

# Appendix

**Prepared By:** Mattea Welch, Benjamin Grant, and Christopher Deutschman

**In Consultation With:** Clare McElcheran, Adam Badzynski, Jennifer A.H. Bell, Andrew Hope, Robert C. Grant, Tran Truong, Kelly Lane, Patti Leake, Divya Sharma, Ian Stedman, Laleh Seyyed-Kalantari, Mike Lovas, Jeremy Petch, Benjamin Haibe-Kains, and James A. Anderson

<b>1. General.....</b>	<b>2</b>
1.1. Fairness Definitions.....	2
Bias.....	3
Inequity.....	3
1.2. Common Methodological Biases.....	3
<b>2. Problem Identification and Study Design.....</b>	<b>5</b>
2.1. Define the problem and the solution.....	5
Assembling your cross-functional team.....	5
Patient engagement and participation.....	5
Non-technical solutions.....	5
2.2. Focus on Equity in the Problem Space.....	6
Explore and document baseline inequities.....	6
Literature review (social determinants of health, sources of bias and inequity).....	6
Consult with knowledgeable partners.....	7
Working with First Nations, Indigenous, and Métis Research and Data.....	7
2.3. Outcome Measurements and Data Requirements.....	8
Assess AI Output Integration into Clinic.....	8
Outcome measurement.....	8
Data requirements and availability.....	9
<b>3. Model Training and Development.....</b>	<b>9</b>
3.1. Appropriateness of Retrospective Data.....	9
Identify current clinical benchmarks.....	9
Assessing dataset for disparities and ‘fit for purpose’.....	10
3.2. Defining Objectives and Metrics.....	10
Appropriate statistical measures of performance.....	10
Defining equity objectives.....	11
Measures of Fairness.....	11
Fairness metrics.....	11
Open-Source fairness metric libraries.....	12
3.3. Model Training and Testing.....	13
Acquisition of a third-party model.....	13
Data Representativeness and Quality.....	14
Mitigation methodologies.....	15

Data pre-processing.....	15
ML training methods.....	15
Prediction post-processing.....	16
<b>4. Silent Deployment and Clinical Trial.....</b>	<b>17</b>
4.1. Prospective Deployment Preparation.....	17
Approvals for prospective deployment.....	17
Retrospective to prospective mapping.....	18
Prospective Data Representativeness.....	18
4.2. Prospective Model Evaluation (Silent Mode).....	18
Silent Deployment Considerations.....	18
Many approaches to silent mode.....	18
Threshold selection.....	19
Assessing Silent Mode Performance.....	19
Define auditing methods.....	19
4.3. Prospective Clinical Trial.....	20
Education materials.....	20
Recruit end-users for future clinical integration.....	21
<b>5. Operationalization and Lifecycle Monitoring.....</b>	<b>22</b>
5.1 Preparation and Documentation.....	22
Comprehensive documentation.....	22
Internal Documentation:.....	22
1. Model Development and Technical Specifications.....	22
2. Deployment Documentation.....	23
3. Regulatory and Compliance Documentation.....	23
4. Monitoring and Post-Deployment Validation.....	23
5.2. Communication and Education.....	25
Modifying trial educational materials.....	25
Communication plan.....	25
5.3. AI-solution Rollout and Monitoring.....	26
Regular monitoring and reporting.....	26
5.4. Updating or Decommissioning of AI-solution.....	26

# 1. General

## 1.1. Fairness Definitions

While both ‘bias’ and ‘inequity’ have myriad meanings, for the purposes of this document, the terms will be defined as follows:

## Bias

An umbrella term for both social biases and methodological biases. Social bias, which can be conscious or unconscious, can be defined as “discrimination for, or against, a person or group, or a set of ideas or beliefs, in a way that is prejudicial” <sup>1</sup>. Methodological bias can be defined as “systematic errors stemming from choices made during the research process” <sup>2</sup>, such as selection bias, observer bias, response bias, and publication bias <sup>3</sup>[Popovic 2024]. In the machine learning sphere, there is often an intersection between social and methodological biases (for example, data based on nursing assessments may have a methodological observer bias, where part of the inter-observer variability is due to each nurse’s individual social biases). Within the context of AI, a biased solution is one that has different performance for different subgroups of patients or inequitable impact on different subgroups of patients.

## Inequity

A term defined broadly as “unjust differences between populations in the access, use, quality, and outcomes of care.” <sup>4</sup>. The key difference between ‘inequities’ and ‘disparities’ is that ‘inequity’ is a value-laden term – the differences described are unjust or unfair – whereas ‘disparity’ merely describes that a difference is present. Serious scrutiny is required to determine in which cases a ‘disparity’ should or should not be considered an ‘inequity’. Inequities are *systematic* differences in the opportunities groups have to achieve optimal health <sup>5</sup>. These systematic differences are influenced by social biases and by structural differential access opportunities.

## 1.2. Common Methodological Biases

Below is a list of commonly found methodological biases that can be present at different times of an AI-solutions development and deployment. This list is not meant to be a comprehensive overview of all potential biases, but to be an introduction to some of the most commonly found ones.

- 1) **Selection bias:** Occurs when the training data is not representative of the broader patient population, resulting in skewed results. This can lead to models that perform well on certain subgroups but fail on others, especially if certain patient demographics are underrepresented (e.g., minority groups).<sup>6,7</sup> This can be the result of biased sampling, measurement errors, or incomplete data collection, leading to non-representative datasets. In healthcare, this can occur if electronic health records (EHRs) or datasets are incomplete or contain errors, leading to skewed models. For example, if data is collected only from a specific hospital type or region, the model may not generalize well to other settings.

Missing data also causes biases when missing in a systematic way, leading to skewed results. For instance, if patients who are healthier are more likely to have missing records (i.e. less tests performed), models may underpredict positive outcomes.

- 2) **Confirmation bias:** The tendency to search for, interpret, or prioritize information that confirms one's preexisting beliefs or hypotheses, while disregarding contradictory evidence. e.g. a belief that a population of patients is not impacted by inequitable care, and therefore a thorough literature search is not conducted.<sup>8,9</sup>
- 3) **Incorporation bias:** A subtype of verification bias, where the predictive model includes features or data points that are part of the outcome it is trying to predict. For example, if an ML model predicts a diagnosis and uses test results that are part of the diagnostic process itself, it may artificially inflate model performance.<sup>10</sup>
- 4) **Survivorship bias:** Arises when only patients who have survived or remained in care and included in the study and those who have died or dropped out are ignored. In ML model predictions, this can lead to over-optimistic performance in predicting patient outcomes, as the model ignores negative outcomes. For qualitative research related to patient experiences, this will also, likely, result in over-optimistic evaluation of an AI-solution.<sup>11,12</sup>
- 5) **Model overfitting:** When a ML model is too complex, it may fit the noise or anomalies in the training data rather than the underlying patterns. This leads to excellent performance on the training set but poor generalization to new, unseen patient data.<sup>13,14</sup>
- 6) **Spectrum effects:** Occur when a model's performance varies across different patient subgroups, such as those with different disease severities. This bias can result in a model performing well for a mild disease but poorly for severe cases, leading to misleading evaluations of model accuracy.<sup>15</sup>
- 7) **Automation bias:** This bias occurs when clinicians overly trust ML model outputs without critically evaluating them. Over-reliance on automation can lead to errors, especially if the model is flawed or the input data is incorrect.<sup>16</sup>
- 8) **Unconscious bias:** Also known as implicit bias, unconscious bias can be introduced if the data or model development reflects the prejudices of the research team. For example, models might inadvertently favor certain patient groups over others if the underlying data contains implicit biases related to race, gender, or socioeconomic status.<sup>17</sup>

## 2. Problem Identification and Study Design

## 2.1. Define the problem and the solution

### Assembling your cross-functional team

The cross-functional team composition and involvement of specific individuals should depend on the scope, topic and the current stage of the project. The team is not a formal review committee, but one that ensures each stage of development is considered by an appropriate expert. Over the lifetime of the project, the team should, at some point, include ML researchers, clinicians, data analysts, statisticians, equity and ethics experts, patient representatives, clinical champions and end-users, as well as ethics, privacy and strategy representatives.

### Patient engagement and participation

Patient voices, represented by patient partners, family and caregivers, and/or professional patient advocates, should be present within and beyond the cross-functional team. Patient partnerships are critical to make sure that patients are guiding the research in identifying unmet needs and improving outcomes for those needs. Patient voices can be present in the whole spectrum of AI research, from problem identification and study design through to clinical implementation and lifecycle monitoring.

Many institutions and disease specific advocacy groups will have their own supports for patient partners, so it is important to look for local initiatives. In Canada, the Canadian Institutes of Health Research has a framework for patient-oriented research and engagement (<https://cihr-irsc.gc.ca/e/48413.html>) - similar initiatives may exist in other countries.

### Non-technical solutions

The allure of AI in healthcare often creates a "shiny object syndrome"<sup>18</sup> where research teams and organizations rush to implement cutting-edge technology simply because it's novel and exciting, rather than because it's the most effective solution. An AI-first mindset can lead to overlooking simpler, proven interventions that might better serve patients and healthcare workers. When enthusiasm for AI's potential overshadows careful consideration of alternatives, organizations risk investing substantial resources into complex technical solutions for problems that might be better solved through basic process improvements, better staffing, or enhanced communication systems. It's worth remembering that sometimes what appears to be a technology problem is actually a human systems problem in disguise, and no amount of sophisticated AI can fix underlying issues with workflow, staffing, or resource allocation.

When addressing healthcare challenges, it's crucial to first explore fundamental non-technical and non-AI-solutions that may be more effective and sustainable. These

approaches may require fewer resources, face fewer implementation challenges, and can be more easily adapted to local contexts compared to AI-solutions.

## 2.2. Focus on Equity in the Problem Space

### Explore and document baseline inequities

A key component of responsible AI development and deployment is mitigating bias, fairness and inequity. In order to do this, there need to be clearly defined equity objectives and fairness metrics to measure success. These definitions will be specific to the population, problem space, and solution that is being addressed by a particular AI-solution. Exploring and understanding baseline inequities is paramount to developing a fair and equitable model, and can be done in a few different ways (please see sections below). Taken together, these learnings lay the foundation for quantitative analysis of biases and inequities within your target population once you have access to local retrospective data to explore in [Section 3.2](#).

#### *Literature review (social determinants of health, sources of bias and inequity)*

The equitable and compassionate components of this framework rely on an extensive review of documented biases in healthcare literature. However, it should be acknowledged that biases identified in the literature are not guaranteed to be exhaustive, nor always fully representative in the collected data or healthcare system, particularly when the evolution of protected attributes over time is considered (e.g. newly adopted gender definitions that are more granular and representative, or new discoveries in a disease definition that increases specificity). Key patient factors associated with healthcare inequity include race<sup>19–23</sup>, age<sup>24–27</sup>, sex<sup>28–30</sup>, geographical location of residence<sup>31,32</sup>, patient support systems<sup>33,34</sup>, primary language, income level, and other social determinants of health. Each of these elements independently may impact healthcare access, treatment, and/or outcomes, and often the effects compound when elements are present together, creating unique barriers for those with intersectional identities. Common intersections may include age/sex, and ethnicity/sex. Accounting for intersectional identities and intersections of other health-related groupings remains challenging<sup>35</sup>; there are far more combinations of intersecting demographic factors than can be practically addressed. Broadening bias assessment to include intersectional identities is essential, but often constrained by data availability.

#### *Consult with knowledgeable partners*

There will be gaps in the healthcare literature regarding what biases exist in a given problem space. By consulting with individuals who understand the patient experience in the specific problem space, as recommended in our framework, the investigator may begin to fill in these gaps. There are numerous stakeholders that can inform the team's understanding of these biases, including patients, patient partners, families/caregivers, patient navigators, and healthcare providers. The goal of this consultation process is to identify any biases or inequities, real or perceived, that must be considered and either successfully mitigated or identified as a limitation to the work.

## Working with First Nations, Indigenous, and Métis Research and Data

For AI-solutions that are focused on First Nations, Indigenous, and/or Métis populations, or are known to utilize First Nations, Indigenous, or Métis data, special care should be taken to respect the data sovereignty of these groups. If the AI-solution will use data from specific individual nations, the project team must seek approval from appropriate leaders from those nations prior to commencing the project.

To bolster the researcher's own capacity to respect/assert the principles of data sovereignty, it is recommended that all members of the research team complete The Fundamentals of OCAP® course that was developed by the First Nations Information Governance Centre. OCAP stands for Ownership, Control, Access and Possession, and the course provides valuable information on how to appropriately engage with these communities and their data.

Through this course, participants will learn about the history and motivation behind OCAP, as well as receive practical steps for participating in, or seeking to participate in relevant research studies. As examples, individuals should ask themselves the following questions, among others that are outlined in the course:

1. Was this data collected with the approval and knowledge of the First Nations?
2. Does the project align with the priorities of the First Nation(s) from whom the data is coming?
3. Could undertaking this project/using this data cause harm to the First Nation(s) or its members?
4. Are First Nation(s) members being included in the project from conception to analysis to implementation?

## 2.3. Outcome Measurements and Data Requirements

## Assess AI Output Integration into Clinic

Prior to embarking on the lengthy and costly endeavor of AI-solution development, testing and integration ensure that the solution will have a measurable impact on clinical processes. Two important questions to ask are: (1) What will be done differently in the clinic, either operationally or for patients, with the AI-solution outputs; and (2) do pathways exist to act on AI-solution outputs.

The clinical impact of an AI-solution's output may be affected by a variety of scenarios including the state of the care facilities outside of healthcare (eg. an output recommending discharge of patients to long term care centers which do not have vacancies), available treatments (eg. an output identifying patients who will become septic which has not had advances in treatment in recent decades), and healthcare resources (eg. an output that predicts emergency room visits which are currently at capacity). The ability to address scenarios to improve the impact of AI-solutions ranges in feasibility and control of the research team and should be thoughtfully considered.

## Outcome measurement

Defining the ideal outcome and determining what should be measured to assess that outcome are two separate steps. First, the ideal outcome should be defined based on the ideal clinical state after AI-solution integration. For example, for an AI-solution meant to predict and intervene in post-surgical pain crisis, the ideal clinical state post-deployment might be 'more patients with appropriately managed pain'. After the ideal clinical state has been defined, the most appropriate outcome measurement(s) can be identified. In this example, measurement may be simple, such as self-reported patient pain or pre- and post-deployment number of analgesic orders per patient .

The challenge lies in making sure that the measurement(s) chosen is/are fair and equally assessable across all affected subpopulations. Patients of colour are often prescribed pain medication at a lower rate than white patients, despite being in equal or higher levels of pain. When they are prescribed analgesics, patients of colour are also more likely to be given a lower dose of the medication or given a different regiment entirely. If the AI-solution is making predictions of the likelihood of pain crisis based in part on historic analgesic orders, the prediction will not be appropriately sensitive for patients of colour.



In this example, measuring outcomes such as ‘rate of analgesics ordered across target subpopulations’ or ‘average patient self-reported pain pre- and post-deployment per target subpopulation’ may be more appropriate to capture how the AI-solution is affecting different groups. Without taking care to do measurement by subpopulation, there may still be an overall increase in orders or overall decrease in reported pain after surgery, but those trends may be primarily influenced by the experiences of white patients, and the unequal increase in orders/dosing would lead to an overall inequitable clinical state.

## Data requirements and availability

Before moving on to model training and development, it is important to consider the data that is required to allow your model to perform effectively and equitably. This includes determining specific outcome measures, the features that need to be included in the model, and whether they are collected in a format, and for a purpose, that is adequate for your specific problem and population. Most clinicians, researchers, and AI specialists are not particularly familiar with the sources, types, structures, and availability of data, so it is worth doing this investigation up front to avoid significant problems at later stages.

It is also important when identifying data elements and data sources to understand the context in which the data was gathered. How and why the data was collected and the intended purpose/user of the data is important information when assessing potential biases and limitations with the data.

# 3. Model Training and Development

## 3.1. Appropriateness of Retrospective Data

### Identify current clinical benchmarks

Clinical benchmarks are essential for evaluating new machine learning models in healthcare because they ensure relevance and impact. By setting a comparison standard grounded in real-world clinical practice, these benchmarks help establish whether a model provides tangible benefits over existing methods. This relevance ensures models address actual clinical needs rather than producing technically interesting but clinically irrelevant results.

Furthermore, benchmarks foster trust and safety. They align models with regulatory standards, reducing risks and proving the model's utility before deployment. Comparative benchmarking also helps stakeholders gauge the model's effectiveness against current methods, ensuring that any advancements are not just statistically

significant but clinically meaningful, thereby justifying the model's use in healthcare settings.

### Assessing dataset for disparities and 'fit for purpose'

Biases that should be considered during this phase of the project include sampling, convergence and participation bias. Additionally, using known societal inequities and biases that were identified during the literature search of the problem space, assess your retrospective dataset to ensure it is properly representative of these known issues. This can be achieved most simply by looking at dataset distributions, and minority and majority class representations. Furthermore, for protected or minority groups, a statistical power analysis should be performed to ensure there are enough individuals in the dataset to garner meaningful information.

## 3.2. Defining Objectives and Metrics

### Appropriate statistical measures of performance

Defining appropriate statistical measures of performance for an AI-solution is crucial to ensuring its effectiveness and reliability. The choice of metrics should align with the specific problem domain, objectives, and nature of the dataset. For instance, in binary classification tasks, measures like AUC-ROC (Area Under the Receiver Operating Characteristic Curve)<sup>36,37</sup> and PR-AUC (Precision-Recall Area Under the Curve)<sup>38,39</sup> are often used. However, selecting between these metrics depends on the data characteristics and known imbalances. AUC-ROC provides an overall view of the model's ability to discriminate between classes, but it may not be as informative in cases of class imbalance. PR-AUC, on the other hand, focuses on the precision and recall trade-off, making it a better choice when the positive class is rare or when false positives and false negatives have differing consequences. Failure to select the right metric can result in a misleading evaluation of the model's performance and potentially biased performances for the majority class.

Some metrics also have known dependencies that must be considered to ensure accurate interpretation of performance. For example, the Dice Similarity Coefficient (DICE), commonly used in medical image segmentation tasks, is influenced by the size or volume of the segmentation<sup>40,41</sup>. Larger segmentations may artificially inflate the DICE score, while smaller ones might unfairly penalize it. Understanding these dependencies is essential to avoid obscuring the true performance of the model. Practitioners should complement such metrics with additional measures or normalization techniques to account for these biases and provide a more holistic evaluation.

Another key aspect is the validation strategy employed to assess the AI model. Techniques like k-fold cross-validation help to mitigate overfitting and provide a robust estimate of the model's generalizability<sup>42,43</sup>. Using too few folds might lead to high variance in performance estimates, while excessively large numbers of folds can increase computational costs without significant gains in reliability. An inadequate validation process can lead to over-optimistic results on retrospective data that fail to generalize to new prospective data, undermining the solution's practical applicability.

Overfitting is another common pitfall that arises when statistical measures are not appropriately applied or understood<sup>13</sup>. Overfitting occurs when a model learns the training data too well, capturing noise instead of generalizable patterns. This often happens when performance metrics are optimized exclusively on the training data or when complex models are used without adequate regularization. For instance, reporting high accuracy on training data while neglecting poor performance on unseen data can give a false sense of success. Similarly, reliance on a single metric, such as accuracy, in imbalanced datasets can obscure significant deficiencies, such as a model's inability to correctly identify minority class instances.

Misusing statistical measures—whether by selecting inappropriate metrics, inadequately validating models, or overfitting to training data—can lead to incorrect conclusions and poor real-world performance. By thoughtfully addressing these considerations, practitioners can create AI-solutions that are both accurate and trustworthy.

## Defining equity objectives

Equity objectives for a project are informed by current inequities related to the problem space, and will fall on a spectrum from maintaining current inequity levels to reducing them; an AI-solution should never make inequity levels worse. To obtain the identified equity objective, appropriate and complementary fairness metrics are required.

## Measures of Fairness

Selecting appropriate measures of fairness is imperative to the assessment of developed AI-solutions. A few example measures are highlighted below, but investigators are encouraged to do a review of the literature to see if any newer and more relevant measures have been developed.<sup>44–47</sup>

### *Fairness metrics*

Designed to analytically assess equality and equity issues in order to give insight into the nature of the models performance.

- 1) Group Unawareness: Group unawareness means that a model does not use a sensitive variable during prediction<sup>48</sup>. For example, in a simple linear regression, the coefficient of the sensitive variable would be set to 0. Group Unawareness can be assessed using metrics such as SHAP, or by looking at the statistical significance of a univariable Ordinary Least Squares Regression that has been fit to predict the outcome of interest can be predicted using a single feature. Caution when using Group Unawareness is needed in scenarios where the sensitive variable may be highly correlated with a proxy variable.
- 2) Statistical Parity/Demographic Parity: measurement of whether a model predicts positive outcome at equal rates for each segment of a subgroup.<sup>49</sup>
- 3) Equal Opportunity: measures whether individuals from each segment of a protected class, who are eligible/qualified, have the same probability of receiving a positive outcome (e.g. being offered participation in a clinical trial). However, assessment of eligibility/qualification can be subjective, and Equal Opportunity does not address biases in the assessment process.<sup>46,50</sup>
- 4) Equal Odds: this metric is used to quantify whether equal true positive rates and false positive rates exist between different groups. It is more restrictive than equal opportunity making it more appropriate when there are existing biases in the data. However, this is a very restrictive metric and may reduce the model's performance.<sup>51,52</sup>
- 5) Positive Predicted Value (PPV) - Parity: given a positive prediction, the precision is equal across different groups. For example, if a model positively predicts that a patient should receive "treatment X", the probability of this treatment being successful is equal in all segments of a protected class.<sup>53</sup>
- 6) False Positive Rate (FPR) - Parity: this metric is the opposite of PPV-Parity and wants to ensure that each segment of a protected class has the same false positive rate. For example, if the model positively predicts that a patient should receive "treatment X", the probability of this treatment not being successful is equal in all segments of the protected class.<sup>54</sup>

### *Open-Source fairness metric libraries*

- 1) [FairLearn](#): An open-source python package designed for metric calculation and reduction of bias in algorithms.
- 2) [Fairness Indicators](#): A package from Google designed to work with TensorFlow. Can be used for evaluation and visualization of group disparities.
- 3) [AIF360](#): An open-source package used to detect and mitigate biases. It is known for its extensive documentation and tutorials.
- 4) [Themis-ML](#): An open-source package for easy integration with scikit-learn. It is used mainly for binary classes and evaluation.

## 3.3. Model Training and Testing

### Acquisition of a third-party model

Assessment of an AI-model, or AI-solution, acquired from a third party for deployment in a healthcare institution is imperative for the safety of patients. Comprehensive questioning is required across multiple domains. Some general questions that should be asked include:

1. General Information
  - a. What is the intended purpose and scope of the model?
  - b. Has the model been used in clinical settings similar to ours?
  - c. What are the specific clinical problems it aims to solve?
  - d. Were patients and end-users consulted during the conception and development of this model?
2. Development and Validation
  - a. What data was used to train and validate the model? (e.g., size, sources, geographic diversity, demographic representation) Are we able to do an internal assessment of the data?
  - b. What validation processes were followed? (e.g., external validation, cross-validation)
  - c. What are the key performance metrics, and how do they vary across subpopulations of interest?
3. Bias and Fairness Assessment
  - a. Was the training data representative of the populations the model will serve? This may be challenging if the training data cannot be accessed and compared to local data.
  - b. What methods, if any, were used to detect and mitigate biases in the development phase?
  - c. Are there known performance disparities across demographic groups (e.g., age, sex, race, ethnicity)?
  - d. What fairness frameworks or metrics were used to evaluate the model?
  - e. Does the model include safeguards to minimize inequities in its recommendations?
  - f. Are there transparency mechanisms to report when biases or unfair outcomes are detected post-deployment?

4. Safety and Risk Management
  - a. What safeguards are in place to ensure patient safety in case of errors?
  - b. Has the model undergone stress testing for edge cases or rare clinical scenarios?
  - c. What adverse outcomes or unintended consequences were identified during testing?
  - d. Is there a protocol for monitoring and reporting errors or adverse events post-deployment?
5. Compliance and Regulatory Adherence
  - a. Does the model comply with relevant regulations?
  - b. Are there documented audit trails for data usage and model outputs?
6. Operational Integration
  - a. What are the technical requirements for integration with our existing systems (EHR, PACS, etc.)?
  - b. What resources are needed for deployment and maintenance?
  - c. What user training and support are provided?
7. Post-Deployment Monitoring
  - a. What tools or processes are available for continuous monitoring of model performance, clinical impact, and operational and regulatory adherence?
  - b. How frequently should the model be retrained or updated?
  - c. How is feedback from clinicians and patients incorporated into updates?
  - d. Are there mechanisms to adjust or recalibrate the model based on observed disparities?
8. Ownership and Intellectual Property
  - a. Who owns the model and any updates made during its use?
  - b. What are the terms of data sharing, if applicable?

## Data Representativeness and Quality

When the retrospective dataset is split into training and testing cohorts (through e.g. randomized stratified splitting, bootstrapping, time wise split), investigators should take care to assure each cohort retains the same representation of known biases and inequities (i.e. the distributions should be the same when compared to the full retrospective dataset).

Data quality should also be assessed at this step. Data missingness and consistency in variable naming are two such tests that should be completed prior to commencing ML training. Additionally, if using longitudinal data, an understanding of whether there were any changes in clinical practice over time is required to avoid temporal bias (e.g. variable naming standards, standard treatment regimens, pandemics affecting patient presentation).

## Mitigation methodologies

Similarly to the fairness and equity measures section presented in subsection 3.3 of our appendix, this section is used only to highlight a few different methods to improve model fairness. Additional methods can be found in some of the open-source fairness metric libraries mentioned above. A literature review is recommended if none of the methods below are appropriate, or to determine if there are newer and more relevant methods that have been developed.

### *Data pre-processing*

In some scenarios, it is appropriate to change or adjust the dataset to be fairer before training an ML model.

- 1) Relabeling and perturbation: These methods involve changing the end-point label or features in a dataset. Relabeling attempts to balance the dataset by changing the end-point label, while perturbation involves varying the features/variables to create a more balanced representation of the data. Two examples of these methods are disparate impact remover<sup>55</sup> and “massaging”<sup>56</sup>. However, relabeling or perturbing data can introduce inaccuracies and distort the data's original distribution. This may lead to models learning incorrect or overly simplified relationships, so these techniques must be thoroughly validated against the original data distributions.
- 2) Sampling: A dataset can be sampled up or down by adding or removing samples, respectively, to change the sample distributions to one that is more balanced. Up sampling the minority class can be done using duplication of existing samples or synthetic data, but caution is warranted since the model may start to overfit to duplicated or synthetic examples. Down sampling can involve removing majority group samples, but similarly, caution is warranted since data complexity can be reduced. Methods such as Synthetic Minority Over-sampling Technique (SMOTE)<sup>57</sup> attempt to balance these methods.

### *ML training methods*

These methods are used to modify or alter ML training algorithms to improve model fairness.



- 1) Regularization and constraints: these methods alter the loss function of an algorithm. During regularization, an extra term is used to penalize discrimination, while constraints are used to limit the allowed bias level according to a certain loss function. Prejudice Remover<sup>58</sup>, Exponentiated Gradient Reduction<sup>59</sup>, Grid Search Reduction<sup>59</sup> and Meta Fair Classifier<sup>60</sup> are all examples of these techniques. A potential drawback of these methods is the difficulty in correctly defining the penalization terms or constraints. Overly strict constraints might lead to underfitting and reduced model performance, while poorly chosen terms can fail to adequately address bias. Additionally, these techniques may require significant computational resources and careful tuning, which can be challenging in practice.
- 2) Adversarial learning involves training two models that compete to improve their performance<sup>61</sup>. Specifically, one model attempts to predict the true label of a dataset, while the other model attempts to exploit a known fairness issue using equality metrics. A drawback of adversarial learning is its complexity and the risk of instability during training, as the competing objectives of the models can lead to convergence issues. Adversarial models may inadvertently reduce overall predictive accuracy if fairness constraints conflict significantly with optimizing performance. Furthermore, adversarial models are not appropriate in scenarios where there are known differences between subgroups. As an example, when developing auto-segmentation models with a known performance difference between males and females adversarial learning should not be used since morphological differences are known to exist between sexes <sup>62</sup>.

### *Prediction post-processing*

Post-processing methods act on the model predictions and are used when access to training data or the model is limited. These methods would be more appropriately used in scenarios where the model is commissioned from an external group.

- 1) Classifier correction: A trained ML model is adapted to remove discrimination based on equalized odds and equality of opportunity constraints. Calibrated Equalized Odds<sup>63</sup> is an example of one of these methods. However, classifier correction depends on accurately identifying sources of bias and setting appropriate constraints. Poorly defined constraints can lead to either insufficient fairness improvements or a decrease in the model's predictive accuracy. Additionally, applying such corrections post-training may not address deeper issues of bias inherent in the data, limiting their effectiveness in mitigating unfairness.



- 2) Output correction: model outputs are modified to obtain fairer distribution of the data. Reject Option based Classification<sup>64</sup> is an example of this type of method that assigns more favourable outcomes to protected groups based upon low confidence regions of the classifier. While output correction can improve fairness metrics, it may distort predicted probabilities and reduce trust in model predictions. This approach addresses bias at the output level without resolving biases present in the training data or the model itself, potentially leading to superficial fairness improvements that fail to generalize to new datasets. Furthermore, aggressive corrections can harm overall model performance and usability for certain tasks.

## 4. Silent Deployment and Clinical Trial

### 4.1. Prospective Deployment Preparation

#### Approvals for prospective deployment

Similar to [Section 3.1.](#), access to and use of prospective clinical data requires institutional approval of an REB or QI protocol submission. Which stream to choose depends on the specifics of the project, including the type of data that is required, the sensitivity of that data, the intended audience of the model outputs, and the nature of any potential interventions based on the model outputs.

Some important questions to consider:

1. Which clinical system(s) will provide prospective data?
2. Do you require data on demand (live data feed) or on a schedule?
3. Is there sufficient infrastructure to meet your data needs?
4. Are there limitations on how the prospective data can be used (e.g., Silent Mode data cannot be used to develop new models)?
5. What are the potential interventions that could be prompted by your deployed model and could they have a direct impact on clinical decision-making?

Depending on the nature of the data that is required and how the model will be used,

## Retrospective to prospective mapping

Healthcare data and EHR systems are constantly changing. In many cases, AI models are developed and tested using historical data, and/or from historical records systems. In these cases, every effort must be made to evaluate how this data corresponds to the data expected from defined sources of prospective healthcare data, and clearly map the differences. It is also important to consider spectrum effects and any unintended disparities in data.

While navigating a change in EMR system between model development and silent deployment is a dramatic (albeit common) scenario, similar approaches are a necessary part of any prospective deployment. Even within any given EMR, important definitions such as procedure codes, names of fields, and other critical data may change over time without warning. After any initial data mapping is complete, prospective data needs to be compared to historical data to ensure consistency and accuracy. Questions you can ask yourself are:

1. Are all of the features included in your model being captured?
2. Are there any unexplained shifts in your data? (missingness, volumes, values, etc.)

This is an iterative process until the prospective data quality meets expectations, but will continue as part of lifecycle monitoring.

## Prospective Data Representativeness

In ideal scenarios the prospective data used during Silent Mode and Prospective Deployment will have demographic distributions and data collection methods that are similar. However, there are certain scenarios where retrospective and prospective datasets will differ, but the model still meets defined objectives and metrics. In this scenario, the model is robust to certain acceptable variable/data variations. These acceptable variations should be assessed and documented so that the model does not undergo unnecessary retraining or decommissioning.

## 4.2. Prospective Model Evaluation (Silent Mode)

### Silent Deployment Considerations

#### *Many approaches to silent mode*

There are many different potential approaches and phases to silent mode, and what is best will completely depend on the specific solution being evaluated. As a general guide, silent mode involves: (1) developing required workflow integration components that consider human factors and user experience; (2) validation of the models performance; (3) reliability and integration within the clinical workflow under real-world

settings; and (4) identification of potential biases, usability issues and any unintended consequences.

### *Threshold selection*

Some AI-solutions will require the selection of a statistical threshold, above and below which different actions are taken. This threshold determines the binarization point of a ML prediction. This threshold should be selected in collaboration with clinical-end-users and the project's cross-functional team. It should also be chosen to balance false positive rates and false negative rates, to minimize unnecessary operational burden and interventions, and delayed diagnosis or intervention, respectively<sup>65</sup>. These thresholds should be selected with consideration of subgroups.

## Assessing Silent Mode Performance

Silent mode performance should broadly assess how the model performs in a real-world setting, as well as how the solution is integrated into current clinical work-flows. The cross-functional team should be consulted if the model fails to meet any of the defined statistical measures of performance, equity objectives of fairness metrics to determine if the model should be adapted, mitigation strategies should be implemented, or if the AI-Solution is not meeting the intent of the project and should be discontinued.

## Define auditing methods

In collaboration with the cross-functional team, an auditing plan should be generated prior to solution rollout. The frequency of audits will be determined by the team. Audits of the solution are comprehensive and are not meant to replace regular assessment of model performance. An audit of the solution should consider the following:

- a. Ethics, security and privacy check of solution and data being used.
- b. Changes in standard practice, structural changes, or other sources of data drift, that may affect the data being used.
- c. Model performance across patient subgroups.
- d. Whether or not equity objectives and fairness metrics are still appropriate and being met.
- e. Feedback from end-users regarding experience, questions, adoption and any additional recommendations about the solution.
- f. Failures that have occurred since the last audit (including if they were properly communicated and addressed)

## 4.3. Prospective Clinical Trial

### Education materials

Developing education materials to provide during the Clinical Trial of an AI-solution in healthcare is critical to ensure all stakeholders understand the technology, its goals, and their roles. This section highlights a sampling of materials a cross-functional team may need to develop for their clinical trial, depending on the stakeholders involved.

1. end-users:
  - a. User Manuals and Quick Start Guides:
    - i. Include examples of AI outputs and how to interpret them.
    - ii. Provide mechanisms for users to report issues or provide feedback.
  - b. Decision Support Documentation:
    - i. Highlight evidence supporting the AI model's recommendations.
    - ii. Address limitations, biases, and areas of uncertainty.
    - iii. Describe situations where the model should not be used or where caution is needed.
  - c. Compliance and Safety Guidelines:
    - i. Clarify the trial's regulatory compliance and safeguards to prevent patient harm.
2. For Patients (and Caregivers, if applicable):
  - a. Simplified Information Sheets or Videos:
    - i. Describe the AI-solution, its role in their care, and expected benefits.
  - b. Trust-Building FAQs:
    - i. Address concerns about AI (e.g., "Will AI replace my doctor?").
    - ii. Reassure participants about clinician oversight and privacy safeguards as relevant.
  - c. Patient Feedback Channels:

- i. Provide materials explaining how participants can offer feedback or raise concerns.

## Recruit end-users for future clinical integration

When assessing the AI-Solution during a Clinical Trial, obtaining feedback from end-users and patients involved in the trial is essential to evaluate whether clinical integration is effective. Various methods can be used to obtain this feedback including surveys, interviews, focus groups, or direct observation. Here are some examples of feedback to consider:

### Feedback from end-users:

1. Is the AI-solution intuitive and user-friendly?
2. Are the AI insights/actionable recommendations accurate and clinically meaningful?
3. Do the AI insights/actionable recommendations align with evidence-based practices or clinical guidelines?
4. Does the AI tool save time or reduce manual workload? Does it increase confidence in developing a care plan? Does it improve equitable care?
5. Is it reducing cognitive load or decision fatigue?
6. Is the rationale behind AI outputs clear and explainable?
7. *If the end-user is a clinician:* Do you feel confident relying on the tool?
8. *If the end-user is a clinician:* Are you noticing measurable improvements in patient care quality or outcomes?
9. Was adequate training provided on how to use this tool?
10. Is ongoing technical support accessible and helpful?

### Feedback from Patients:

1. Are you aware that AI is part of your care process?
2. Do you understand how the AI contributes to your treatment?
3. Do you feel the tool is enhancing your care (e.g., faster results, personalized treatments, fewer errors or delays)?
4. Are there any concerns about depersonalization of care?
5. Do you feel your data is secure and used responsibly?
6. Are you comfortable with AI being used in decision-making?
7. Do you trust the collaboration between clinicians and AI?

## 5. Operationalization and Lifecycle Monitoring

### 5.1 Preparation and Documentation

#### Comprehensive documentation

Comprehensive documentation bridges the gap between development, deployment, and real-world use. Documentation should be generated both for internal usage (to ensure safe, ethical, and effective implementation within the local care system) and external publishing (to foster transparency, reproducibility, and collaboration).

Some best practices to keep in mind for both internal and external documentation include: (1) generating plain language summaries for accessibility to non-technical stakeholders; (2) ensuring all documentation aligns with FAIR principles (Findable, Accessible, Interoperable, and Reusable)<sup>66</sup> when publishing; and (3) including a revision history of any updates and improvements to the AI-solution.

Below are some recommendations for documentation for internal usage and external publishing:

#### Internal Documentation:

1. Model Development and Technical Specifications
  - a. Model Overview:
    - i. Objectives and intended use cases.
    - ii. Description of the problem the AI addresses.
  - b. Model Architecture:
    - i. Detailed explanation of algorithms, architecture, and training pipeline.
    - ii. Hyperparameters and optimization techniques.
  - c. Data Sources and Preprocessing:
    - i. Description of datasets used for training, validation, and testing.
    - ii. Details on data cleaning, augmentation, and handling missing values.
    - iii. Consider use of Data Labels<sup>67</sup>
  - d. Fairness, Equity and Performance Metrics:
    - i. Intended equity objective(s) and utilized fairness metric(s)
    - ii. Performance evaluation metrics (e.g., accuracy, precision, recall, AUC-ROC).
    - iii. Comparison with baseline methods.
  - e. Ethics and Bias:
    - i. Description of steps taken to minimize bias and promote fairness.

- ii. Description of known biases and inequities in AI-Solution.
- f. Versioning:
  - i. Model version history and updates.
- 2. Deployment Documentation
  - a. System Integration:
    - i. How the AI integrates with existing healthcare infrastructure (including e.g. what data sources are accessed, who the primary end-users are, etc).
  - b. Operational Guidelines:
    - i. Deployment process, runtime environment, and maintenance requirements.
  - c. Interfaces:
    - i. User interface guides for clinicians or administrators.
- 3. Regulatory and Compliance Documentation
  - a. Data Privacy and Security:
    - i. Compliance with regulations like GDPR, HIPAA, or equivalent.
    - ii. Explanation of data storage, access controls, and encryption methods.
  - b. Risk Management:
    - i. Identified risks and mitigation strategies.
- 4. Monitoring and Post-Deployment Validation
  - a. Performance Monitoring Plans:
    - i. Procedures for tracking model performance in real-world settings.
  - b. Update and Retraining Guidelines:
    - i. When and how the model should be updated or retrained.
  - c. Audit Logs:
    - i. Documentation of decision-making processes for accountability.

#### External Documentation (For Publishing to Journals or Open Access Repositories):

The TRIPOD+AI<sup>68</sup> statement is an excellent resource that should be followed for external publishing of an AI-Solution. As a starting point, we recommend the following sections for inclusion in any external publication:

- 1. Research and Model Development
  - a. Introduction and Background:
    - i. Problem statement, clinical relevance, and literature review.
  - b. Methods:
    - i. Detailed description of the AI development process, including:
      - 1. Model architecture and algorithms.
      - 2. Training pipeline and dataset descriptions.

3. Statistical methods used for validation.
  4. Explanation of fairness assessments and steps to mitigate bias
- c. Results:
    - i. Intended equity objective and utilized fairness metric
    - ii. Evaluation metrics (e.g., accuracy, precision, recall, AUC-ROC), statistical significance tests, and comparative analysis..
    - iii. Comparison with baseline methods or alternatives.
    - iv. Visualization of results (e.g., confusion matrices, ROC curves).
  - d. Discussion:
    - i. Interpretation of findings, limitations, and implications for clinical practice.
    - ii. Intended users
    - iii. Known limitations
2. Dataset Documentation (if sharing data)
    - a. Data Description:
      - i. Features, data sources, and collection methods.
    - b. Data Preprocessing Steps:
      - i. Cleaning, normalization, and other transformations.
    - c. Data Dictionary:
      - i. Definitions of variables and labels.
    - d. Ethical and Legal Considerations:
      - i. How patient privacy was protected.
      - ii. Any restrictions on dataset usage.
  3. Open Access Repositories
    - a. Code Repository (e.g., GitHub, GitLab):
      - i. Source code with comprehensive comments and clear organization.
      - ii. Instructions for reproducing results (e.g., environment setup, dependencies).
    - b. Model Repository (e.g., Hugging Face, Zenodo):
      - i. Pre-trained models with detailed usage instructions.
      - ii. Associated metadata for easy reference.
  4. FAQs:
    - a. Common questions from reviewers, researchers, or practitioners.



## 5.2. Communication and Education

### Modifying trial educational materials

Well-designed educational materials help build confidence, trust and engagement across all stakeholders, enabling a smooth clinical integration of the AI-Solution. During this step the education materials made during the Clinical Trial should be modified for a broader audience. Some recommended additions to the above section on education materials (section 4.3) include:

1. end-users:
  - a. User Manuals and Quick Start Guides:
    - i. Expand to explain system functionalities, workflows and common troubleshooting steps.
  - b. Clinical Workflow Integration Training:
    - i. Provide interactive sessions or videos on how to use the AI-Solution.
  - c. FAQ and Troubleshooting Tips:
    - i. Address common concerns and technical issues.
2. For Organizational Stakeholders (e.g. administrators, IT):
  - a. Implementation Guides:
    - i. Highlight infrastructure, security, and resource requirements.
  - b. Compliance and Governance Briefs:
    - i. Outline regulatory, ethical, and operational responsibilities.

### Communication plan

The communication plan should assure that the team has the following information readily available, complete and free from reporting bias:

- a. Equity objectives
- b. Fairness metrics
- c. Model performance across defined subgroups
- d. Information about non-technical components of AI-solution
- e. Contact information of solution management team
- f. Educational material to avoid automation bias
- g. Educational material on system usage

Communication plans should encompass dissemination of regular reports and audits, patient and end-user educational materials, and adverse event reporting. Plans should also be generated in advance for events such as AI-solution update, AI-solution failure, and AI-solution decommissioning.

## 5.3. AI-solution Rollout and Monitoring

### Regular monitoring and reporting

The metrics that should be monitored and reported on will depend on the final AI-solution. However, it is recommended that the AI-solution monitoring focuses on performance, safety, compliance and operational impact. As examples: (1) model performance and data quality could be monitored by looking at error rates (false positives and false negatives), drift detection (in both data and model performance), and statistical metrics; (2) Clinical outcomes could be monitored by looking at workflow integration and adverse events, such as patient safety issues; (3) Operational metrics, security and privacy can be monitored by looking at uptime and response times, as well as usage metrics, data breaches and access logs; (4) finally, periodic revalidation of the AI-solution may be beneficial using updated data or in the case of a system update that impacts the software and hardware used by the AI-solution.

Utilizing the developed communication plan, regularly report on the AI-Solution's impact and performance, as determined by audits (section 4.2), feedback from end-users and patients, and regular monitoring metrics. These regular reports should be comprehensive and include disparities in AI-Solution performance, current clinical adoption and impact, as well as any adverse events that have occurred.

## 5.4. Updating or Decommissioning of AI-solution

If at any point an AI-solution is failing to meet the set objectives and requirements (i.e. fairness and performance metrics are worsening, institutional requirements are no longer being met, or end-users and/or patients are no longer seeing benefit from the solution), the cross-functional team should be consulted. During this consultation there are three possible next steps: Pausing the AI-Solution, Updating the AI-Solution, or Decommissioning the AI-Solution.

Each of these steps will look different depending on the institution and AI-Solution. Broadly speaking, it should be determined what each of these steps would look like from a clinical and regulatory perspective and who should be informed. Additionally, depending on the decided step, the results and/or changes will need to be communicated.

## References

1. Webster, C. S., Taylor, S., Thomas, C. & Weller, J. M. Social bias, discrimination and inequity in healthcare: mechanisms, implications and recommendations. *BJA Educ.* **22**, 131–137 (2022).
2. Pinto, M. F. Methodological and Cognitive Biases in Science: Issues for Current Research and Ways to Counteract Them. (2023).
3. Popovic, A. & Huecker, M. R. Study bias. in *StatPearls* (StatPearls Publishing, Treasure Island (FL), 2025).
4. Baumann, A. A. & Cabassa, L. J. Reframing implementation science to address inequities in healthcare delivery. *BMC Health Serv. Res.* **20**, 190 (2020).
5. National Academies of Sciences, Engineering, and Medicine *et al.* The Root Causes of Health Inequity. in *Communities in Action: Pathways to Health Equity* (National Academies Press (US), 2017).
6. Cortes, C., Mohri, M., Riley, M. & Rostamizadeh, A. Sample selection bias correction theory. in *Lecture Notes in Computer Science* 38–53 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2008).
7. selection bias.  
<https://www.cancer.gov/publications/dictionaries/cancer-terms/def/selection-bias> (2011).
8. Nickerson, R. S. Confirmation bias: A ubiquitous phenomenon in many guises. *Rev. Gen. Psychol.* **2**, 175–220 (1998).
9. Pilat, D. & Krastev, S. Confirmation Bias - Biases & Heuristics. *The Decision Lab*  
<https://thedecisionlab.com/biases/confirmation-bias> (2021).
10. Kea, B., Hall, M. K. & Wang, R. Recognising bias in studies of diagnostic tests part 2: interpreting and verifying the index test. *Emerg. Med. J.* **36**, 501–505 (2019).
11. Elston, D. M. Survivorship bias. *J. Am. Acad. Dermatol.* **0**, (2021).
12. Pasqualetti, F. *et al.* The impact of survivorship bias in glioblastoma research. *Crit. Rev.*

- Oncol. Hematol.* **188**, 104065 (2023).
13. Hawkins, D. M. The problem of overfitting. *J. Chem. Inf. Comput. Sci.* **44**, 1–12 (2004).
  14. Park, Y. & Ho, J. C. Tackling overfitting in boosting for noisy healthcare data. *IEEE Trans. Knowl. Data Eng.* **33**, 2995–3006 (2021).
  15. Usher-Smith, J. A., Sharp, S. J. & Griffin, S. J. The spectrum effect in tests for risk prediction, screening, and diagnosis. *BMJ* **353**, i3139 (2016).
  16. Abdelwanis, M., Alarafati, H. K., Tammam, M. M. S. & Simsekler, M. C. E. Exploring the risks of automation bias in healthcare artificial intelligence applications: A Bowtie analysis. *Journal of Safety Science and Resilience* **5**, 460–469 (2024).
  17. Sabin, J. A. Tackling implicit bias in health care. *N. Engl. J. Med.* **387**, 105–107 (2022).
  18. Bova, D. Do You Have ‘Shiny Object’ Syndrome? What It Is and How to Beat It. *Entrepreneur*  
<https://www.entrepreneur.com/living/do-you-have-shiny-object-syndrome-what-it-is-and-how-to/288370> (2017).
  19. Guo, L. N., Lee, M. S., Kassamali, B., Mita, C. & Nambudiri, V. E. Bias in, bias out: Underreporting and underrepresentation of diverse skin types in machine learning research for skin cancer detection-A scoping review. *J. Am. Acad. Dermatol.* **87**, 157–159 (2022).
  20. Seyyed-Kalantari, L., Zhang, H., McDermott, M. B. A., Chen, I. Y. & Ghassemi, M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nat. Med.* **27**, 2176–2182 (2021).
  21. Khor, S. *et al.* Racial and ethnic bias in risk prediction models for colorectal cancer recurrence when race and ethnicity are omitted as predictors. *JAMA Netw. Open* **6**, e2318495 (2023).
  22. Gichoya, J. W. *et al.* AI recognition of patient race in medical imaging: a modelling study. *Lancet Digit. Health* **4**, e406–e414 (2022).
  23. Reeder-Hayes, K. E. *et al.* Structural racism and treatment delay among Black and White

- patients with breast cancer. *J. Clin. Oncol.* JCO2302483 (2024).
24. Ring, A., Harder, H., Langridge, C., Ballinger, R. S. & Fallowfield, L. J. Adjuvant chemotherapy in elderly women with breast cancer (AChEW): an observational study identifying MDT perceptions and barriers to decision making. *Ann. Oncol.* **24**, 1211–1219 (2013).
  25. Protière, C., Viens, P., Rousseau, F. & Moatti, J. P. Prescribers' attitudes toward elderly breast cancer patients. Discrimination or empathy? *Crit. Rev. Oncol. Hematol.* **75**, 138–150 (2010).
  26. Ring, A. The influences of age and co-morbidities on treatment decisions for patients with HER2-positive early breast cancer. *Crit. Rev. Oncol. Hematol.* **76**, 127–132 (2010).
  27. Haase, K. R. *et al.* A scoping review of ageism towards older adults in cancer care. *J. Geriatr. Oncol.* **14**, 101385 (2023).
  28. Cirillo, D. *et al.* Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare. *NPJ Digit. Med.* **3**, 81 (2020).
  29. Lee, M. S., Guo, L. N. & Nambudiri, V. E. Towards gender equity in artificial intelligence and machine learning applications in dermatology. *J. Am. Med. Inform. Assoc.* **29**, 400–403 (2022).
  30. Larrazabal, A. J., Nieto, N., Peterson, V., Milone, D. H. & Ferrante, E. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 12592–12594 (2020).
  31. Turner, M. *et al.* A cancer geography paradox? Poorer cancer outcomes with longer travelling times to healthcare facilities despite prompter diagnosis and treatment: a data-linkage study. *Br. J. Cancer* **117**, 439–449 (2017).
  32. Ambroggi, M., Biasini, C., Del Giovane, C., Fornari, F. & Cavanna, L. Distance as a barrier to cancer diagnosis and treatment: Review of the literature. *Oncologist* **20**, 1378–1385 (2015).

33. Maly, R. C., Umezawa, Y., Leake, B. & Silliman, R. A. Mental health outcomes in older women with breast cancer: impact of perceived family support and adjustment. *Psychooncology* **14**, 535–545 (2005).
34. Bevan, J. L. & Pecchioni, L. L. Understanding the impact of family caregiver cancer literacy on patient health outcomes. *Patient Educ. Couns.* **71**, 356–364 (2008).
35. Mickel, J. The Importance of Multi-Dimensional Intersectionality in Algorithmic Fairness and AI Model Development. (2023).
36. Hajian-Tilaki, K. Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation. *Caspian J. Intern. Med.* **4**, 627–635 (2013).
37. Çorbacioğlu, Ş. K. & Aksel, G. Receiver operating characteristic curve analysis in diagnostic accuracy studies: A guide to interpreting the area under the curve value: A guide to interpreting the area under the curve value. *Turk. J. Emerg. Med.* **23**, 195–198 (2023).
38. Czakon, J. F1 Score vs ROC AUC vs Accuracy vs PR AUC: Which Evaluation Metric Should You Choose? *neptune.ai* <https://neptune.ai/blog/f1-score-accuracy-roc-auc-pr-auc> (2022).
39. Richardson, E. *et al.* The receiver operating characteristic curve accurately assesses imbalanced datasets. *Patterns (N. Y.)* **5**, 100994 (2024).
40. Bertels, J., Robben, D., Vandermeulen, D. & Suetens, P. Theoretical analysis and experimental validation of volume bias of soft Dice optimized segmentation maps in the context of inherent uncertainty. *Med. Image Anal.* **67**, 101833 (2021).
41. Taha, A. A. & Hanbury, A. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. *BMC Med. Imaging* **15**, 29 (2015).
42. Gorriz, J. M. *et al.* Is K-fold cross validation the best model selection method for Machine Learning? *arXiv [stat.ML]* (2024).
43. Jung, Y. & Hu, J. A K-fold averaging cross-validation procedure. *J. Nonparametr. Stat.* **27**, 167–179 (2015).

44. Carey, S., Pang, A. & Kamps, M. de. Fairness in AI for healthcare. *Future Healthc. J.* **11**, 100177 (2024).
45. Gao, J. *et al.* What is fair? Defining fairness in machine learning for health. *arXiv [cs.LG]* (2024).
46. Rajkomar, A., Hardt, M., Howell, M. D., Corrado, G. & Chin, M. H. Ensuring fairness in machine learning to advance health equity. *Ann. Intern. Med.* **169**, 866–872 (2018).
47. Chen, R. J. *et al.* Algorithmic fairness in artificial intelligence for medicine and healthcare. *Nat. Biomed. Eng.* **7**, 719–742 (2023).
48. Fabris, A., Esuli, A., Moreo, A. & Sebastiani, F. Measuring fairness under unawareness of sensitive attributes: A quantification-based approach. *J. Artif. Intell. Res.* **76**, 1117–1180 (2023).
49. Bou, V. Achieving Demographic Parity across multiple artificial intelligence applications: A new approach for real-time bias mitigation. *Preprints* (2024)  
doi:10.20944/preprints202412.0468.v1.
50. Davoudi, A. *et al.* Fairness gaps in Machine learning models for hospitalization and emergency department visit risk prediction in home healthcare patients with heart failure. *Int. J. Med. Inform.* **191**, 105534 (2024).
51. Roberts-Licklider, K. & Trafalis, T. Machine learning techniques with fairness for prediction of completion of drug and alcohol rehabilitation. *arXiv [cs.LG]* (2024).
52. Feng, Q., Du, M., Zou, N. & Hu, X. Fair Machine Learning in Healthcare: A Review. *arXiv [cs.LG]* (2022).
53. D'Souza, G., Zhang, H. H., D'Souza, W. D., Meyer, R. R. & Gillison, M. L. Moderate predictive value of demographic and behavioral characteristics for a diagnosis of HPV16-positive and HPV16-negative head and neck cancer. *Oral Oncol.* **46**, 100–104 (2010).
54. Kim, S.-Y. *et al.* Deep learning-based computer-aided diagnosis in screening breast

- ultrasound to reduce false-positive diagnoses. *Sci. Rep.* **11**, 395 (2021).
55. Feldman, M., Friedler, S., Moeller, J., Scheidegger, C. & Venkatasubramanian, S. Certifying and removing disparate impact. *arXiv [stat.ML]* (2014).
  56. Kamiran, F. & Calders, T. Classifying without discriminating. in *2009 2nd International Conference on Computer, Control and Communication* 1–6 (IEEE, 2009).
  57. View of SMOTE: Synthetic Minority Over-sampling Technique.  
<https://www.jair.org/index.php/jair/article/view/10302/24590>.
  58. Kamishima, T., Akaho, S., Asoh, H. & Sakuma, J. Fairness-aware classifier with prejudice remover regularizer. in *Machine Learning and Knowledge Discovery in Databases* 35–50 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2012).
  59. Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J. & Wallach, H. A reductions approach to fair classification. *arXiv [cs.LG]* (2018).
  60. Celis, L. E., Huang, L., Keswani, V. & Vishnoi, N. K. Classification with fairness constraints: A meta-algorithm with provable guarantees. *arXiv [cs.LG]* (2018).
  61. Yang, J., Soltan, A. A. S., Eyre, D. W., Yang, Y. & Clifton, D. A. An adversarial training framework for mitigating algorithmic biases in clinical machine learning. *NPJ Digit. Med.* **6**, 55 (2023).
  62. Fan, Y. *et al.* Quantification of mandibular sexual dimorphism during adolescence. *J. Anat.* **234**, 709–717 (2019).
  63. Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J. & Weinberger, K. Q. On Fairness and Calibration. *arXiv [cs.LG]* (2017).
  64. Franc, V., Prusa, D. & Voracek, V. Optimal strategies for reject option classifiers. *J. Mach. Learn. Res.* **24**, 1–49 (2023).
  65. Ruopp, M. D., Perkins, N. J., Whitcomb, B. W. & Schisterman, E. F. Youden Index and optimal cut-point estimated from observations affected by a lower limit of detection. *Biom. J.* **50**, 419–430 (2008).



66. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* **3**, 160018 (2016).
67. Chmielinski, K. S. *et al.* The Dataset Nutrition Label (2nd Gen): Leveraging Context to Mitigate Harms in Artificial Intelligence. *arXiv [cs.LG]* (2022).
68. Collins, G. S. *et al.* TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* **385**, e078378 (2024).