Appendix

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

Table of Contents

Table of Contents 1
1. General 4
1.1. Fairness Definitions 4
Bias 4
Inequity 4
Responsibility 4
1.2. Common Methodological Biases 5
1.3. Financial Considerations 7
2. Problem Identification and Study Design 8
2.1. Define the problem and the solution 8
Assembling your cross-functional team 8
Patient engagement and participation 8
Non-technical solutions 8
2.2. Focus on Equity in the Problem Space 9
Explore and document baseline inequities 9
Literature review (social determinants of health, sources of bias and inequity) 9
Consult with knowledgeable partners 10
Working with Health Information and Data from Vulnerable Populations 10
Health Information and Data from Black Individuals 11
First Nations, Indigenous, and Métis Health Information and Data 11
2.3. Outcome Measurements and Data Requirements 12

Assess AI Output Integration into Clinic 12
Outcome Measurement 12
Data requirements and availability 13
3. Model Training and Development 14
3.1. Appropriateness of Retrospective Data 14
Identify current clinical benchmarks 14
Assessing dataset for disparities and 'fit for purpose' 14
3.2. Defining Objectives and Metrics 14
Appropriate statistical measures of performance 14
Defining Equity Objectives 16
Measures of AI-Solution Fairness 17
Fairness metrics 17
Open-Source fairness metric libraries 18
3.3. Model Training and Testing 18
Acquisition of a third-party model 18
Data Representativeness and Quality 20
Mitigation methodologies 20
Data pre-processing 21
ML training methods 21
Prediction post-processing 22
4. Silent Deployment and Clinical Evaluation 23
4.1. Prospective Deployment Preparation 23
Approvals for prospective deployment 23
Retrospective to prospective mapping 23
Prospective Data Representativeness 24
Defining AI-Solution Failures and Adverse Events 24
4.2. Prospective Model Assessment (Silent Mode) 25
Silent Deployment Considerations 25
Many approaches to silent mode 25

Threshold selection 25
Assessing Silent Mode Performance 25
Define auditing methods 25
4.3. Prospective Clinical Evaluation 26
Education materials 26
Recruit end-users for future clinical integration 27
5. Operationalization and Lifecycle Monitoring 28
5.1 Preparation and Documentation 28
Comprehensive documentation 28
Regulatory/Compliance Documentation 28
Deployment Documentation for End Users 29
Methodological Documentation 31
1. Model Development and Technical Specifications 32
2. Deployment Documentation 32
3. Regulatory and Compliance Documentation 33
4. Monitoring and Post-Deployment Validation 33
5.2. Communication and Education 34
Modifying educational materials 34
Communication plan 35
Communicating to Patients and Care Partners 35
5.3. AI-Solution Rollout and Monitoring 36
Regular monitoring and reporting 36
Gathering Feedback for Continuous Quality Improvement 37
5.4. Updating or Decommissioning of AI-Solution 37
6. Community Tools 39
7. Version History 39
References 41

1. General

1.1. Fairness Definitions

While ‘bias’, ‘inequity’, and ‘responsibility’ 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” (Webster et al. 2022). Methodological bias can be defined as “systematic errors stemming from choices made during the research process”(Pinto 2023), such as selection bias, observer bias, response bias, and publication bias(Popovic and Huecker 2025). 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”(Baumann and Cabassa 2020). 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(National Academies of Sciences, Engineering, and Medicine et al. 2017). These systematic differences are influenced by social biases and by structural differential access opportunities.

Responsibility

A ‘responsible’ AI solution is one that is applied in a way that is the most appropriate (just, ethical) - but what is most appropriate for a given AI solution in a given situation is not static. The proceeding framework is adaptable to most definitions of ‘responsibility’ that could be chosen. An AI solution that is developed and deployed following all of the steps provided in this framework should implicitly follow, at minimum, the principals of equity, fairness, scientific rigor, and democratic engagement, but other values will be infused by choices made throughout the AI lifecycle (such as which patient populations are served or prioritized, how fairness is defined and to what extent the definition is enforced, how

information about the AI is disseminated to end users, and so on). Identifying which values are important to the project at the outset is critical to making these decisions downstream.

An AI solution could be considered responsible if at the end, it functions in alignment with the core principles of clinical ethics(Varkey 2021) (beneficence, non-maleficence, autonomy, and justice); or if it meets the standards of a pre-existing set of ethical standards such as those put forth by the Coalition for Health AI(Coalition for Health AI, n.d.) (usefulness, fairness, safety, transparency, and privacy). Alternatively, the guiding definition for responsibility for your solution may be that the AI contributes to the common good (Coeckelbergh 2024). However, when using a broader philosophical definition of responsibility, care must be taken to operationalize it appropriately. For example, in the case of defining responsible AI as contributing to the common good, who is the "common" - the patients served by your institution? All residents in the city of your institution? The nation? Humanity? After defining "common", what is the definition of "good"? Who gets to have input on that definition?

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, validation 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. It should also be noted, that as AI methodologies evolve, so do the types of biases that may be present. As an example, as LLMs have become more commonplace, biases as they relate to the presentation of information and interaction between humans and machines requires consideration (Robinig and Wallner 2026).

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)(Cortes et al. 2008; "Selection Bias" 2011). 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 (Nickerson 1998; Pilat and Krastev 2021).
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(Kea et al. 2019).
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(Elston 2021; Pasqualetti et al. 2023).
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(Hawkins 2004; Park and Ho 2021).
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(Usher-Smith et al. 2016).
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(Abdelwanis et al. 2024).
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(Sabin 2022).
9) **Presentation Bias**: If information generated by an AI is not presented in a neutral and transparent way, user behavior may be influenced, resulting in subsequent

biased actions. This is particularly important when utilizing LLMs (Mehrabi et al. 2021).

10) **Interaction Bias:** Users that iteratively engage with AI systems (e.g. chatbots) contribute to bias by bringing their own implicit biases and expectations. These beliefs and expectations affect how they interact with AI systems by influencing what they accept, what they question, and how they present system queries. In turn, the AI system begins to present outputs in a way that matches the user's beliefs, further contributing to confirmation and implicit biases (Robinig and Wallner 2026).

1.3. Financial Considerations

Addressing bias, fairness and inequity in AI and ML problem spaces requires additional resources. It is encouraged to consider this guideline in the early stages of problem development to define a realistic plan for incorporating responsible AI strategies into the project. Responsible AI development resources can be incorporated into grant applications through budgeting and through collaborators. In the case that additional funds are not available for addressing bias, fairness and equity in development, some resources are available through open-source community tools (Section 6).

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 and/or caregiver representatives, clinical champions and end-users.

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 must be present in the whole spectrum of AI research, from problem identification and study design through to clinical implementation and lifecycle monitoring. This becomes even more imperative as we begin to develop and deploy chatbots and triaging solutions that may interact directly with patients or caregivers.

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 (Strategy for Patient-Oriented Research - SPOR) - similar initiatives exist in other countries. There are a number of validated tools to assess how well patients were engaged in the research process, such as the PEIRS-22 tool (Hamilton et al. 2021) and the REST (Goodman et al. 2022) survey. These can be helpful both to validate that patient and caregiver partners have been successfully brought into the cross-functional team, and to serve as a resource for thinking about how to engage partners in the first place.

Non-technical solutions

The allure of AI in healthcare often creates a "shiny object syndrome" (Bova 2017) 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, unfairness, 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 (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 (Guo et al. 2022; Seyyed-Kalantari et al. 2021; Khor et al. 2023; Gichoya et al. 2022; Reeder-Hayes et al. 2024), age (A. Ring et al. 2013; Protière et al. 2010; Alistair Ring 2010; Haase et al. 2023), sex (Cirillo et al. 2020; Lee et al. 2022; Larrazabal et al. 2020), geographical location of residence(Turner et al. 2017; Ambroggi et al. 2015), patient support systems(Maly et al. 2005; Bevan and Pecchioni 2008), 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 (Mickel 2023); 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 and caregiver experience in the specific problem space, as recommended in our framework, the team 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 (see Section 2.1 - Assembling your cross-functional team). 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 Health Information and Data from Vulnerable Populations

In all healthcare AI contexts, there is a probability that your dataset will include health information from or about individuals who belong to underserved/vulnerable populations. It is always best practice to be mindful of whose data you are stewarding and to what extent you are obliged to work with people from the community through the course of your project. For more vulnerable groups, proportionally more consideration must be given to reduce the risk of data misuse or harm. This is **especially** critical if your AI-Solution either a) primarily ingests data from one or more specific vulnerable populations or b) will primarily be used to serve one or more specific vulnerable populations.

Here we provide overarching information on working with health information from Black communities and from Indigenous, First Nations, and Metis communities, as these groups have been particularly vulnerable to medical and health research-related harms(Sheikh et al. 2023). It is the study team's responsibility to consider which other vulnerable populations may be impacted by their AI-Solution and how to manage these impacts. We recommend paying attention to groups that are politically vulnerable (for example, refugee populations, unsheltered individuals, transgender and gender diverse individuals, individuals with experience in the carceral system).

Health Information and Data from Black Individuals

In 2021, in response to concerns around race-based data collection during the Covid-19 pandemic, the Black Health Equity Working Group developed a framework to support data sovereignty in Black communities. The Engagement, Governance, Access, and Protection (EGAP) framework supports researchers and healthcare workers to "understand where to start if you are involved in the collection, management, analysis and use of race-based data for health purposes"(Ca 2021). The framework specifically calls out the importance of community engagement across all four pillars, as well as a focus on meaningful and systemic dismantling of structural racism, as key components of success.

An implementation guide for the EGAP framework was developed by the Wellesley Institute (a Toronto-area health thinktank) in 2024 (Demeke et al. 2024). In addition to providing more background on the development and importance of the EGAP framework, this guide helps walk teams through its application from developing organizational readiness through to ongoing data access governance.

First Nations, Indigenous, and Métis Health Information 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 limited if systemic barriers exist that prevent recommended actions. These barriers may include lack of available interventions (e.g., an output recommending the patient switches treatment lines, when there are no treatments remaining that the patient is naïve to), resource limitations within the institution (e.g., an output that recommends referral to palliative care when the department does not have capacity for more patients), or resource limitations within the broader health system (e.g., an output that recommends discharging patients to long term care centres when these facilities do not have capacity). The utility of AI-Solutions in such situations must 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 these patients 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. Many 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 to consider whether PHI is required and if so, for exactly what purpose, or whether the data needs to be de-identified or anonymized (Section 3.1).

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; this 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 only statistically significant but clinically meaningful as well, 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. Reliability and currentness of the data should also be assessed. This is particularly important when dealing with LLMs that utilize RAG-based systems that can suffer from stale knowledge sources(Koduri 2026).

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)(Hajian-Tilaki 2013; Çorbacıoğlu and Aksel 2023) and PR-AUC (Precision-Recall Area Under the Curve)(Czakon 2022; Richardson et al. 2024) 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 (Bertels et al. 2021; Taha and Hanbury 2015). 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(Gorriz et al. 2024; Jung and Hu 2015). 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 (Hawkins 2004). 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

While the specific equity objectives for a project are informed by the inequity(/inequities) present in the problem space, the overarching type of objective can be broadly categorized as:

1) Preserving the current clinical state (i.e. the AI-solution doesn't worsen inequities after it is implemented).
2) Improving the clinical state for all groups as equally as possible, following the principles of formal justice.
3) Preferentially improving the clinical state for a historically underserved group, following the principles of distributive justice.

To exemplify how an equity objective might be defined, consider the scenario presented in Outcome Measurement, in which an AI-Solution is deployed to assist in managing post-surgical pain crises.In this theoretical example, there has been a historic disparity between white patients (whose pain has been treated appropriately most but not all of the time) and Black patients (who have more often been undertreated for more severe pain) (Hoffman et al. 2016; Nguemeni Tiako et al. 2026).

If the equity objective was to make sure that the clinical state stayed the same post AI-deployment, it might be expressed as an objective like "the rates of analgesic prescriptions for white patients and for Black patients does not decrease post-deployment". This would ensure that the model did not worsen under-prescribing of pain medication for Black patients based on historic trends, but it may or may not close the gap between white and Black patients' pain management.

If the equity objective was to produce a clinical state where all groups are benefited as equally as possible, it may be expressed as an objective that "a higher proportion of patients receive analgesics *and* report lower pain post-deployment". This objective captures individual benefit regardless of subgroup belonging, but it risks obscuring if the AI-solution is favoring white or Black patients unduly.

If the equity objective was to focus on benefiting the historically underserved group the most (in this case, Black patients), the objective might be expressed as "the rate of analgesics prescribed to Black patients increased, *and* self-reported pain of Black patients decreased, post-deployment". This objective would ensure that medication is

being prescribed to Black patients more often in the appropriate situations, but it would not specifically measure the effect of the AI-solution on white patients.

In addition to choosing a type of equity objective, it may be valuable for teams to choose a magnitude of impact to attach to their objective. For example, instead of phrasing their objective as "a higher proportion of patients receive analgesics *and* report lower pain post-deployment", they may want to state their objective as "10% more patients receive analgesics, *and* the average patient treated with analgesics reports 20% lower pain post-deployment". Having these concrete magnitudes attached to the objectives allows for a clearer delineation of when the AI-solution is succeeding in deployment, and when it is in need of updating or decommissioning.

Decisions on what type of equity objectives to choose, and what (if any) magnitude to assign to them, should be made in consultation with the entire cross-functional team. While patient and caregiver partners should be involved in the entire lifecycle of the project, it is particularly important to engage them in deciding what the equity objectives are, as this measure is directly related to the magnitude of clinical impact the AI-solution will have.

Measures of AI-Solution 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 (Carey et al. 2024; Gao et al. 2024; Rajkomar et al. 2018; Chen et al. 2022).

Fairness metrics

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

1) Group Unawareness: Group unawareness means that a model does not use a sensitive variable during prediction(Fabris et al. 2023). 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(Bou 2024).

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 (Rajkomar et al. 2018; Davoudi et al. 2024).
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 (Roberts-Licklider and Trafalis 2024; Q. Feng et al. 2022).
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 (D'Souza et al. 2010).
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 (Kim et al. 2021).

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.
5) UnBias-Plus: Tool kit to assess biased language in text

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 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 Section 3.3 of this 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 (Feldman et al. 2014) and "massaging"(Kamiran and Calders 2009). 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)("View of SMOTE: Synthetic Minority Over-Sampling Technique," n.d.) 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(Kamishima et al. 2012), Exponentiated Gradient Reduction(Agarwal et al. 2018), Grid Search Reduction(Agarwal et al. 2018) and Meta Fair Classifier(Celis et al. 2018) 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(Yang et al. 2023). For example, 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(Fan et al. 2019).

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(Pleiss et al. 2017) 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(Franc et al. 2023) 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 Evaluation

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. Which type of institutional approval to seek 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 (REB and QIRC for reference).

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 the deployed model, and could they have a direct impact on clinical decision-making?

Retrospective to prospective mapping

Healthcare data and EHR systems are dynamic and subject to ongoing change. AI models are frequently developed and evaluated using historical datasets and legacy record systems. Due to this, developers should carefully assess the extent to which these data reflect the prospective healthcare data sources for which the model is intended and clearly characterize any differences. Potential spectrum effects and unintended data disparities should also be evaluated, as they may affect model performance and generalizability.

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. 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. This is also relevant for third party, or vendor obtained, models.

Defining AI-Solution Failures and Adverse Events

Before any prospective deployment of an AI-solution, the team should define pre-specified thresholds for AI failures and critical incidents. An AI *failure* refers to a deviation from the target level of performance that remains within the range reasonably anticipated during research and development. In contrast, a *critical incident or adverse event* refers to a deviation that exceeds expected limits or that causes, or has the potential to cause, harm.

For example, if a medical chatbot has been implemented to allow clinicians to ask for up-to-date medication recommendations for their patients, there may be failure if:

- when prompted by the clinician, the chatbot returns no answer at all
- when asked for a list of options, the chatbot returns a medication recommendation that is nonsense (hallucinated)

However, there may be an incident if:

- when prompted by the clinician, the chatbot returns offensive/derogatory language
- when asked about a certain medication, the chatbot returns the names and contact information for all patients enrolled in a clinical trial related to that medication (a privacy incident)

Some situations are clear cut as either failures or critical incidents (for example, any physical harms to patients would be critical incidents). However, some situations may be less clear. The entire cross-functional team should be engaged to deliberate on the

bounds of expected and unexpected AI-solution behaviour, and in particular, ethicists and patient and caregiving partners' expertise should be drawn on. There are situations in which a clinician or researcher may perceive deviations in the AI solution to be a minor issue, but a patient on the receiving end may feel harmed.

4.2. Prospective Model Assessment (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 depend on the AI 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 model's performance using prospective data; (3) reliability testing and integration within the clinical workflow under real-world settings; and (4) identification of potential biases, usability issues and any unintended consequences. Importantly, silent mode is not designed to impact patient outcomes and care must be taken to isolate AI solution outputs from clinical decision making.

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(Ruopp et al. 2008). These thresholds should be selected with consideration of identified patient subgroups.

Assessing Silent Mode Performance

Silent mode evaluation should assess both how the model performs in a real-world clinical setting and how effectively the AI-solution integrates into existing clinical workflows. If the model does not meet predefined performance targets, equity objectives, or fairness metrics, the cross-functional team should be consulted to determine whether the model requires adaptation, additional mitigation strategies should be implemented, or the AI-Solution is unlikely to achieve its intended purpose and should be discontinued.

Define auditing methods

In collaboration with the cross-functional team, an auditing plan should be generated prior to clinical 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 Evaluation

Education materials

Developing education materials to provide during the Clinical Evaluation 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 Evaluation, depending on the stakeholders involved.

1. For 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 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 and risks.
    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.
    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 Evaluation, obtaining feedback from end-users and patients 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)(Wilkinson et al. 2016) when publishing; and

(3) including a revision history of any updates and improvements to the AI-Solution.

Below are some more specific recommendations for documentation for internal usage and external publishing.

For healthcare purposes, AI documentation can be categorized into one of three types: (1) Regulatory and compliance documentation; (2) user and installation guides; and (3) methodological documentation.

Regulatory/Compliance Documentation

In the context of healthcare software and AI deployment, regulation and compliance refer to the frameworks, standards, and legal requirements that ensure digital health technologies are safe, effective, secure, and used responsibly. These requirements govern areas such as patient privacy, data security, clinical safety, transparency, risk management, and software quality, helping developers and healthcare organizations ensure that AI-enabled tools meet regulatory expectations while maintaining patient trust and protecting health outcomes.

Currently, there is no comprehensive AI-specific legislative framework governing healthcare applications in Canada. Furthermore, formal clinical trials are not universally required before AI systems are implemented in healthcare settings, although regulatory, validation, and governance requirements may still apply depending on the intended use and risk level of the technology. However, as policy landscape continues to evolve rapidly. Recent developments, including Ontario's Enhancing Digital Security and Trust Act, 2024 (Bill 194) and related responsible AI initiatives, signal increasing expectations for

transparency, accountability, risk management, and oversight of AI systems used within the public sector, including healthcare organizations. In many use cases, healthcare-based AI software impacts patient care in a similar manner to that of tools which are regulated as medical devices and, according to Health Canada, can be classified as a medical device(Health Canada 2021). Regardless of legal obligation, documentation of AI methods, data collection, validation, and testing is important.

In lieu of regulation, there are internation standards available for best practices which can serve as a guide. International standards (ISO) for AI have been drafted and should be reviewed before deployment[REF].

For further information, refer to ISO 42001:2023 for International Standard for Information technology – Artificial intelligence – Management System, ISO 13485:2016 for the international standard for Medical Devices, and ISO 14155:2026 for Clinical investigation of medical devices for human subjects — Good clinical practice.

Data Retention

One key component of regulatory documentation is the length and method of retaining data related to a healthcare tool. There are institutional, national, and international guidelines for data retention which vary based on the type of healthcare technology. It is important to research and confirm all policies that apply to all institutions or clinical settings where the AI solution will be deployed.

Health Canada does not have a data retention policy specific to AI at the present time; however, established guidelines for medical treatments and devices can be used to provide guidance on reasonable best practices. Under the Food and Drug Regulations and Natural Health Products Regulations, Health Canada requires all clinical trial data to be retained in a secure location for 15 years (Health Canada 2026a). For medical devices, regulations require all data pertaining to the device to be stored in a secure location for a minimum of 2 years post-distribution of the device or up to the life-cycle of the device (Health Canada 2026b).

Deployment Documentation for End Users

User and installation guides are important to ensure the AI tool is deployed and utilized correctly. Performance metrics, training, validation, and deployment of AI models are based on a specific application for many AI models. Even an AI tool or model that is

developed following best practices to ensure fair, ethical and unbiased outcomes for patients can be subject to poor clinical outcome if utilized incorrectly. It is vital that the user is aware of the correct usage, limitations of the model, applicable clinical use cases, and appropriate patient data.

Installation Guide

Comprehensive installation guides ensure that AI models are correctly integrated into clinical workflow, properly ingest patient data, and integrate seamlessly with existing clinical tools and software. Ensure that installation guides are clear, concise, and complete.

Important features of installation documentation include:

1. Descriptions of the applicable data to be used in the model, including the correct format and source for the data.
2. A description of how model output should be integrated with existing healthcare infrastructure (if applicable).
3. The dependencies that are required for the installation. A list of software dependencies and versions should be recorded.
4. Any regular or scheduled maintenance requirements with instructions for performing maintenance procedures.

Installation guidelines should be written in common language, with step-by-step instructions and screenshots to ensure the reader can perform the installation accurately. Different operating systems and common variations of the user's baseline infrastructure should be considered.

User-Guide and User-Facing Model Documentation

A user-guide is important to ensure that the AI tool is used and interpreted correctly once the end user has access to the tool. It should use common language (i.e. minimal jargon) and should be easily understood by a lay person with no knowledge of AI or AI methodology. This documentation should encompass:

1. What the tool does and its intended use. It should be clear how the tool is intended to be used in the workflow and how the output is interpreted. For example, it should be clear if the tool is intended as decision support, to automate a process, or to make clinical decisions. It is also important to clarify if the tool modifies the standard workflow (e.g. by replacing a human).

2. How the output should be interpreted to guide decisions or improve workflow, and the user's role in interpretation of the results.
3. Methods and processes for the user to validate or verify the output of the AI model with original data and/or review the accuracy of the model output.
4. Step-by-step user's manual to ensure that the user understands the tool's workflow and features.

In addition, the user should have access to clear, jargon-free methodological and performance documentation, including any significant model performance differences between patient sub-groups or use-cases.

Methodological Documentation

It is important to ensure that the methods used to create the AI solution are documented thoroughly. Methodology encompasses all elements of the technological design of the AI solution. For third party AI solutions, or AI solutions designed for commercial sale, detailed methodology documentation may be proprietary information and therefore cannot be shared with the user. However, when possible, methodological documentation is recommended.

Methodological documentation should include information about the underlying model architecture, algorithms, and training pipeline. It should detail the datasets used for training, validation and testing, including any methods for data augmenting, cleaning, or modification used.

Performance of the model should be included in methodological documentation. When reporting performance metrics, context is important to ensure that the results are correctly interpreted. The performance metric reported will depend on the type of AI solution used, and thus a description of the metric should be included. The documentation should report if the metric is calculated from retrospective or prospective data, as well as a description of the data used for determining performance. It is recommended that both retrospective and prospective metrics are calculated and reported.

Finally, a frame of reference for a given metric is highly beneficial for interpretation. For example, for AUROC, one may inform the reader that a measurement of 0.5 means that the solution is correct 50% of the time, and is equivalent to a random guess.

If the AI model is published in an academic journal, this information will likely be documented in the accompanying manuscript. However, an internal summary document

containing a simplified description available for end-users is beneficial as the user may not be able to correctly interpret or even access the manuscript. Key items to include in the summary document are:

1. Detailed explanation of algorithms, architecture, and training pipeline.
2. Description of datasets used for training, validation, and testing.
3. Performance evaluation metrics (e.g., accuracy, precision, recall, AUC-ROC).
4. Description of known biases and inequities in AI-Solution.

Please find a list of considerations for documentation below.

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.
        iii. Description of anticipated clinical impact.
    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(Chmielinski et al. 2022)
    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 (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:

When publishing in journals or open access repositories please consider the following documentation. Additionally, the TRIPOD+AI(Collins et al. 2024) statement is an excellent resource that should be followed for external publishing of an AI-Solution.

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 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 Prospective Clinical Evaluation should be modified for a broad 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 enure that the team has the following information readily available:

1. Equity objectives, as defined at the start of the project.
2. Fairness metrics, as defined at the start of the project.
3. Model performance (general and in relation to fairness metrics) across defined subgroups.
4. Information about non-technical components of AI-Solution.
5. Contact information of solution management team.
6. Educational material to avoid automation bias.
7. Educational material on system usage.

It is important to ensure that the information is complete and, to the extent possible, free from reporting bias(Sinha and Montori 2006). Reporting bias could manifest in a variety of ways. For example, regular performance reports could under-report unexpected or undesirable results or overemphasize expected or desirable results, or educational materials could obfuscate the intended use of AI-Solution in relation to how it is actually being deployed. In many cases, these reporting errors are not made maliciously, which makes them all the more important to stay alert for.

Communication plans should encompass dissemination of regular reports and audits, and patient and end-user educational materials. Plans should also be generated in advance for events such as adverse event reporting (see Communicating About Adverse Events), AI-Solution update, AI-Solution failure, and AI-Solution decommissioning.

Communicating to Patients and Care Partners

For explainers, reports, and educational materials that will be provided to patients and care partners, consideration should be given to making the language and figures lay-accessible. There are many resources on the web that can be used to help convert medical jargon into more lay-friendly synonyms, such as the CDC's "Everyday Words for

Public Health Communication" search tool. While many organizations encourage developing materials to be understandable to those with a Grade 6 reading level, there are no readability formulas that reliably evaluate the interpretability of complex information(Government of Canada et al. 2026). Instead of focusing on the *readability* of patient and care partner materials, it is recommended that teams focus on *usability*, by using plain language wherever possible. The International Plain Language Federation considers a document to be in plain language if the communication's "wording, structure, and design are so clear that the intended readers can easily find what they need, understand what they find, and use that information." (International Plain Language Federation 2019) Plain language can be facilitated with intentional word choice, well placed headings and subtitles, and graphic design.

In addition to being lay-friendly, teams should strive to make patient and care partner materials accessible, for example for those with low vision/vision loss. There are simple steps teams can take to make material accessible for individuals with low vision, as laid out by the Canadian National Institute for the Blind's Clear Print Guidelines. These include printing in black and white, keeping font size to a minimum of 12 to 18 pt font, using a medium heaviness font throughout most of the printed material, and printing on matte or non-glossy papers to minimize glare. Teams may note that some of the principles of plain language communication synergize well with principles of accessible design.

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; and
4. 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 (J. Feng et al. 2022).

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

Gathering Feedback for Continuous Quality Improvement

Very little literature exists on how to implement patient and end-user feedback in clinically integrated AI-solutions. However, there is existing literature on integrating such feedback into other developed clinical workflows(Nordberg et al. 2021), which can be used as a scaffold.

It should be expected that if feedback is gathered from an appropriately diverse group of end users, some feedback will conflict – for example, for an AI-solution meant to flag patients at risk for pain crisis, some nurses may want to be alerted frequently to an unlikely chance of crisis, while others may only want rare alerts for highly likely crises. Opportunity to improve the quality of the AI-solution overall lies in resolving these previously unaddressed tensions(Nordberg et al. 2021). Taking the previous example, the cross functional team may decide to implement two versions of the model depending on the time of day, so that nurses are alerted frequently during the day to control patients' pain, but less frequently during the night, when pain control measures might only serve to disturb the patients' sleep. Alternatively, the team might give only one nurse per shift access to the less sensitive model, and ask that nurse to decide, based on how well resourced the floor is, whether or not to follow up on low-certainty alerts.

Time must be taken to gather feedback specifically from patients affected by the AI-solution deployment. Patient feedback should be used to provide face validity for the impact of the AI-solution – if patients are noticing a primarily negative change, the team needs to evaluate if the clinical integration of the AI-solution is going as planned. Additionally, by continuously evaluating feedback from patients, change in the clinical environment and subsequent model drift may be detected earlier than by using numeric indicators of performance alone.

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.

6. Community Tools

There are many free tools and resources to help in identifying, mitigating and evaluating bias and fairness. Below is a table containing some resources.

Resource	Summary	Source
Data Nutrition Labels	The Data Nutrition Project aims to create a standard label for interrogating datasets. Website provides tools to evaluate data quality, fairness and bias in data.	https://datanutrition.org/
JarvAIs	An auto-machine learning tool for data quality assurance, training, and explaining data models. Includes input data bias analysis and model bias analysis.	https://github.com/pmcdi/jarvais
Med-ImageNet	DICOM image data processing tool. Unifies data format and indexes TCIA for easy search and download.	https://github.com/bhklab/med-imagenet
FairLearn	An open-source python package designed for metric calculation and reduction of bias in algorithms.	https://fairlearn.org/
Fairness Indicators	A package from Google designed to work with TensorFlow. Can be used for evaluation and visualization of group disparities.	https://www.tensorflow.org/tfx/guide/fairness_indicators
AIF360	An open-source package used to detect and mitigate biases. It is known for its extensive documentation and tutorials.	https://aif360.res.ibm.com/
Themis-ML	An open-source package for easy integration with scikit-learn. It is used mainly for binary classes and evaluation.	https://www.themisai.io/

7. Version History

Revision	Release Date	Author/Title	Change Description
1.0	Mar-20-2026	Mattea Welch, Benjamin Grant, Christopher Deutschman, and Clare McElcheran	Original release.
1.1	Nov-2-2026	Mattea Welch, Clare McElcheran, and Christopher Deutschman	Revised documentation section to include regulatory bodies Added LLM examples throughout Added EGAP framework

References

Abdelwanis, Moustafa, Hamdan Khalaf Alarafati, Maram Muhanad Saleh Tammam, and Mecit Can Emre Simsekler. 2024. “Exploring the Risks of Automation Bias in Healthcare Artificial Intelligence Applications: A Bowtie Analysis.” *Journal of Safety Science and Resilience* 5 (4): 460–469.

Agarwal, Alekh, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. 2018. “A Reductions Approach to Fair Classification.” In *arXiv [cs.LG]*. March 6. arXiv. http://arxiv.org/abs/1803.02453.

Ambroggi, Massimo, Claudia Biasini, Cinzia Del Giovane, Fabio Fornari, and Luigi Cavanna. 2015. “Distance as a Barrier to Cancer Diagnosis and Treatment: Review of the Literature.” *The Oncologist* 20 (12): 1378–1385.

Baumann, Ana A., and Leopoldo J. Cabassa. 2020. “Reframing Implementation Science to Address Inequities in Healthcare Delivery.” *BMC Health Services Research* 20 (1): 190.

Bertels, Jeroen, David Robben, Dirk Vandermeulen, and Paul Suetens. 2021. “Theoretical Analysis and Experimental Validation of Volume Bias of Soft Dice Optimized Segmentation Maps in the Context of Inherent Uncertainty.” *Medical Image Analysis* 67 (101833): 101833.

Bevan, Jennifer L., and Loretta L. Pecchioni. 2008. “Understanding the Impact of Family Caregiver Cancer Literacy on Patient Health Outcomes.” *Patient Education and Counseling* 71 (3): 356–364.

Bou, Victoria. 2024. “Achieving Demographic Parity across Multiple Artificial Intelligence Applications: A New Approach for Real-Time Bias Mitigation.” In *Preprints*. December 5. https://doi.org/10.20944/preprints202412.0468.v1.

Bova, Dan. 2017. “Do You Have ‘Shiny Object’ Syndrome? What It Is and How to Beat It.” Entrepreneur, February 9. https://www.entrepreneur.com/living/do-you-have-shiny-object-syndrome-what-it-is-and-how-to/288370.

Ca, Blackhealthequity. 2021. *Engagement, Governance, Access, and Protection (EGAP): A Data Governance Framework for Health Data Collected from Black Communities*. https://blackhealthequity.ca/wp-content/uploads/2021/03/Report_EGAP_framework.pdf.

Carey, Siân, Allan Pang, and Marc de Kamps. 2024. “Fairness in AI for Healthcare.” *Future Healthcare Journal* 11 (3): 100177.

Celis, L. Elisa, Lingxiao Huang, Vijay Keswani, and Nisheeth K. Vishnoi. 2018. “Classification with Fairness Constraints: A Meta-Algorithm with Provable Guarantees.” In *arXiv [cs.LG]*. June 15. arXiv. http://arxiv.org/abs/1806.06055.

Chen, Richard J., Tiffany Y. Chen, Jana Lipkova, et al. 2022. “Algorithm Fairness in AI for Medicine and Healthcare.” In *arXiv*. March 24. arXiv. https://doi.org/10.48550/arXiv.2110.00603.

Chmielinski, Kasia S., Sarah Newman, Matt Taylor, et al. 2022. “The Dataset Nutrition Label (2nd Gen): Leveraging Context to Mitigate Harms in Artificial Intelligence.” In *arXiv [cs.LG]*. January 10. arXiv. http://arxiv.org/abs/2201.03954.

Cirillo, Davide, Silvina Catuara-Solarz, Czuee Morey, et al. 2020. “Sex and Gender Differences and Biases in Artificial Intelligence for Biomedicine and Healthcare.” *Npj Digital Medicine* 3 (1): 81.

Coalition for Health AI. n.d. “Responsible Health AI for All.” Accessed October 26, 2026. https://www.chai.org/.

Coeckelbergh, Mark. 2024. “Artificial Intelligence, the Common Good, and the Democratic Deficit in AI Governance.” *AI and Ethics* 5 (2): 1491–1497.

Collins, Gary S., Karel G. M. Moons, Paula Dhiman, et al. 2024. “TRIPOD+AI Statement: Updated Guidance for Reporting Clinical Prediction Models That Use Regression or Machine Learning Methods.” *BMJ (Clinical Research Ed.)* 385 (April): e078378.

Çorbacıoğlu, Şeref Kerem, and Gökhan Aksel. 2023. “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.” *Turkish Journal of Emergency Medicine* 23 (4): 195–198.

Cortes, Corinna, Mehryar Mohri, Michael Riley, and Afshin Rostamizadeh. 2008. “Sample Selection Bias Correction Theory.” In *Lecture Notes in Computer Science*. Lecture Notes in Computer Science. Springer Berlin Heidelberg.

Czakon, Jakub. 2022. “F1 Score vs ROC AUC vs Accuracy vs PR AUC: Which Evaluation Metric Should You Choose?” Neptune.ai, July 21. https://neptune.ai/blog/f1-score-accuracy-roc-auc-pr-auc.

Davoudi, Anahita, Sena Chae, Lauren Evans, et al. 2024. “Fairness Gaps in Machine Learning Models for Hospitalization and Emergency Department Visit Risk Prediction in Home Healthcare Patients with Heart Failure.” *International Journal of Medical Informatics* 191 (105534): 105534.

Demeke, Jemal, Brenda Roche, Camille Orridge, and Kwame McKenzie. 2024. “Engagement, Governance, Access and Protection (EGAP) Implementation Guide.” Wellesley Institute, July 29. https://www.wellesleyinstitute.com/engagement-governance-access-and-protection-egap-implementation-guide/.

D’Souza, Gypsyamber, Hao H. Zhang, Warren D. D’Souza, Robert R. Meyer, and Maura L. Gillison. 2010. “Moderate Predictive Value of Demographic and Behavioral Characteristics for a Diagnosis of HPV16-Positive and HPV16-Negative Head and Neck Cancer.” *Oral Oncology* 46 (2): 100–104.

Elston, Dirk M. 2021. “Survivorship Bias.” *Journal of the American Academy of Dermatology* 0 (0). https://doi.org/10.1016/j.jaad.2021.06.845.

Fabris, Alessandro, Andrea Esuli, Alejandro Moreo, and Fabrizio Sebastiani. 2023. “Measuring Fairness under Unawareness of Sensitive Attributes: A Quantification-Based Approach.”

The Journal of Artificial Intelligence Research 76 (April): 1117–1180.

Fan, Yi, Anthony Penington, Nicky Kilpatrick, et al. 2019. “Quantification of Mandibular Sexual Dimorphism during Adolescence.” *Journal of Anatomy* 234 (5): 709–717.

Feldman, Michael, Sorelle Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2014. “Certifying and Removing Disparate Impact.” In *arXiv [stat.ML]*. December 11. arXiv. http://arxiv.org/abs/1412.3756.

Feng, Jean, Rachael V. Phillips, Ivana Malenica, et al. 2022. “Clinical Artificial Intelligence Quality Improvement: Towards Continual Monitoring and Updating of AI Algorithms in Healthcare.” *Npj Digital Medicine* 5 (1): 66.

Feng, Qizhang, Mengnan Du, Na Zou, and Xia Hu. 2022. “Fair Machine Learning in Healthcare: A Review.” In *arXiv [cs.LG]*. June 29. arXiv. http://arxiv.org/abs/2206.14397.

Franc, Vojtech, Daniel Prusa, and Vaclav Voracek. 2023. “Optimal Strategies for Reject Option Classifiers.” *Journal of Machine Learning Research: JMLR* 24 (11): 1–49.

Gao, Jianhui, Benson Chou, Zachary R. McCaw, et al. 2024. “What Is Fair? Defining Fairness in Machine Learning for Health.” In *arXiv [cs.LG]*. June 13. arXiv. http://arxiv.org/abs/2406.09307.

Gichoya, Judy Wawira, Imon Banerjee, Ananth Reddy Bhimireddy, et al. 2022. “AI Recognition of Patient Race in Medical Imaging: A Modelling Study.” *The Lancet. Digital Health* 4 (6): e406–e414.

Goodman, Melody S., Nicole Ackermann, Zoé Haskell-Craig, Sherrill Jackson, Deborah J. Bowen, and Vetta L. Sanders Thompson. 2022. “Construct Validation of the Research Engagement Survey Tool (REST).” *Research Involvement and Engagement* 8 (1): 26.

Gorriz, Juan M., R. Martin Clemente, F. Segovia, J. Ramirez, A. Ortiz, and J. Suckling. 2024. “Is K-Fold Cross Validation the Best Model Selection Method for Machine Learning?” In *arXiv [stat.ML]*. January 29. arXiv. http://arxiv.org/abs/2401.16407.

Government of Canada, Public Services, Procurement Canada, and Translation Bureau. 2026. “Readability Formulas, Programs and Tools: Do They Work for Plain Language?” October 26. https://our-languages.canada.ca/en/blogue-blog/readability-formulas-eng.

Guo, Lisa N., Michelle S. Lee, Bina Kassamali, Carol Mita, and Vinod E. Nambudiri. 2022. “Bias In, Bias out: Underreporting and Underrepresentation of Diverse Skin Types in Machine Learning Research for Skin Cancer Detection-A Scoping Review.” *Journal of the American Academy of Dermatology* 87 (1): 157–159.

Haase, Kristen R., Schroder Sattar, Sophie Pilleron, et al. 2023. “A Scoping Review of Ageism towards Older Adults in Cancer Care.” *Journal of Geriatric Oncology* 14 (1): 101385.

Hajian-Tilaki, Karimollah. 2013. “Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation.” *Caspian Journal of Internal Medicine* 4 (2): 627–635.

Hamilton, Clayon B., Alison M. Hoens, Annette M. McKinnon, et al. 2021. “Shortening and

Validation of the Patient Engagement In Research Scale (PEIRS) for Measuring Meaningful Patient and Family Caregiver Engagement." *Health Expectations: An International Journal of Public Participation in Health Care and Health Policy* 24 (3): 863–879.

Hawkins, Douglas M. 2004. "The Problem of Overfitting." *Journal of Chemical Information and Computer Sciences* 44 (1): 1–12.

Health Canada. 2021. "Guidance Document: Software as a Medical Device (SaMD): Definition and Classification." October 5. https://www.canada.ca/en/health-canada/services/drugs-health-products/medical-devices/application-information/guidance-documents/software-medical-device-guidance-document.html.

Health Canada. 2026a. "Guidance Document: Part C, Division 5 of the Food and Drug Regulations 'Drugs for Clinical Trials Involving Human Subjects' (GUI-0100): Labelling and Records - Canada.ca." March 2. https://www.canada.ca/en/health-canada/services/drugs-health-products/compliance-enforcement/good-clinical-practices/guidance-documents/guidance-drugs-clinical-trials-human-subjects-gui-0100/labelling-records.html.

Health Canada. 2026b. "Guidance on Medical Device Establishment Licensing (GUI-0016) (effective December 14, 2026): Other Regulatory Requirements." June 17. https://www.canada.ca/en/health-canada/services/drugs-health-products/compliance-enforcement/establishment-licences/directives-guidance-documents-policies/dec-guidance-medical-device-establishment-licensing-0016/other-regulatory-requirements.html.

Hoffman, Kelly M., Sophie Trawalter, Jordan R. Axt, and M. Norman Oliver. 2016. "Racial Bias in Pain Assessment and Treatment Recommendations, and False Beliefs about Biological Differences between Blacks and Whites." *Proceedings of the National Academy of Sciences of the United States of America* 113 (16): 4296–4301.

International Plain Language Federation. 2019. "Plain Language Definitions - International Plain Language Federation." International Plain Language Federation -, June 28. https://www.iplfederation.org/plain-language-definitions/.

Jung, Yoonsuh, and Jianhua Hu. 2015. "A K-Fold Averaging Cross-Validation Procedure." *Journal of Nonparametric Statistics* 27 (2): 167–179.

Kamiran, Faisal, and Toon Calders. 2009. "Classifying without Discriminating." *2009 2nd International Conference on Computer, Control and Communication*, February, 1–6.

Kamishima, Toshihiro, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2012. "Fairness-Aware Classifier with Prejudice Remover Regularizer." In *Machine Learning and Knowledge Discovery in Databases*. Lecture Notes in Computer Science. Springer Berlin Heidelberg.

Kea, Bory, M. Kennedy Hall, and Ralph Wang. 2019. "Recognising Bias in Studies of Diagnostic Tests Part 2: Interpreting and Verifying the Index Test." *Emergency Medicine Journal: EMJ* 36 (8): 501–505.

Khor, Sara, Eric C. Haupt, Erin E. Hahn, Lindsay Joe L. Lyons, Veena Shankaran, and Aasthaa Bansal. 2023. "Racial and Ethnic Bias in Risk Prediction Models for Colorectal Cancer Recurrence When Race and Ethnicity Are Omitted as Predictors." *JAMA Network Open* 6 (6): e2318495.

Kim, Soo-Yeon, Yunhee Choi, Eun-Kyung Kim, et al. 2021. “Deep Learning-Based Computer-Aided Diagnosis in Screening Breast Ultrasound to Reduce False-Positive Diagnoses.” *Scientific Reports* 11 (1): 395.

Koduri, Sivarama Krishna Akhil. 2026. “Dynamic Memory Updating in RAG: Lifelong Learning and Adaptation.” In *Social Science Research Network*. January 1. https://doi.org/10.2139/ssrn.6000815.

Larrazabal, Agostina J., Nicolás Nieto, Victoria Peterson, Diego H. Milone, and Enzo Ferrante. 2020. “Gender Imbalance in Medical Imaging Datasets Produces Biased Classifiers for Computer-Aided Diagnosis.” *Proceedings of the National Academy of Sciences of the United States of America* 117 (23): 12592–12594.

Lee, Michelle S., Lisa N. Guo, and Vinod E. Nambudiri. 2022. “Towards Gender Equity in Artificial Intelligence and Machine Learning Applications in Dermatology.” *Journal of the American Medical Informatics Association: JAMIA* 29 (2): 400–403.

Maly, Rose C., Yoshiko Umezawa, Barbara Leake, and Rebecca A. Silliman. 2005. “Mental Health Outcomes in Older Women with Breast Cancer: Impact of Perceived Family Support and Adjustment.” *Psycho-Oncology* 14 (7): 535–545.

Mehrabi, Ninareh, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. “A Survey on Bias and Fairness in Machine Learning.” *ACM Computing Surveys* 54 (6): 115:1–115:35.

Mickel, Jennifer. 2023. *The Importance of Multi-Dimensional Intersectionality in Algorithmic Fairness and AI Model Development*. May. https://hdl.handle.net/2152/122644.

National Academies of Sciences, Engineering, and Medicine, Health and Medicine Division, Board on Population Health and Public Health Practice, et al. 2017. “The Root Causes of Health Inequity.” In *Communities in Action: Pathways to Health Equity*. National Academies Press (US).

Nguemeni Tiako, Max Jordan, Eden Engel-Rebitzer, Ari Friedman, et al. 2026. “Racial Disparities in Unmet Pain Treatment Preference, Pain Treatment Satisfaction and Subsequent Opioid Misuse: A Secondary Analysis of a National Multisite RCT.” *Journal of General Internal Medicine* 41 (4): 1041–1047.

Nickerson, Raymond S. 1998. “Confirmation Bias: A Ubiquitous Phenomenon in Many Guises.” *Review of General Psychology: Journal of Division 1, of the American Psychological Association* 2 (2): 175–220.

Nordberg, Samuel S., Andrew A. McAleavey, and Christian Moltu. 2021. “Continuous Quality Improvement in Measure Development: Lessons from Building a Novel Clinical Feedback System.” *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation* 30 (11): 3085–3096.

Park, Yubin, and Joyce C. Ho. 2021. “Tackling Overfitting in Boosting for Noisy Healthcare Data.” *IEEE Transactions on Knowledge and Data Engineering* 33 (7): 2995–3006.

Pasqualetti, Francesco, Alessandro Barberis, Sofia Zanotti, et al. 2023. “The Impact of

Survivorship Bias in Glioblastoma Research." *Critical Reviews in Oncology/hematology* 188 (104065): 104065.

Pilat, Dan, and Sekoul Krastev. 2021. "Confirmation Bias - Biases & Heuristics." The Decision Lab, December 15. https://thedecisionlab.com/biases/confirmation-bias.

Pinto, Manuela Fernández. 2023. *Methodological and Cognitive Biases in Science: Issues for Current Research and Ways to Counteract Them*. https://philpapers.org/rec/PINMAC-2.

Pleiss, Geoff, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q. Weinberger. 2017. "On Fairness and Calibration." In *arXiv [cs.LG]*. September 6. arXiv. http://arxiv.org/abs/1709.02012.

Popovic, Aleksandar, and Martin R. Huecker. 2025. "Study Bias." In *StatPearls*. StatPearls Publishing.

Protière, Christel, Patrice Viens, Frédérique Rousseau, and Jean Paul Moatti. 2010. "Prescribers' Attitudes toward Elderly Breast Cancer Patients. Discrimination or Empathy?" *Critical Reviews in Oncology/hematology* 75 (2): 138–150.

Rajkomar, Alvin, Michaela Hardt, Michael D. Howell, Greg Corrado, and Marshall H. Chin. 2018. "Ensuring Fairness in Machine Learning to Advance Health Equity." *Annals of Internal Medicine* 169 (12): 866–872.

Reeder-Hayes, Katherine E., Bradford E. Jackson, Tzy-Mey Kuo, et al. 2024. "Structural Racism and Treatment Delay among Black and White Patients with Breast Cancer." *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*, August 6, JCO2302483.

Richardson, Eve, Raphael Trevizani, Jason A. Greenbaum, Hannah Carter, Morten Nielsen, and Bjoern Peters. 2024. "The Receiver Operating Characteristic Curve Accurately Assesses Imbalanced Datasets." *Patterns (New York, N.Y.)* 5 (6): 100994.

Ring, A., H. Harder, C. Langridge, R. S. Ballinger, and L. J. Fallowfield. 2013. "Adjuvant Chemotherapy in Elderly Women with Breast Cancer (AChEW): An Observational Study Identifying MDT Perceptions and Barriers to Decision Making." *Annals of Oncology* 24 (5): 1211–1219.

Ring, Alistair. 2010. "The Influences of Age and Co-Morbidities on Treatment Decisions for Patients with HER2-Positive Early Breast Cancer." *Critical Reviews in Oncology/hematology* 76 (2): 127–132.

Roberts-Licklider, Karen, and Theodore Trafalis. 2024. "Machine Learning Techniques with Fairness for Prediction of Completion of Drug and Alcohol Rehabilitation." In *arXiv [cs.LG]*. April 23. arXiv. http://arxiv.org/abs/2404.15418.

Robinig, Wolfgang, and Johannes P. Wallner. 2026. "Bias in AI: A Framework for Understanding the Sources of Bias in Artificial Intelligence." *AI and Ethics* 6 (4): 438.

Ruopp, Marcus D., Neil J. Perkins, Brian W. Whitcomb, and Enrique F. Schisterman. 2008. "Youden Index and Optimal Cut-Point Estimated from Observations Affected by a Lower

Limit of Detection." *Biometrical Journal* 50 (3): 419–430.

Sabin, Janice A. 2022. "Tackling Implicit Bias in Health Care." *The New England Journal of Medicine* 387 (2): 105–107.

"Selection Bias." 2011. February 2. https://www.cancer.gov/publications/dictionaries/cancer-terms/def/selection-bias.

Seyyed-Kalantari, Laleh, Haoran Zhang, Matthew B. A. McDermott, Irene Y. Chen, and Marzyeh Ghassemi. 2021. "Underdiagnosis Bias of Artificial Intelligence Algorithms Applied to Chest Radiographs in under-Served Patient Populations." *Nature Medicine* 27 (12): 2176–2182.

Sheikh, Fatima, Alison E. Fox-Robichaud, and Lisa Schwartz. 2023. "Collecting Race-Based Data in Health Research: A Critical Analysis of the Ongoing Challenges and next Steps for Canada." *Canadian Journal of Bioethics* 6 (1): 75–80.

Sinha, Mukesh K., and Victor M. Montori. 2006. "Reporting Bias and Other Biases Affecting Systematic Reviews and Meta-Analyses: A Methodological Commentary." *Expert Review of Pharmacoeconomics & Outcomes Research* 6 (5): 603–611.

Taha, Abdel Aziz, and Allan Hanbury. 2015. "Metrics for Evaluating 3D Medical Image Segmentation: Analysis, Selection, and Tool." *BMC Medical Imaging* 15 (1): 29.

Turner, Melanie, Shona Fielding, Yuhan Ong, et al. 2017. "A Cancer Geography Paradox? Poorer Cancer Outcomes with Longer Travelling Times to Healthcare Facilities despite Prompter Diagnosis and Treatment: A Data-Linkage Study." *British Journal of Cancer* 117 (3): 439–449.

Usher-Smith, Juliet A., Stephen J. Sharp, and Simon J. Griffin. 2016. "The Spectrum Effect in Tests for Risk Prediction, Screening, and Diagnosis." *BMJ (Clinical Research Ed.)* 353 (June): i3139.

Varkey, Basil. 2021. "Principles of Clinical Ethics and Their Application to Practice." *Medical Principles and Practice: International Journal of the Kuwait University, Health Science Centre* 30 (1): 17–28.

"View of SMOTE: Synthetic Minority Over-Sampling Technique." n.d. Accessed September 25, 2024. https://www.jair.org/index.php/jair/article/view/10302/24590.

Webster, Craig S., Saana Taylor, Courtney Thomas, and Jennifer M. Weller. 2022. "Social Bias, Discrimination and Inequity in Healthcare: Mechanisms, Implications and Recommendations." *BJA Education* 22 (4): 131–137.

Wilkinson, Mark D., Michel Dumontier, Ijsbrand Jan Aalbersberg, et al. 2016. "The FAIR Guiding Principles for Scientific Data Management and Stewardship." *Scientific Data* 3 (1): 160018.

Yang, Jenny, Andrew A. S. Soltan, David W. Eyre, Yang Yang, and David A. Clifton. 2023. "An Adversarial Training Framework for Mitigating Algorithmic Biases in Clinical Machine Learning." *Npj Digital Medicine* 6 (1): 55.