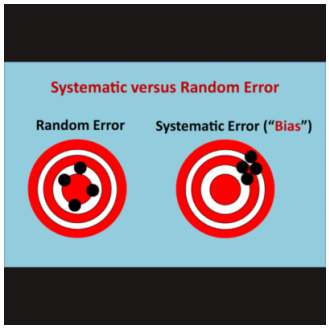
HealthManagement
Promoting Management and Leadership

Strategies for Addressing Bias in Artificial Intelligence for Medical Imaging



Strategies for Addressing Bias in Artificial Intelligence for Medical Imaging

The understanding of bias in artificial intelligence (AI) involves recognising various definitions within the AI context. Bias can refer to unequal treatment based on preexisting attitudes, with distinctions between intentional and unintentional bias. Cognitive bias involves systematic errors in judgement, often stemming from reliance on mental shortcuts. In AI, biases can arise from data limitations, model assumptions, or statistical discrepancies, leading to inaccurate predictions. Systematic error, such as demographic disparities in training data affecting model performance, contrasts with random error, like inconsistencies in image quality impacting measurements. Addressing bias requires consideration at various stages of the AI life cycle: data handling, model development, evaluation, and deployment. An article recently published in RadioGraphics simplifies technical discussions for non-experts, highlighting bias sources in radiology and proposing mitigation strategies to promote fairness in AI applications.

Identifying potential sources of bias in AI for medical imaging

Identifying biases in AI for medical imaging entails looking beyond pixel data to include metadata and text-based information. DICOM metadata and radiology reports can introduce bias if they contain errors or inaccuracies. For example, using patient demographic data or image acquisition details as labels for training models may inadvertently reinforce biases present in the metadata. Moreover, studies have shown that AI models can infer demographic information like race from radiographs, even when such details are not explicitly provided. These latent associations may be difficult to detect, potentially exacerbating existing clinical disparities. Dataset heterogeneity poses another challenge. Training models on datasets from a single source may not generalise well to populations with diverse demographics or varying socioeconomic contexts. External validation on datasets from different sources is crucial to ensure the model's generalizability.

Class imbalance is a common issue, especially in datasets for rare diseases or conditions. Overrepresentation of certain classes, such as positive cases in medical imaging studies, can lead to biassed model performance. Similarly, sampling bias, where certain demographic groups are underrepresented in the training data, can exacerbate disparities.

Data labelling introduces its own set of biases. Annotator bias arises from annotators projecting their own experiences and biases onto the labelling task. This can result in inconsistencies in labelling, even with standard guidelines. Automated labelling processes using natural language processing tools can also introduce bias if not carefully monitored.

Label ambiguity, where multiple conflicting labels exist for the same data, further complicates the issue. This ambiguity can stem from differences in annotators' interpretations or instructions, leading to systematic errors in model training.

Additionally, label bias occurs when the available labels do not fully represent the diversity of the data, leading to incomplete or biassed model training. Care must be taken when using publicly available datasets, as they may contain unknown biases in labelling schemas.

Overall, understanding and addressing these various sources of bias is essential for developing fair and reliable AI models for medical imaging.

Guarding Against Bias in AI Model Development

In model development, preventing data leakage is crucial during data splitting to ensure accurate evaluation and generalisation. Data leakage occurs when information not available at prediction time is included in the training dataset, such as overlapping training and test data. This can

lead to falsely inflated performance during evaluation and poor generalisation to new data. Data duplication and missing data are common causes of leakage, as redundant or global statistics may unintentionally influence model training.

Improper feature engineering can also introduce bias by skewing the representation of features in the training dataset. For instance, improper image cropping may lead to over- or underrepresentation of certain features, affecting model predictions. For example, a mammogram model trained on cropped images of easily identifiable findings may struggle with regions of higher breast density or marginal areas, impacting its performance. Proper feature selection and transformation are essential to enhance model performance and avoid biassed development.

Model Evaluation: Choosing Appropriate Metrics and Conducting Subgroup Analysis

In model evaluation, selecting appropriate performance metrics is crucial to accurately assess model effectiveness. Metrics such as accuracy may be misleading in the context of class imbalance, making the F1 score a better choice for evaluating performance. Precision and recall, components of the F1 score, offer insights into positive predictive value and sensitivity, respectively, which are essential for understanding model performance across different classes or conditions.

Subgroup analysis is also vital for assessing model performance across demographic or geographic categories. Evaluating models based solely on aggregate performance can mask disparities between subgroups, potentially leading to biassed outcomes in specific populations. Conducting subgroup analysis helps identify and address poor performance in certain groups, ensuring model generalizability and equitable effectiveness across diverse populations.

Addressing Data Distribution Shift in Model Deployment for Reliable Performance

In model deployment, data distribution shift poses a significant challenge, as it reflects discrepancies between the training and real-world data. Models trained on one distribution may experience declining performance when deployed in environments with different data distributions. Covariate shift, the most common type of data distribution shift, occurs when changes in input distribution occur due to shifting independent variables, while the output distribution remains stable. This can result from factors such as changes in hardware, imaging protocols, postprocessing software, or patient demographics. Continuous monitoring is essential to detect and address covariate shift, ensuring model performance remains reliable in real-world scenarios.

Mitigating Social Bias in AI Models for Equitable Healthcare Applications

Social bias can permeate throughout the development of AI models, leading to biassed decision-making and potentially unequal impacts on patients. If not addressed during model development, statistical bias can persist and influence future iterations, perpetuating biassed decision-making processes.

AI models may inadvertently make predictions on sensitive attributes such as patient race, age, sex, and ethnicity, even if these attributes were thought to be de-identified. While explainable AI techniques offer some insight into the features informing model predictions, specific features contributing to the prediction of sensitive attributes may remain unidentified. This lack of transparency can amplify clinical bias present in the data used for training, potentially leading to unintended consequences.

For instance, models may infer demographic information and health factors from medical images to predict healthcare costs or treatment outcomes. While these models may have positive applications, they could also be exploited to deny care to high-risk individuals or perpetuate existing disparities in healthcare access and treatment.

Addressing biassed model development requires thorough research into the context of the clinical problem being addressed. This includes examining disparities in access to imaging modalities, standards of patient referral, and follow-up adherence. Understanding and mitigating these biases are essential to ensure equitable and effective AI applications in healthcare.

Navigating Social Bias in Deployed AI Models

Biassed models can perpetuate social bias through various mechanisms once deployed. Privilege bias may arise, where unequal access to AI solutions leads to certain demographics being excluded from benefiting equally. This can result in biassed training datasets for future model iterations, limiting their applicability to underrepresented populations.

Automation bias exacerbates existing social bias by favouring automated recommendations over contrary evidence, leading to errors in interpretation and decision-making. In clinical settings, this bias may manifest as omission errors, where incorrect AI results are overlooked, or commission errors, where incorrect results are accepted despite contrary evidence.

Radiology, with its high-volume and time-constrained environment, is particularly vulnerable to automation bias. Inexperienced practitioners and resource-constrained health systems are at higher risk of overreliance on AI solutions, potentially leading to erroneous clinical decisions based on biased model outputs.

The acceptance of incorrect AI results contributes to a feedback loop, perpetuating errors in future model iterations. Certain patient populations, especially those in resource-constrained settings, are disproportionately affected by automation bias due to reliance on AI solutions in the absence of expert review.

Challenges and Strategies for AI Equality

Inequity refers to unjust and avoidable differences in health outcomes or resource distribution among different social, economic, geographic, or demographic groups, resulting in certain groups being more vulnerable to poor outcomes due to higher health risks. In contrast, inequality refers to unequal differences in health outcomes or resource distribution without reference to fairness.

AI models have the potential to exacerbate health inequities by creating or perpetuating biases that lead to differences in performance among certain populations. For example, underdiagnosis bias in imaging AI models for chest radiographs may disproportionately affect female, young, Black, Hispanic, and Medicaid-insured patients, potentially due to biases in the data used for training.

Concerns about AI systems amplifying health inequities stem from their potential to capture social determinants of health or cognitive biases inherent in real-world data. For instance, algorithms used to screen patients for care management programmes may inadvertently prioritise healthier White patients over sicker Black patients due to biases in predicting healthcare costs rather than illness burden. Similarly, automated scheduling systems may assign overbooked appointment slots to Black patients based on prior no-show rates influenced by social determinants of health.

Addressing these issues requires careful consideration of the biases present in training data and the potential impact of AI decisions on different demographic groups. Failure to do so can perpetuate existing health inequities and worsen disparities in healthcare access and outcomes.

Metrics to Advance Algorithmic Fairness in Machine Learning

Algorithm fairness in machine learning is a growing area of research focused on reducing differences in model outcomes and potential discrimination among protected groups defined by shared sensitive attributes like age, race, and sex. Unfair algorithms favour certain groups over others based on these attributes. Various fairness metrics have been proposed, differing in reliance on predicted probabilities, predicted outcomes, actual outcomes, and emphasis on group versus individual fairness. Common fairness metrics include disparate impact, equalised odds, and demographic parity. However, selecting a single fairness metric may not fully capture algorithm unfairness, as certain metrics may conflict depending on the algorithmic task and outcome rates among groups. Therefore, judgement is needed for the appropriate application of each metric based on the task context to ensure fair model outcomes.

Strategies for Mitigating Bias in AI Development and Deployment: A Multidisciplinary Approach

To mitigate bias in AI development and deployment, it's crucial to engage stakeholders representing diverse perspectives, including radiologists, non-radiologist clinicians, statisticians, data scientists, informaticists, technologists, and department administrators. This interdisciplinary team should thoroughly define the clinical problem, considering historical evidence of health inequity, and assess potential sources of bias.

After assembling the team, thoughtful dataset curation is essential. This involves conducting exploratory data analysis to understand patterns and context related to the clinical problem. The team should evaluate sources of data used to train the algorithm, including large public datasets composed of subdatasets.

Addressing missing data is another critical step. Common approaches include deletion and imputation, but caution should be exercised with deletion to avoid worsening model performance or exacerbating bias due to class imbalance.

A prospective evaluation of dataset composition is necessary to ensure fair representation of the intended patient population and mitigate the risk of unfair models perpetuating health disparities.

Additionally, incorporating frameworks and strategies from non-radiology literature can provide guidance for addressing potential discriminatory actions prompted by biased AI results, helping establish best practices to minimize bias at each stage of the machine learning lifecycle.

Best Practices for Ensuring Fairness and Performance in AI Deployment and Monitoring

To prevent data leakage during model training and testing, it's essential to ensure proper splitting of training, validation, and test datasets at the patient level. Splitting data at lower levels like image, series, or study still poses risks of leakage due to shared features among adjacent data points. When testing the model, involving data scientists and statisticians to determine appropriate performance metrics is crucial. Additionally, evaluating model performance in both aggregate and subgroup analyses can uncover potential discrepancies between protected and non-protected groups.

For model deployment and post-deployment monitoring, anticipating data distribution shifts and implementing proactive monitoring practices are essential. Continuous monitoring allows for the identification of degrading model performance and associated factors, enabling corrective actions such as adjusting for specific input features driving data shift or retraining models. Implementing a formal governance structure to supervise model performance aids in prospective detection of AI bias, incorporating fairness and bias metrics for evaluating models for clinical implementation.

Addressing equitable bias involves strategies such as oversampling underrepresented populations or using generative AI models to create synthetic data. However, caution is needed to avoid perpetuating stereotypes or model collapse. Attempting to generalise models developed on specific populations to other groups can introduce inequitable bias and worsen health disparities, highlighting the importance of monitoring model performance across different demographic groups.

Understanding and addressing bias in imaging AI is essential for its responsible development and deployment. While the article doesn't delve deeply into technical aspects of AI bias, it serves as a consolidated reference for stakeholders involved in creating or acquiring AI solutions. The goal is to identify and address bias before its impact becomes evident, thus promoting fairness and effectiveness in AI applications.

Source & Image Credit: **RadioGraphics**

Published on : Thu, 25 Apr 2024