secondary analysis of data. This requires considering permission for secondary analysis, subject privacy, and data quality.

Consent to Secondary Analysis. Obtaining explicit informed consent from every data subject for each specific reuse of deidentified data would be costly. Some have therefore argued that patients should be asked to provide broad consent to the use of their deidentified data for future secondary research projects (12). More radically, some argue that after deidentification there is no owner of the data and that consent thus should be replaced by an enhanced duty of care to ensure data use is appropriate (13). Although there is currently no settled view on the appropriate ethical standard to apply to secondary research, these proposals suggest 3 issues researchers should consider.

First, individual subjects should be notified that their data will be included in AIDM datasets. This notification may be accomplished by notification of the possibility of reuse during initial data collection (i.e., during medical treatment) or at the time of each specific reuse. When possible, notification should include a statement of the risks raised by the intended reuse and the opportunity to refuse or withdraw (14). When this is not possible, educational programs to raise public awareness regarding the use of anonymized medical data should be adopted (2). It is likely that a single mechanism is insufficient given that different data sources may be more or less ethically sensitive (e.g., cranial imagery or whole-body CT). Instead, multiple opportunities to notify data subjects will better support the trustworthiness of AIDM development.

Second, data subjects should be provided the opportunity to withdraw their data from research datasets. Currently, many medical imaging datasets are fully anonymized on collection, allowing
sharing of the dataset without human subject research oversight. Although this arguably promotes subject confidentiality, it also deprives subjects of control of their data, a question of particular concern when public datasets are used to create closed-source AIMDs. Though there may not be the opportunity to remove data from a trained model, some subjects may wish to remove their imagery from datasets to prevent further use. Providing opportunities for blanket withdrawal of data from secondary datasets is a key mechanism for demonstrating trustworthiness and working to repair damaged relationships with oppressed communities. The medical imaging research community should consider exploring models—such as those adopted in genomics and biobanking (15)—that allow for this kind of withdrawal of consent. These kinds of data repositories require infrastructure that securely maintains coded links between subject contact information and subjects’ deidentified data alongside staff to facilitate withdrawal of data from the repository. This is more costly than open access, permanently anonymized datasets that allow unrestricted access.

Third, in the absence of individual consent or withdrawal, researchers should develop other mechanisms to remain accountable to subject populations. A data subject advisory board may be formed to express concerns and exercise oversight related to the use of deidentified data. This may be particularly important in the context of data from vulnerable or historically marginalized populations. Marginalized communities may have a warranted mistrust of medical researchers because of serious historical abuses. Black and indigenous scholars have thus suggested enhanced ethical obligations with respect to the use of health data from marginalized populations, including seeking explicit consent for each subject or engaging in community consultation (16,17).

Privacy and Reidentification. Clinical data used for the development of AIMDs may contain sensitive personal health information. Nuclear medicine and molecular imaging generate datasets that include whole-body scans, diagnoses, and neurologic imaging—all of which may expose highly sensitive information about a patient’s sex, behavior, life expectancy, or insurability. Moreover, there are particular sensitivities surrounding data connected to behavioral or neurologic health. Creating a framework to guide approaches to security, deidentification, and privacy for nuclear medicine and molecular imaging data is thus necessary if we are to increase public trust in the use of these data for AIMD development (2,18). We highlight 3 strategies for increasing the privacy of data subject information.

First, whereas many AI-based approaches include patient demographic features as inputs into decisions, the inclusion of features that could directly or indirectly identify a subject in training data should be limited to cases for which there is a specific rationale for the use of that feature. Given the increasing integration of data mining technologies into our daily lives and the high dimensionality of AIMD datasets, caution should also be exercised with respect to indirect identifiers. For instance, the National Institutes of Health’s All of Us precision medicine database identifies a tier of “data elements that may not, in their own right, readily identify individual participants but that may increase the risk of unapproved re-identification” (19). Limiting the dimensionality of a dataset may thus be a key privacy safeguard. However, deidentification carries with it risks of degrading performance. For instance, quasiidentifiers such as patient weight raise reidentification risk (20) but enable calculations (e.g., SUVs) that are essential for making sense of molecular images. Although it is not always easy to determine a priori whether a feature will influence AIMD performance, if early proof-of-concept projects show that particularly sensitive features have a low degree of influence, then there is a strong case for excluding them from training datasets used in later work.

Second, when indirectly identifiable information (i.e., weight or CT of the head) is required, differential privacy techniques can be used to obscure patient-level identifiers while preserving population-level regularities (21). Differential privacy aims to remove the identifiability of individual subjects while preserving the underlying structure of the data. However, differential privacy may also involve performance trade-offs given that identifiable data may be necessary to ensure model utility in the clinical environment.

Finally, even when researchers use differential privacy techniques, the storage of large amounts of patient data—so-called data lakes—carries the risk of malicious intrusion. Although such data lakes promise to improve the speed and efficacy of AIMD development, they also offer tempting attack surfaces for malicious actors. One tool for minimizing this risk may be federated learning, whereby AIMDs can be trained on a federation of smaller datasets, without formally aggregating the datasets (22). A combination of federated learning and differential privacy techniques may help minimize the risks of reidentification and help build community trust in the secondary reuse of data.

Social Value and the Quality of Data

The clinical performance, fairness, and trustworthiness of AIMDs are ultimately dependent on the quality of the dataset used to train the model (7). Avoiding data pollution—whereby severe sampling, measurement, or population biases are introduced into datasets—is thus a fundamental responsibility of researchers seeking to design AIMDs (23). We suggest the following 3 considerations to reduce the risk of data pollution.

First, even before data are collected, the definition of input and target constructs can have a profound impact on the performance of an AIMD. For instance, it has been suggested that datasets that include heterogeneous behavioral definitions of psychiatric disorders may distort estimates of the effect size of neuropsychiatric biomarkers identified through functional MRI (23). Careful definitions of data features is especially important in the context of social categories—such as race, sex, and nationality—which often have very complex relationships with other patient-level variables (24).

Second, the use of datasets derived from clinical data should be approached cautiously (25). Without a research protocol governing collection, diagnoses and chart notes made by multiple clinicians may not exhibit interrater variability. If data from a single site are reanalyzed, the clinical population and procedures may introduce selection or measurement biases into the dataset. If data from multiple sites are aggregated, differences in clinical practices across sites—including diagnostic variability—may make data incomparable. Although data augmentation techniques can ameliorate these issues (26), if the construct variability, measurement bias, or selection bias of a dataset is too extreme, researchers should consider constructing a de novo research dataset for training the AIMD.

Third, datasets should be representative of the populations that the AIMD will serve. Structural barriers to health care mean Hispanic/Latinx and Black populations are frequently underrepresented in datasets used for AI training (27). Moreover, as populations in the United States, Canada, and the European Union become more diverse, historical research data used for algorithm development may fail to reflect current demographics (28). Both these issues create...
the risk of severe selection bias in training datasets. Although representative datasets do not guarantee algorithmic fairness, without explicit efforts to diversify datasets, AIMDs will perpetuate and widen health inequalities (29).

MODEL DEVELOPMENT

Once data have been collected, researchers and developers must responsibly use those data to train and evaluate AIMDs. In previous work, we identified best practices for the development of AIMDs to ensure that they are task-specific, are interpretable to users, and are generalizable to different populations (26). In what follows, we discuss the ethical justification for these best practices—including patient welfare, patient autonomy, and fairness for marginalized populations—and the complexities that arise when pursuing these ethical goals.

Safety and Efficacy

Developers have an obligation to develop AIMDs that are safe and effective for the contexts in which they will be used. This obligation is derived from the duties of nonmaleficence and beneficence that clinicians owe to their patients and suggests at least 2 considerations for developers.

First, design and training of the model must be task-specific and informed by the needs and expertise of domain experts. Often the goal of machine learning developers has been to produce models that satisfy highly abstracted performance criteria (e.g., root mean squared error for generating images and Dice scores for segmenting images) within a convenient test dataset. Although this kind of method development enables innovation at a fast pace and low cost, it may result in harmful errors once deployed in a clinical context (30). Safety and efficacy at the bedside are more likely to be promoted if AIMDs are developed with input from domain experts, high-quality datasets, and performance metrics that are tightly aligned to the specific clinical task (26).

Second, model development should promote the generalizability of the model to the full range of clinical contexts in which it is expected to be deployed. Although AI systems are predominantly developed and used in high-resource settings, developers should expect that models may be deployed in clinical contexts with fewer resources (31). Thus, algorithms would ideally be robust to a range of patient populations, hospital resources, and sociotechnical support systems (32). The effort and training data required to make truly generalizable models are not trivial, however, and there is likely a tradeoff between the speed of innovation and the generalizability of the resultant AIMDs. Therefore, as discussed in a prior AI task force paper on evaluation best practices (33), limitations on the performance of the AIMD for particular populations, imaging equipment, or tasks must be clearly specified.

User Autonomy

Developers also have an obligation to develop AIMDs that facilitate joint decision-making by patient and clinician, by ensuring that models are interpretable to clinicians in ways that facilitate informed conversations with patients. Interpretability or explainability refers to a cluster of techniques that aim to provide clinicians with an understanding of the salience of different inputs into the AIMD’s predictions and the degree of uncertainty with respect to those predictions. In a companion piece (3), we explore these strategies and note that—although of technical interest to developers—they often fail to provide clinicians with the right kinds of explanation (34) and may in fact worsen automation bias (35). Ethical patient care requires shared decision-making in which the AIMD output is but one of many sources of relevant information (36). Thus, rather than explainability, clinicians may simply require transparency around the intended uses and accuracy of an AIMD to fully inform patients of risks and benefits.

Fairness

Basic principles of procedural and substantive justice require that AIMDs treat subgroups within a population fairly. It is well established, however, that machine learning models in medicine can exhibit race, sex, or socioeconomic biases (18,37). The sources of these biases are complex and may stem from problem selection, target specification, sampling bias, model architecture, and structural bias reflected in training and input data (8). Two obligations of developers arise in this context: an obligation to ensure that AIMDs are procedurally fair and an obligation to ensure that AIMDs strive toward distributive fairness.

Procedural Fairness. Standard principles of medical ethics require that patients be treated with equal consideration, regardless of their race, sex, religion, or other protected characteristics (38). In the context of AIMDs, this is normally understood as a prohibition on the use of these features to make predictions or decisions about patients during deployment. This has led many developers to attempt to exclude race or sex as features within a training dataset. Two considerations suggest that ensuring procedural fairness requires more nuance.

First, the high dimensionality of AIMD input data, alongside the pervasive impact of race and sex on people’s lives, means almost all features will be partially correlated with race or sex. For instance, in the United States, race is correlated with residential address, insurance status, and prior access to PET/CT. Although procedural fairness may require identifying and removing close correlates of race and sex, identifying and eliminating all the features correlated with race or sex is likely to severely reduce accuracy. In this respect, some have suggested that rather than a focus on whether sensitive features cause predictions, procedural fairness requires that models be well calibrated for all racial and sex groups (39,40). In this view, ensuring that predictions have the same meaning (i.e., predictive value) for different subpopulations ensures that subsequent decision-making treats these individuals with equal consideration.

Second, explicit use of sensitive features during model training may be justifiable as a way of measuring or improving the performance of algorithms for marginalized groups. Historically, some medical decision-making tools have used features such as patient race or ethnicity as direct inputs. Although the use of race was initially rooted in racist biologic essentialism, the medical community’s understanding of the influence of race and ethnicity on diseases has progressed. Some now argue that race can act as a proxy for the influence of racist oppression on patient’s health, opening the possibility that race (or other sensitive attributes) can be used to improve predictive accuracy for these groups (41). Notably, however, the categorically driven capture of this information (e.g., patients identified as Black, White, Asian, or other) can often be imprecise and can in fact contribute to the reification of social stereotypes.

Distributive Fairness. Another potential concern is the differential impact of AIMD systems on marginalized populations. AIMDs have tremendous potential to mitigate health disparities in the United States; however, discrepancies in access and use of AIMDs in patients who differ by race or sex have raised concerns regarding AI contributions to health-care disparities (42). There are well-documented inequities in medical imaging studies on the
topics of coronavirus disease 2019, mammography, lung cancer screening, and missed care opportunities (43). For instance, privately insured Hispanic/Latinx and Black women are more likely to obtain mammography at facilities with less favorable characteristics, which subsequently correlates with higher breast cancer mortality rates in this population (44). Likewise, common training and test datasets for AIMDs (e.g., Medical Information Mart for Intensive Care, version IV) reflect societal biases in the diagnosis and treatment or patients—biases that can distort algorithms optimized to perform well in those test datasets (45). To underscore this issue, a recent study demonstrated a 4-fold discrepancy between Black and non-Hispanic White patients in their access to prostate cancer molecular imaging with 68Ga-PSMA-11 PET (46). Since PET is not immune to these challenges, neither are PET-based AI systems. We identify 3 points within the development pipeline at which technical or policy interventions may promote distributive fairness.

First, preprocessing strategies attempt to remediate biases introduced during the training phase. For instance, the development and characterization of standardized testing datasets that are more broadly representative than those available to single research teams is a high priority (e.g., the All of Us research program (15)). Such broadly representative datasets may limit sampling bias and societal biases that are localized to a particular study site. It is, however, fundamentally misleading to think that unbiased or debiased training sets will eliminate the potential for disparate impact. Even though preprocessing strategies are likely important considerations in the development of successful AIMD, they are not sufficient.

Second, inprocessing and postprocessing seek to remediate outcome disparities by fine-tuning model performance in different subgroups. This is often achieved by constraining the model to satisfy a statistical metric of fairness (i.e., demographic parity, error rate parity, or calibration parity). Although these metrics may sometimes help diagnose problems (39), they suffer 2 important limitations as methods for achieving fairness. First, they cannot be mutually satisfied under ordinary conditions (47) and thus require a value judgment about the appropriate metric to adopt in a given domain. Second, satisfying these metrics does not reliably reduce health disparities. For instance, in the famed study by Obermeyer et al. (39), the underlying inequality was generated by an inappropriate proxy for health risk (i.e., health expenditure), which would not show up as unfair on these tests. Thus, although failing to satisfy these fairness metrics may provide evidence of inappropriate bias, satisfying them is not sufficient to ensure that AIMDs reduce health disparities.

Finally, ensuring distributive fairness requires careful consideration of the context of deployment. For instance, an AIMD trained using imagery from a high-resolution PET/CT scanner may perform well in all demographic subgroups in a test dataset. Nonetheless, if it is with imagery derived from lower-resolution PET/CT scanners, it may exhibit higher error rates. This may increase disparities between high- and low-resource contexts despite a putatively fair algorithm. As we discuss in a companion piece (3), training, imaging infrastructure, and structural biases all contribute to the effect of an AIMD on health disparities.

**EVALUATION AND TRANSPARENCY**

In previous work, we identified best practices for the evaluation of AIMDs to ensure that they are evaluated on clinical tasks and that their performance is transparently declared for different clinical contexts and populations (33). In this section, we discuss the ethical foundations of these best practices.

**Clinical Evaluation**

AIMDs are part of complex clinical ecosystems, are fine-tuned to optimize very specific tasks, and often must team with humans across different organizations. Deploying AIMDs without evaluation in these complex clinical environments risks harming patients through inappropriate diagnosis or treatment, may waste resources by deploying AIMDs with low efficacy in real-world settings, undermines the ability of clinicians to judge the risks and benefits of interventions, and risks exacerbating health inequalities.

Recognizing this, McCradden et al. outlined 3 key phases of an evidence-based clinical validation process: algorithmic validation, silent trial testing, and prospective clinical evaluation (10). Algorithmic validation (outlined in 2.1.1 of their article) is necessary, but not sufficient, for deployment. A silent trial (also known as a shadow trial [silent period]) is critical to both scientific and ethical goals: it establishes whether the model can perform in a live environment and characterizes that performance; it provides the necessary ethical justification to proceed with prospective clinical testing. Prospective clinical evaluation can take many forms (e.g., from an observational trial to a randomized controlled trial); trial design decisions should be guided by ethical constraints and to generate the kind of knowledge the research team is looking to establish. This third phase of validation is where researchers can reliably establish facts about the model’s performance in a clinical workflow on specific outcomes. Evaluation at this last stage is crucial before widespread deployment of an AIMD in routine clinical imaging.

**Performance Transparency**

A lack of information about an AIMD’s performance and limitations undermines clinicians’ ability to provide informed medical care to their patients. Moreover, marginalized populations, especially communities that were historically discriminated against or victims of medical experimentations, may naturally be more hesitant to adopt AI applications in their medical care if their performance is not clearly communicated. Resistance to trust AIMDs can subsequently lead to hesitancy to seek medical care from imaging centers that have fully implemented AI technology into the workflow and patient care (48). These concerns recommend 3 different considerations for ensuring the transparency of AIMDs.

First, developers should make clear the intended use and clinically validated performance characteristics of the AIMD to all end users. Simply omitting performance information, or providing a standard caveat emptor disclaimer, is inadequate. For physicians to perform their responsibility to operate in the best interests of their patient, they must know that the AIMD performs accurately in the clinical context in which the physician is operating. Recent proposals to provide performance information have included the creation of model cards that define the provenance of the data used to train the model, key performance characteristics (e.g., area under the receiver-operating characteristic curve and ratios of false-positive rate to false-negative rate) within various populations, and known limits or biases (49). Importantly, as we identified above, whereas many AIMD systems are currently validated using only retrospective datasets, model cards for use in clinical contexts should provide evidence from clinical trials. Unfortunately, the fact that there is a paucity of AI trial reports conforming with established reporting standards (Consolidated Standards of Reporting Trials–Artificial Intelligence, Standard Protocol Items: Recommendations for Interventional Trials–Artificial Intelligence, etc.) compromises the quality of the information that clinicians must rely on to guide adoption and practice. This lack should be remediated.
Second, AIMDs should ideally incorporate methods for alerting users to the degree of uncertainty associated with AIMD predictions. Such methods could provide physicians with a better understanding of the results and their clinical implications (50). When the AIMD is not confident in its prediction, the AIMD must notify the human operator of the increased risk of AI-induced errors (51). Standards organizations and professional societies can assist in providing test datasets to obtain this quantitative information, which is particularly important when considering AIMDs applied to inherently quantifiable technologies such as PET (32,33).

Third, the performance of the AIMD in demographic subgroups should be clearly communicated. For an AI system to be just, clinicians and operators need to be able to explain and understand the system’s limitations, as the clinicians and operators are ultimately accountable and therefore responsible for addressing whether an AI system is unjust or biased (39). Model cards provide one such mechanism for achieving this transparency but must be combined with training schemes that alert clinicians to the limits of an AIMD’s efficacy.

CONCLUSION

The development of AIMDs holds tremendous potential to improve the accuracy and efficiency of medical imaging (50). In the rush to develop these tools, however, researchers should not forget their obligation to consider the autonomy and well-being of data subjects, clinicians, and patients. Nor should they forget that AIMDs have the power to ameliorate health inequalities only if the AIMDs are carefully and judiciously developed with justice in mind. In this paper, we have highlighted some special ethical challenges that AIMD developers should be aware of at each stage in the AIMD pipeline. By careful cultivation of their own sense of justice—in collaboration with ethicists, health disparity researchers, and community members—we are confident that AIMD developers can realize the full potential of this new technology.

DISCLOSURE

Melissa McCradden acknowledges funding from the SickKids Foundation pertaining to her role as the John and Melinda Thompson Director of AI in Medicine at the Hospital for Sick Children. Abihnau Jha acknowledges support from NIH R01EB031051-02S1. Sven Zuehlkodff is a full-time employee of Siemens Medical Solutions USA, Inc. No other potential conflict of interest relevant to this article was reported.

REFERENCES