An End-to-End Medical Algorithmic Audit of TriageAssist

PREPARED FOR PRIORITIZE™

Dr. AJung Moon, PhD (ajung.moon@mcgill.ca) **Chief Executive Officer**

PREPARED BY VANTAMED+

Miguel Carrillo-Cobián (miguel.carrillocobian@mail.mcgill.ca), **Responsible AI Auditor**
Leah Davis (leah.davis@mail.mcgill.ca) **Responsible AI Auditor**

Address: 845 Rue Sherbrooke Ouest, Montréal, QC H3A 0G4

Executive Summary

This report conducts an algorithmic audit of TriageAssist, a clinical decision support system developed by Prioritize™ for cardiovascular triaging in emergency room settings. Chosen for its contextual relevance and credibility, the audit follows Liu et al.'s (2022) end-to-end medical algorithmic audit framework, analyzing six stages: scoping, mapping, artifact collection, testing, synthesis of key audit matters, and post-audit measures. The primary objective is to assess how TriageAssist performs across various patient demographic subgroups, particularly biological sex and age, given documented disparities in cardiovascular symptom presentation and treatment outcomes. Following this investigation, actionable recommendations for several parties are identified to guide product improvements before broader deployment. The key findings of this audit can be summarized in four points, listed in priority order from highest to lowest:

1. **Data Imbalances:** Female patients, particularly at either age extreme, are significantly underrepresented in the training data, reducing system reliability for these populations.
2. **Usage Workflow Challenges:** Over-reliance on the tool's outputs (e.g., pre-screening use) risks automation bias, undermining human judgment, and increasing clinical liability.
3. **Model Biases:** Predictive performance skews toward middle-aged male patients, with younger and older females at higher misclassification risks.
4. **Lack of Transparency:** The system lacks interpretability features (e.g., confidence scores or usage warnings), limiting appropriate clinical integration.

Based on these findings, several key audit matters emerged for three parties based on the corresponding roles, scope, and capabilities that each requires. Company-wide risk mitigation measures are considered to be the most influential in creating change across Prioritize™. All measures and actions are listed in order of highest to lowest priority:

Company-Wide Risk Mitigation Measures

- Proactively implement corrections before further downstream usage of the system.
- Construct accessible and continual feedback mechanisms for all stakeholders.
- Provide training resources for usage, and more generally, responsible AI practices.
- Align future partnerships and funding initiatives with equity-first institutions.

Developer Actions

- Address data imbalances through specialized collection and preprocessing techniques.
- Conduct comprehensive fairness testing protocols and explainability reviews.
- Incorporate confidence measures, usage limitations, and warnings in the system's outputs.
- Consider integrating support vector machine models given their demonstrated suitability.

Clinical Actions

- Use TriageAssist as a cross-validation tool only after initial clinician assessments.
- Provide additional care and attention in triaging younger, female patients.
- Once implemented, consider the system's confidence scores and warnings in triaging.

Table of Contents

Executive Summary **2**
Table of Contents **3**
Introduction to Vantamed+ **5**
Positionality Statement of Auditors 5
Audit Background **5**
Technological Contextualization of TriageAssist 6
Problem Statement and Audit Objectives 6
Existing Inquiry and Related Work 7
Clinical Algorithm Use and Patient Subgroup Analyses within Cardiovascular Disease 7
Patient Demographic Subgroup Focus: Biological Sex and Age 8
Audit Methodology **9**
General Approach and Rationale 9
Audit Findings **10**
Stage 1: Scoping 10
Intended Audit Impact and Perceived Benefits 10
Intended Audit Usage 12
Stage 2: Mapping 12
Stakeholder Identification 12
Mapping Healthcare Tasks 13
Artificial Intelligence Components and Boundaries of TriageAssist 14
Use Case Scenarios 14
Takeaways: Positioning the Human-in-the-Loop Process 16
Stage 3: Artifact Collection 16
Audit Checklist 16
Dataset Artifact: Heart Failure Prediction Dataset 17
Model Artifact: Binary Classifier Using Traditional Machine Learning Approaches 17
Real-Time Deployment Artifacts 18
Documentation Artifacts: Stakeholder Interviews and Contextual Materials 18
Scientific Literature Artifacts 19
Stage 4: Testing 19
Descriptive Data Analysis 19
Concern 01: Underrepresentation of Female Population 20
Concern 02: Heart Disease Prevalence by Sex 20
Takeaways: General Directionality for Further Testing 21
Exploratory Error Analysis 22
Shapley Additive Explanations (SHAP) Analysis 23
Takeaways: Implications of Exploratory Testing 25
Patient Subgroup Testing 25

Takeaways: Implications of Patient Subgroup Testing 27
Adversarial Testing 28
Takeaways: Implications of Adversarial Testing 29
Stage 5: Synthesis of Key Audit Matters 29
Developer Actions 30
Clinical Actions 33
Company-Wide Management Risk Mitigation Measures 34
Limitations of Audit **36**
Stage 6: Post-Audit Measures 37
Re-Auditing TriageAssist 37
Concluding Statement **37**
References **41**
Appendices **43**
Appendix A: Key Stakeholder Value Mapping 43
Appendix B: Code Repository Access Information 45
Appendix C: Baseline Heart Disease Patient Profile 46

Introduction to Vantamed+

Vantamed+ is a Montreal-based startup specializing in the ethical algorithmic evaluation of digital health technologies. Initially providing web design and marketing solutions tailored to the needs of medical professionals, the company has evolved to build on its deep and multidisciplinary understanding of clinical communication and user-centered design. Vantamed+ is committed to supporting the responsible integration of artificial intelligence (AI) in healthcare settings by ensuring systems are fair, transparent, and best aligned with safe patient care and clinical practice. Guided by the principles of equity, accountability, and collaboration, the company works closely with developers, clinicians, company executives, and other stakeholders to assess the safety, effectiveness, and real-world impact of algorithmic tool usage. This audit reflects Vantamed+'s mission to promote the development of trustworthy medical technologies through rigorous, justice-informed evaluation.

Positionality Statement of Auditors

Leah Davis is conducting this audit through her lens as a female, Canadian, Caucasian, biomedical engineer-in-training. Having several previous research partnerships with Canadian hospital departments, she has significant experience implementing digital tool usage in clinical settings. With several family members working in the nursing profession, she acknowledges her relative biases regarding these tools within the Canadian healthcare system. She remains optimistic about the importance and relevance of incorporating new technologies to empower healthcare diagnoses and treatments, especially for complex medical cases that can leverage the benefits of powerful computational capabilities.

Miguel Carrillo Cobián conducts this audit from the perspective of a male, Caucasian, Latin American engineer. He has experience in both the use and development of machine learning classification systems. With both of his parents practicing medicine in Mexico, he is familiar with the complexity of decision-making in medical settings, especially within the high-stakes environment of emergency rooms. This background brings a strong personal and familial bias when engaging with health-related decisions. Nevertheless, Miguel recognizes the critical importance of both efficiency and accuracy in diagnostic processes.

Audit Background

This report contains a systems-level, third-party, end-to-end medical algorithmic audit for Prioritize™, a Montreal-based health technology start-up. Recently creating an algorithmic clinical decision support system named TriageAssist, Prioritize™ has commissioned this audit within the late development stages and local pilot deployment of this product, before its highly scaled expansion in more hospitals.

Technological Contextualization of TriageAssist

TriageAssist aims to enhance medical professionals' abilities in effectively triaging cardiac-related conditions in emergency room settings. The system receives inputs from clinicians regarding each patient in the triaging phase of care, upon which a binary classification algorithm predicts whether the individual has heart disease. It functions by processing a subset of patient data—primarily vital signs, selected symptoms, some medical history provided, and limited demographic information—returning a binary classification: whether a patient is likely to have heart disease. *Based on the documentation provided, it is assumed to be a classification system rather than a ranking algorithm or large-language model.*

Due to the system's involvement in triaging emergency cardiac use cases, TriageAssist continuously manages high-risk, high-stress, and high-stakes patient scenarios, often life or death, but ultimately, the system aims to enhance patient prioritization and support logistical clinical decisions within highly resource-constrained hospital environments. Given that all human populations can use emergency room services in Canada, this use case is embedded among a complex ecosystem of technical, social, and legal dimensions.

Problem Statement and Audit Objectives

Of all potential areas of interest noted in the provided technological brief and stakeholder interviews, one apparent, and rather significant, triage concern was raised: **the difference in cardiac distress and heart failure symptom onset among different patient subgroups**. The National Center for Advancing Translational Sciences defines a patient subgroup as "a subset of the study population or study sample defined by specific baseline characteristics" (n.d.). Patient subgroups can be divided into demographic, behavioural, or molecular subcategories; however, the patient attributes of concern in this audit are *demographic*. Age and ethnicity were both raised in discussion, however, biological sex (e.g., female or male) was the most predominant concern within cardiac clinical assessments. Ultimately, each patient subgroup has nuanced variability, preventing a one-size-fits-all approach in the algorithmic modelling of this condition. A failure to address differences among subgroups may lead to underdiagnoses, delayed treatments, and ultimately poorer health outcomes, or even death, if not considered carefully. **Based on the information provided, Prioritize™ has yet to consider sex-specific differences or other demographic subgroup intersectionalities in TriageAssist's development, evaluation, or deployment.**

As a result, this audit takes a proactive and protective stance in identifying and mitigating the potential risks of TriageAssist, particularly in assessing and comparing the performance of patient demographic subgroups, particularly within sex-specific discrepancies.

Takeaways: Audit Objectives

1. Evaluate how TriageAssist currently performs across various patient subgroups, particularly demographic features including biological sex and age, within the system's development, evaluation, and deployment workflows.

2. Conclude and deliver an initial, actionable set of evidence-based recommendations at various priority levels to address identified demographic disparities, for a clear path forward in TriageAssist's continued expansion.

Note that biological sex, a physical construct, is different from gender, a social construct. The usage of TriageAssist, including all dataset training, does not consider gender as a demographic variable, but rather biological sex. It is important to note that the system does not have the granularity to denote cardiac characteristics of intersex individuals.

Existing Inquiry and Related Work

To ensure that demographic subgroup analyses warrant an important course of action and scope for Prioritize™, the following section investigates the existing background of this problem, specifically within other clinical algorithms that have attempted to manage it and further justify its importance as a topic of focus.

Clinical Algorithm Use and Patient Subgroup Analyses within Cardiovascular Disease

As the technological description provides, cardiovascular diseases (CVD) are significant causes of concern among global health challenges. Canada, in particular, has observed a steady increase in CVD disability and mortality across all populations since 2010 (Campbell et al., 2021). Noting the improvement needed in effectively improving CVD health outcomes, many companies like Prioritize™ have aimed to develop a clinical decision support system to aid early detection triaging processes. Often, these support systems implement specialized machine learning (ML) algorithms to improve i) a variety of CVD diagnoses, ii) adverse outcome prevention, and iii) personalized treatments for CVD.

However, significant biases across demographic groups have been found in the datasets used to train these support systems, in addition to the modelling and evaluation strategies used to assess their efficacy (Giordano et al., 2021). Risk estimate tools such as STRIDE-HF, an extensive heart failure triage system piloted for emergency room usage, fail to assess or tailor their tool's usage, performance, or outcomes across any demographic characteristics (Sax et al., 2024). Suri et al. (2022) found that across 24 ML-based CVD risk assessment tools, there is a significant exclusion and prioritization of sourcing and applying diverse data to risk modelling, compounded by a lack of research and industry funding to do so and, more generally, a failure of data sharing.

The above examples' findings are concerning, but common across clinical decision support systems. Of 37 studies analyzed in a systematic literature review of multi-purpose clinical decision support system usage, the vast majority indicate high or unclear risks of bias, displaying no consideration for human factors (Vasey et al., 2021). Over five years ago, Paulus and Kent (2020) emphasized how unfair these algorithmic systems are, especially within their evaluation processes. The authors discuss how risk cannot be objectively measured in an individual, predisposing them to be estimated within a larger collective group judged to have "similar enough" characteristics to their own– often resulting in poor performance and health outcomes. These issues remain prevalent in these systems even today, even five years later.

Patient Demographic Subgroup Focus: Biological Sex and Age

Aside from these four examples, which show concerns across all patient subgroups, the repeated and systemic failures of algorithmic clinical support systems have been found more drastically within sex-specific CVD diagnosis and treatment. This is rooted in the historical marginalization and neglect of female healthcare needs, alongside the proliferation of imbalanced societal power dynamics (Arslanian-Engoren, 2002). Male baselines and thresholds, assessment strategies, and triage decision norms have *always* been prioritized in cardiac evaluation systems, resulting in females having fewer data points (in both quality and quantity), clinical trials, and available reporting structures in CVD (Arslanian-Engoren, 2002; Mulvagh et al., 2024). As clinical decision support systems rely on past data, the results have become increasingly ungeneralizable to female patients, showing consistently adverse effects in the consideration of biological sex in the design, analysis, and reporting of clinical tools (Mulvagh et al., 2024). For instance, the American Heart Association developed Pooled Cohort Risk Equations (PCEs) and ML models based on data from the Vanderbilt University Medical Center to predict CVD risks. Despite analyzing different subgroup populations, extreme biases were reported in all risk-based predictions, particularly in providing lower expected rates of CVD occurrence and overall poorer diagnostic accuracy across females (Li et al., 2023). Nolin-Lapalme et al. (2024) continue that sex-specific biases across clinical settings are only becoming more amplified as they build on a series of incorrect correlations among features. These biases become even more apparent when considering the intersectionality of additional demographic attributes.

This perpetuation of bias in clinical systems is problematic for many reasons, especially among female CVD, which is the leading cause of premature female deaths and hospitalization globally, following only birth (Jaffer et al., 2021). Regionally, the province of Quebec has the second-highest annual count of female CVD mortality rates. This is hypothesized for several reasons, as discussed by Jaffer et al. (2021). The first is that female risk factors of CVD, such as diabetes, depression, or hypertension, lead to greater morbidity impacts compared to males. The second is that females have higher chances of heart failure, stroke, or death following a heart attack episode. The third, and most important for companies like Prioritize™, is that females have an extremely atypical presentation of CVD, which has been historically overlooked for decades, affecting the subsequent historical health data collected and trended. As Jaffer et al. (2021) continue, despite females ≥ 52 years of age contributing to most CVD emergency room visits, the majority of

congenital heart diseases occur in young female populations. Based on recent health data trends, CVD outcomes for *all* females have stagnated, particularly at either of these age extremes– the burden of CVD has begun to shift. Ultimately, all of these research studies emphasize how companies aiming to improve CVD must consider the drastic impact of demographic patient characteristics in their systems, especially across biological sex and age, as early and effective triaging amongst females with CVD experience greater rehabilitative benefits and outcomes compared to males (Jaffer et al., 2021).

Takeaways: Framing the Problem Statement

Without a comprehensive assessment of the sociotechnical risks stemming from a demographic lens, TriageAssist risks perpetuating the biases and inconsistencies that fail to consider subgroup differences adequately. Therefore, it is of utmost priority to call attention to these factors within this audit. Conducting an audit aimed at recognizing and evaluating demographic subgroup differences will help uncover existing inequalities, ultimately improving diagnostic accuracy and enhancing the system's ability to equitably serve other diverse populations. Establishing a clearer understanding of how these systems perform across different demographic groups is essential to ensure that the AI-driven triage processes are inclusive, fair, and effective in promoting better health outcomes for *all* patients.

Audit Methodology

General Approach and Rationale

The prior section confirmed the importance of performing an audit that assesses and validates critical demographic differences in the usage, application, and safety of clinical support systems. Hence, this audit must apply various strategies to assess how TriageAssist performs within this scope to conclude its safe operation boundaries. Our team chose to employ Liu et al.'s (2022) "The Medical Algorithmic Audit", specifically designed for applications with clinical contexts. The framework's relevancy lies in its direct emphasis on demographic patient subgroup testing and analyses, due to its coverage in both of the audit objectives and within the documentation provided for this particular use case.

Recently published in an incredibly acclaimed peer-reviewed academic journal, *The Lancet*, the audit builds upon an extremely strong algorithmic audit foundation created by Raji et al.'s (2020) team, which has been cited nearly 1000 times. The audit also holds a heavy industry focus, likely due to the authors' commercial affiliations, such as the second author Ben Glocker, who works at a cardiac health company named *Heartflow*.

The framework's greatest strength is its capacity to systematically guide evaluation across the entire machine learning pipeline– from the datasets selected, the modelling employed, and the deployment strategies promoted. This end-to-end nature enables a much more holistic and

systemic approach, capturing critical interdependencies and reliability throughout the entire workflow. Figure 1 summarizes the six stages of the audit structure: scoping, mapping, artifact collection, testing, synthesis of key audit matters, and post-audit measures.

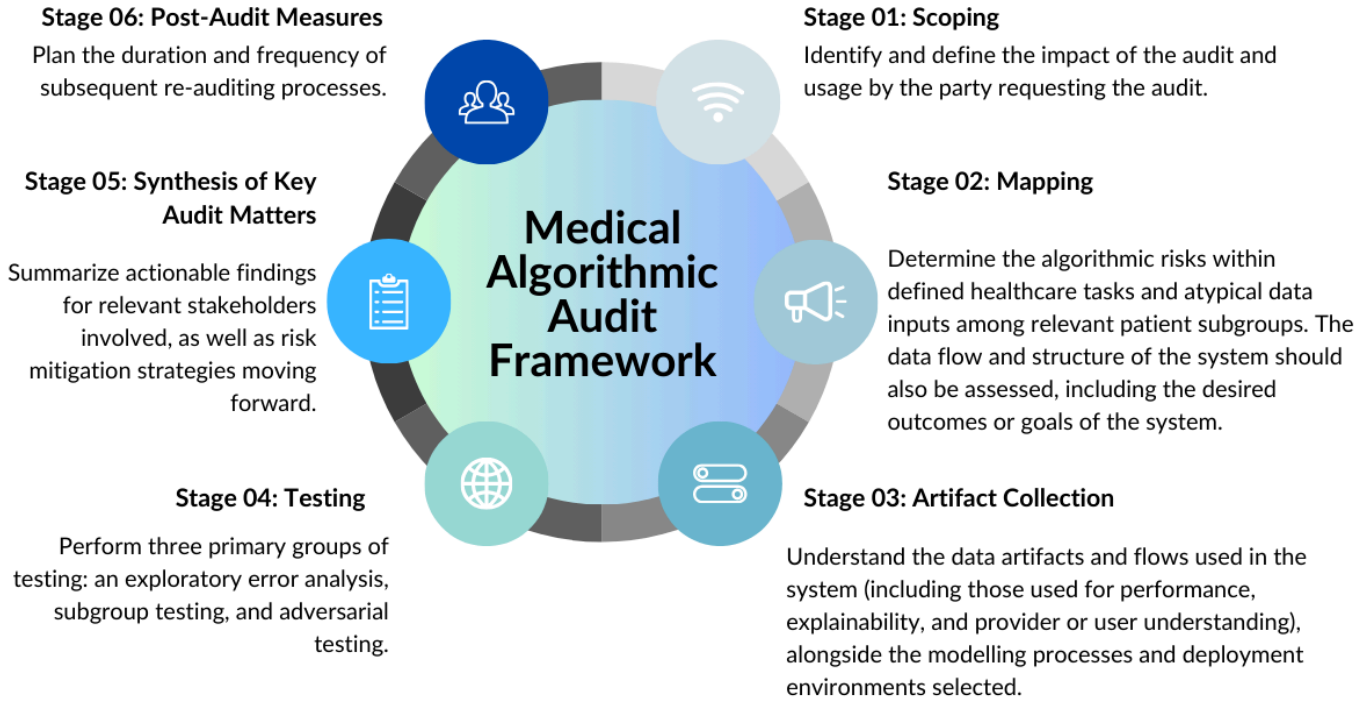

Figure 1: Description of Liu et al.'s (2022) six-stage medical algorithmic auditing process, selected as the framework to audit TriageAssist.

Audit Findings

Stage 1: Scoping

Intended Audit Impact and Perceived Benefits

Ultimately, improving demographic patient disparities within TriageAssist provides five general benefits to Prioritize™. First, it ensures that the company commits to its ethical, corporate, and legal duties to safeguard all patients. Second, it presents evidence-based inquiries enabling large-scale expansion opportunities among hospitals, and opportunities for equity-related funding and strategic partnerships. Third, it appropriately allows the company to take and better manage a protective and proactive stance on a range of reputational and social risks and responsibilities. Fourth, it lends to a better understanding of product usage, deployment, and implementation practices. Finally, it prepares the company for current and downstream market and regulatory considerations in the medical device industry. *Note that in communication with Prioritize™'s Chief Executive Officer (CEO), the first three areas are the primary priorities within their company's aims.* All five of these areas are summarized in Table 1.

Table 1: Five general areas and ten specific points of interest in the importance of analyzing patient demographic subgroups in TriageAssist. In discussion with company executives, the primary priorities for the TriageAssist product have been labelled with an asterisk (*).

General Areas	Related Impacts	Perceived Company Benefits
***Ethical, Corporate, and Legal Duties to Safeguard *All* Patients Using TriageAssist**	1. Ensure the well-being of all patient demographics.	Fulfill ethical and legal obligations while enhancing patient trust and confidence.
	2. Eliminate liability risks for the company and associated clinicians.	Mitigate litigation risks by ensuring comprehensive patient coverage.
***Large-Scale Expansion Among Hospitals, Funding Agencies, and Strategic Partnership Management**	3. Meet the needs of patient subgroups, who comprise a significant portion of the user base, such as females.	Grow the company's size, scale, market positioning, and credibility.
	4. Facilitate effective collaboration and implementation across hospital networks.	Grow the company's size, scale, market positioning, and credibility.
	5. Increase the eligibility for equity-related funding, awards, and grants.	Earn funding, resources, and opportunities to hire more personnel and expand the company's infrastructure.
***Reputational and Social Risks and Responsibilities**	6. Address current market mistrust associated with clinical decision support systems.	Reinforce a positive brand image.
	7. Prevent negative publicity and reputational damage.	Strengthen trust among the public, clinicians, and other relevant stakeholders.
Product Usage, Deployment, and Implementation Practices	8. Identify appropriate product usage patterns.	Improve patient and clinician satisfaction and product efficiency.
	9. Deliver effective instructional materials and training support.	Reduce product misuse and increase clinician proficiency.
Healthcare Market and Regulatory Considerations	10. Actively preparing for market expansion that requires more rigorous regulatory compliance.	Centre the company in an agile position to readily expand into other markets and sectors.

Takeaways: Audit Incentives

Demographic subgroup evaluation provides five incentive areas affecting current development practices, future deployment strategies, and downstream market effects (Table 1). These practices virtually affect all teams at Prioritize™, including software developers, system integrators, marketing agents, human resources and partnership leads, regulatory and quality analysts, legal advisors, project management, and executive members.

Intended Audit Usage

This audit serves as an initial assessment of TriageAssist before its widespread expansion. As requested in the audit briefing, this audit is targeted toward management-level executives in Prioritize™, particularly the CEO, to provide a preliminary evaluation of the most critical product improvements needed. Additional insights may be provided for specific stakeholders to operationalize any course corrections found. We anticipate the company executives using the proposed key audit matters to establish or approve implementation strategies such as assigning task forces, commissioning follow-up evaluations, or integrating changes into strategic roadmaps. The audit specifically looks to inform choices about safety, resource allocation, and/or product, market, or regulatory direction. Furthermore, its results can help in the following capacities: help executive members communicate material issues to different actors, shape internal documentation, update the board of directors on the product's initial pilot deployment findings, and provide avenues for public disclosure.

Stage 2: Mapping

Stakeholder Identification

The stakeholder identification process serves as a critical foundation for assessing any inputs or outputs of the TriageAssist system. It maps out all key actors who are influenced or impacted by the system and positions their values with the potential areas of concern. A stakeholder-first approach ensures that ethical, clinical, and operational dimensions are properly contextualized and that audit findings can be translated into meaningful, actionable recommendations. A detailed summary of all relevant stakeholders, their values, and their alignment among each of the five audit benefits can be found in Appendix A, Table A1. A brief summary follows:

- The provider, **Prioritize™**, is the developer and deployer of the system. Their values, such as market expansion, regulatory compliance, workflow optimization, and public trust, are explored in depth in the previous section, *"Intended Audit Impact and Perceived Benefits"*. The audit's scope is most closely aligned with these priorities.

- The primary clients, **hospitals**, value equitable health outcomes, reduced wait times, and effective support for clinical staff. They seek smooth system integration, patient safety, and reliable training and usage of the system.
- The direct users of the systems, **nurses and doctors**, value accuracy, interpretability, and clarity in system outputs. Their focus is on usability, patient safety, and fairness in patient care delivery and well-being.
- The end beneficiaries, **patients,** value timely, human-centered care and transparent and trustworthy communication.
- The oversight authorities, **regulators,** prioritize public health of *all* patient populations, data privacy, and fairness, with an emphasis on safety and compliance.

Takeaways: Stakeholder Mapping

Mapping each provider, client, user, end beneficiary, and oversight authority to their values, various system-level risks and design trade-offs are present. Ethical and operationally grounded audit recommendations must be interpreted through the lens of those directly affected.

Mapping Healthcare Tasks

While nurses perform a wide array of duties in emergency departments, this audit focuses specifically on their role in assessing the risk level of incoming patients during the triage process. Based on the documentation provided on product usage, nurses interact with the system first, before doctors or other stakeholders. This preliminary step is critical, as it directly informs the urgency and severity of the medical attention a patient receives and subsequently shapes the clinical workflow that follows. The typical sequence of clinical tasks within the current triage process (exclusive of the TriageAssist system) includes four steps (Figure 2).

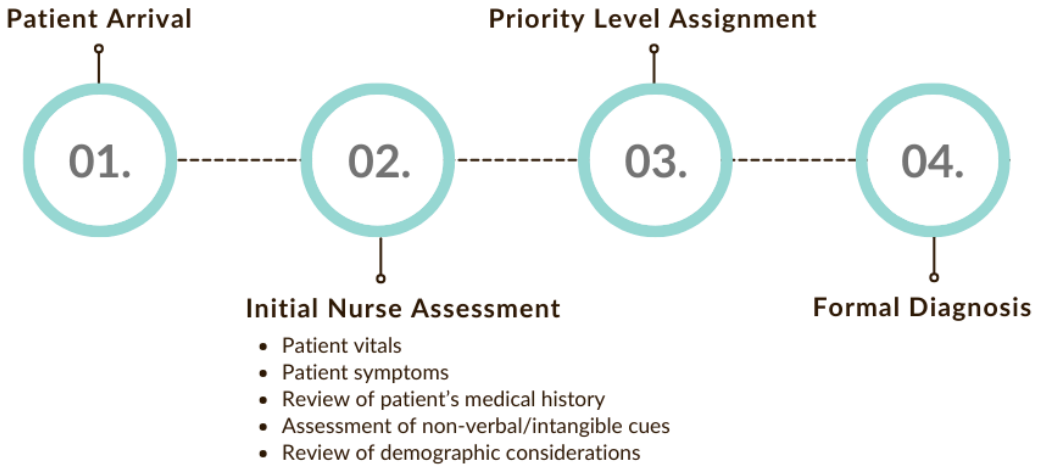

Figure 2: Emergency room heart failure diagnosis workflow without using the TriageAssist system.

Within the following sequences, the audit maps various intervention points where TriageAssist may be introduced, analyzing how it can support or influence clinical decision-making and product workflow patterns.

Artificial Intelligence Components and Boundaries of TriageAssist

As stated previously, TriageAssist is a clinical decision support system designed to assist in emergency cardiac triage. It functions by processing a subset of patient data– primarily vital signs, selected symptoms, some medical history provided, and limited demographic information– and returns a binary classification: whether a patient is likely to have heart disease. *Note that the system often only receives partial information about a patient's symptoms or medical history, as a nurse must enter any details into it; it cannot assess non-verbal or intangible cues or gestures.* Through clinician interviews, two primary use case scenarios emerged, reflecting on how nurses and doctors envision TriageAssist being deployed in practice. The system is designed to be flexible in its application and can be used either *before or after* the nurse's initial patient assessment.

Use Case Scenarios

Use Case 01: Nurse-Initiated Assessment with TriageAssist Cross-Validation

In the first scenario, nurses perform their initial triage assessments, per usual, and subsequently consult TriageAssist as a secondary check, introducing an extra step in the triaging workflow (Figure 3). The algorithm serves as a checkpoint, providing a comparative prediction that the nurse can use to either reinforce or re-evaluate their clinical judgment. This approach maintains the primacy of human evaluation while introducing the AI system as a safety layer.

Such an implementation may improve diagnostic caution and reduce overconfidence, especially in edge cases where symptom presentation is ambiguous. However, it could also extend triage times, particularly in high-volume settings, and may add complexity to workflows without offering significant time savings.

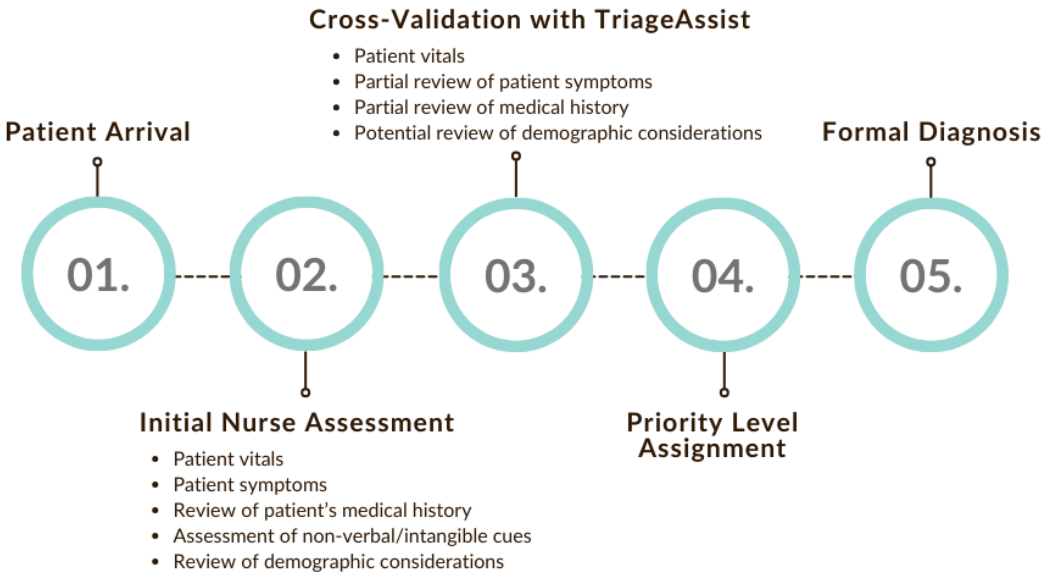

Figure 3: Emergency room heart failure diagnosis workflow using the TriageAssist system to cross-validate triaging decisions.

Use Case 02: TriageAssist Pre-Screening Followed by Nurse Assessment

In the second use case, TriageAssist is employed as a preliminary screening tool. Upon patient arrival, the system immediately evaluates structured input data about the patient and flags individuals deemed high-risk (Figure 4). Nurses then focus their attention on reviewing these flagged cases more thoroughly, using their own clinical expertise to assign a final priority level and begin diagnosis and further planning.

This model may accelerate overall triage efficiency and help staff quickly identify critical patients, particularly in overcrowded and resource-constrained environments. However, there are risks associated with relying on a binary output early in the assessment workflow. Patients not flagged by the system may receive less immediate attention, and nurses may develop an over-reliance on the algorithm if its guidance is not clearly framed as supportive, but rather definitive. This is particularly concerning as it fails to initially consider non-verbal and intangible cues.

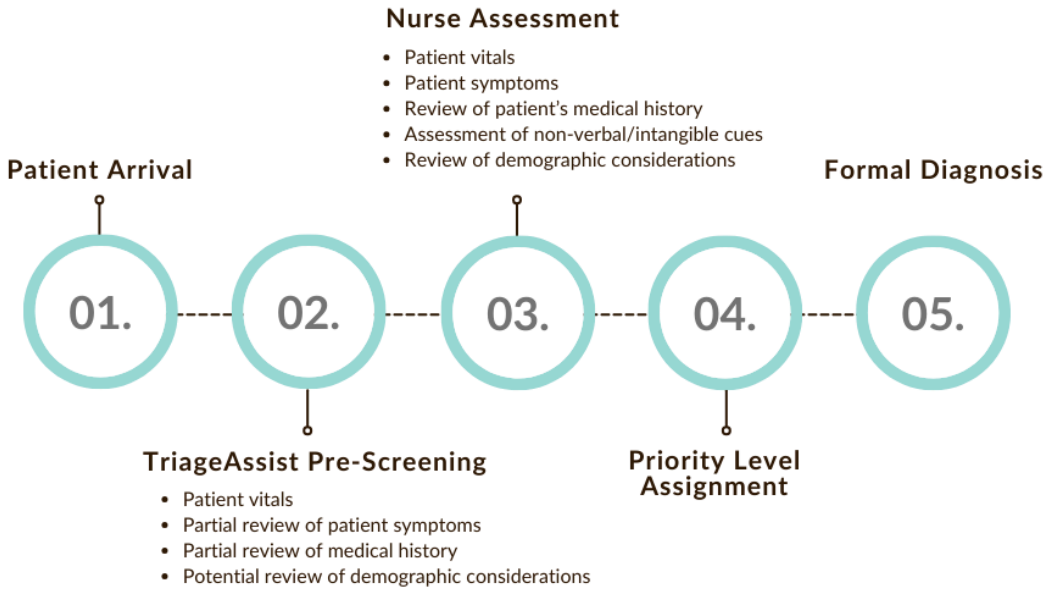

Figure 4: Emergency room heart failure diagnosis workflow using the TriageAssist system as a pre-screening tool.

Takeaways: Positioning the Human-in-the-Loop Process

In both scenarios, the nurse remains the primary decision-maker. TriageAssist is designed to augment, not replace, clinical judgment by offering structured algorithmic insights. A human-in-the-loop approach is essential, particularly given the variability in symptom presentation among demographic subgroups, such as females at either end of the age spectrum, where clinical nuance may fall outside the system's most effective scope.

The choice of implementation, either as a cross-validation or pre-screening tool, should reflect the specific value trade-off between speed and assessment robustness. Each use case offers aspects of automation bias. Faster workflows may support efficiency in high-pressure environments but risk missing subtle presentations. More deliberate, nurse-led assessments with AI support may be slower but offer greater diagnostic reliability. These tensions must be weighed based on clinical priorities and acceptable risk tolerance.

Stage 3: Artifact Collection

Audit Checklist

Before deciding when, where, and how TriageAssist should be used, it is essential to identify and catalog the artifacts currently available for evaluation. This checklist further defines the scope and boundaries of the audit, ensuring clarity on what can be meaningfully analyzed, what gaps remain, and where assumptions may influence human and algorithmic interpretation. A complete

inventory of these artifacts is especially important in early-stage audits where systems may be under development or only partially documented. Six artifacts were identified and assessed as part of this audit.

Dataset Artifact: Heart Failure Prediction Dataset

The dataset that TriageAssist uses is the "H*eart Failure Prediction Dataset*", derived from author Federico Soriano in 2021. It consists of 918 patient records, drawn from a combination of five well-known heart disease datasets from Hungary, Switzerland, the United States, and Germany. Originally totalling 1,190 entries, 272 duplicate records were removed to produce a cleaned, moderately-sized dataset suitable for binary classification tasks.

It includes 11 clinical input features relevant to heart disease risk prediction including age, biological sex, chest pain type (e.g., typical angina), resting blood pressure, cholesterol, fasting blood sugar, resting echocardiogram (the first measurable echocardiogram risk criterion e.g., normal, heart (ST-T) wave abnormality), max heart rate, exercise-induced angina, old peak (the second measurable echocardiogram risk criterion), and ST slope (the third measurable echocardiogram risk criterion). These attributes are often cited in cardiovascular risk stratification literature and are used as input variables in many traditional machine learning models. The dataset also contains the binary output variable of heart disease, indicating a 0 for normal risk and a 1 for a high risk.

Model Artifact: Binary Classifier Using Traditional Machine Learning Approaches

At the time of this audit, no specific model architecture, training configuration, or internal documentation for TriageAssist was made available. As a result, the technical analysis remains agnostic to the exact modelling choices made by the development team. However, Prioritize™ provided several sources indicating the models likely to be used, including sources authored by Rawat (2019), which discuss heart disease prediction methods.

Within the field of clinical risk prediction, particularly for heart disease, certain machine learning models are routinely employed due to their effectiveness in handling structured clinical data. These models are commonly utilized in open-source implementations and academic research that draw on similar datasets. As shown in the references provided, common algorithms in this space include logistic regression, random forest, k-nearest neighbours (KNN), Naive Bayes, and support vector machines (SVM). Each of these models varies in complexity and interpretability. For example, logistic regression is often valued for its transparency and explainability in clinical settings, while ensemble methods such as the random forest model offer greater predictive power but reduced interpretability. Any of these classification techniques can be applied as the modelling artifact for TriageAssist.

Real-Time Deployment Artifacts

At the time of this audit, no deployment data is available as TriageAssist remains in the early development and pilot testing phases, with no finalized or stable deployment environment. As such, this audit does not include performance evaluations under real-world hospital conditions.

Documentation Artifacts: Stakeholder Interviews and Contextual Materials

To better understand TriageAssist's real-world context and ethical implications, interviews were conducted with four key stakeholders: two emergency room nurses, one doctor, and one patient. Their insights offer practical guidance on how the tool is expected to fit into clinical workflows, as stated in the two potential use cases above (Figures 3 and 4), and highlight values that should be used to guide the system's deployment.

Given their similar roles and perspectives, nurses and doctors are grouped into a single stakeholder category, as they are both targeted end users of the technology and share many of the same values and concerns. Patients were analyzed as a separate stakeholder group due to their distinct perspectives and priorities. Other stakeholders are not included in this artifact as they were not interviewed.

Stakeholder 01: Clinicians (Nurses and Doctors)

- **Role of TriageAssist:** View clinical decision support systems as supportive tools, not primary decision-makers. Their ideal use involves post-assessment validation or risk flagging to reinforce clinical judgment.
- **Concerns:** Risks include technology over-reliance and automation bias, a lack of transparency, and a failure to capture non-verbal cues or atypical symptoms, especially in underrepresented or ill-documented patient subgroups.
- **Values:** Emphasize fairness, efficiency, intuitiveness, and adaptability within existing team-based workflows.
- **Implementation Preferences:** Preference for effective human-in-the-loop designs, preserving autonomy while enhancing triage consistency.

Stakeholder 02: Patients

- **Role of TriageAssist:** Patients see clinical decision support systems as tools to improve efficiency and reduce wait times, but believe they should be used to support rather than replace patient care.
- **Concerns:** Patients fear that AI system usage may results in a loss of human connection or empathy with clinicians, reduced communication in their healthcare experiences, and potential inaccuracies in AI assessments.
- **Values**: Highlight efficiency, transparency, and empathy from clinicians, emphasizing clear communication and human interaction practices.

- **Implementation Preferences:** Balanced AI integration that improves speed and yields transparent explanations of their care decisions without sacrificing personal care.

Supporting Material Artifacts

The internal documents provided by Prioritize™ include a brief description outlining the tool's origin during the COVID-19 crisis and a technology description detailing the system's goals around real-time cardiac risk prediction and its planned deployment in emergency room triaging workflows. Both of these documents emphasize the need for strong workflow alignment, equity considerations, and trust-building in product design and rollout.

Scientific Literature Artifacts

The *"Existing Inquiry and Related Work"* section provides a comprehensive review of various literary sources reviewed in addressing existing fairness and bias challenges and limitations, in addition to further limitations in how CVD risk prediction is conducted.

Stage 4: Testing

Considering each of these artifacts, key descriptive statistics of the dataset will first be investigated, followed by a three-part testing suite: an exploratory error analysis, patient subgroup testing, and adversarial testing. The completed analysis for all testing procedures in the following section can be found in our accessible online repository, linked in Appendix B. This repository provides reproducible code and additional insights into the exploratory analysis and fairness evaluations conducted, including additional summary statistics and plots.

Descriptive Data Analysis

An exploratory analysis of the dataset was first conducted to identify structural imbalances and specific subgroup trends, important for subsequent testing stages. This initial step serves to contextualize model behaviour and risk by examining how demographic and clinical features are distributed. Since our approach is model-agnostic (e.g., not tailored to a specific model architecture), the dataset artifact itself represents the most critical piece of information for identifying key issues. Understanding the data's structure and inherent biases is essential for anticipating potential performance disparities, regardless of the specific model used.

Given the assumption that Prioritize™ has not yet explicitly accounted for subgroup dynamics in TriageAssist's development, this analysis focused on disaggregating the data by the only two demographic features available: biological sex and age, for further intersectional comparisons. Two key concerns emerged, described below.

Concern 01: Underrepresentation of Female Population

The dataset reveals a significant sex imbalance, with male patients substantially outnumbering female patients. This disparity is especially pronounced in age ranges below 30 and above 70, suggesting a heavily skewed representation of younger and older female populations. This trend is visualized in Figure 5, where the distribution of age by sex demonstrates general female underrepresentation, and particularly limited data coverage for females in these critical age brackets. Such underrepresentation poses a serious risk for model fairness, as it may lead to lower predictive reliability for underrepresented groups due to insufficient data variability and volume during model training.

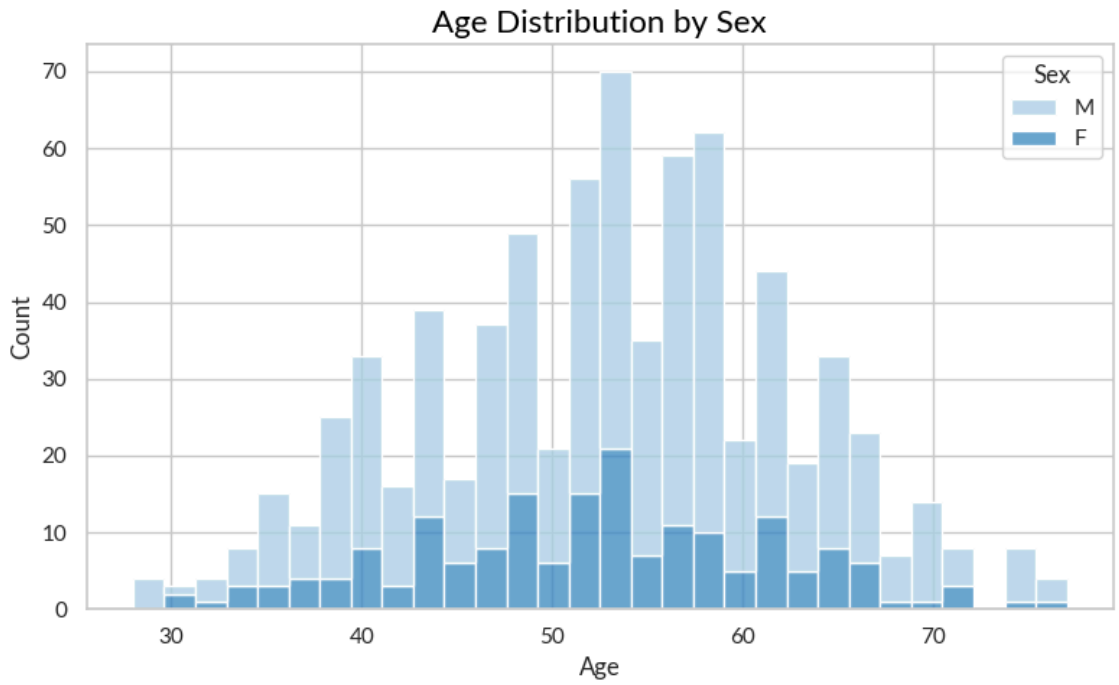

Figure 5: The age distribution plot comparing the biological sex breakdown indicates females by the dark blue bars, and males by the light blue bars. Distribution by age and sex demonstrates general female underrepresentation, and particularly limited data coverage for females below age 30 and above age 70.

Concern 02: Heart Disease Prevalence by Sex

As noted in Figure 6, there is a clear difference in the prevalence of heart disease between males and females. Based on the data imbalance, this is not surprising, however, this poses a very large prevalence bias, as male patients are indicated to have a much higher proportion of heart disease cases than females. Jaffer et al. (2021) establish that CVD differs in overall prevalence between the sexes, in addition to the Public Health Agency of Canada, which found that in 2018, approximately 7% of females and 10% of males over age 20 were diagnosed with heart disease. However, when analyzing Figure 6, approximately 89% of males are classified as having heart disease, and only 11% of females. This shows an extreme prevalence and manifestation mismatch,

which are only shifting toward higher female incidence rates as time progresses. This mismatch between the training data and the collected population statistics could lower the system's predictive reliability– the model's subsequent predictions are therefore at a high risk of perpetuating significant selection biases towards the majority class (e.g., males) (Lapalme et al., 2024).

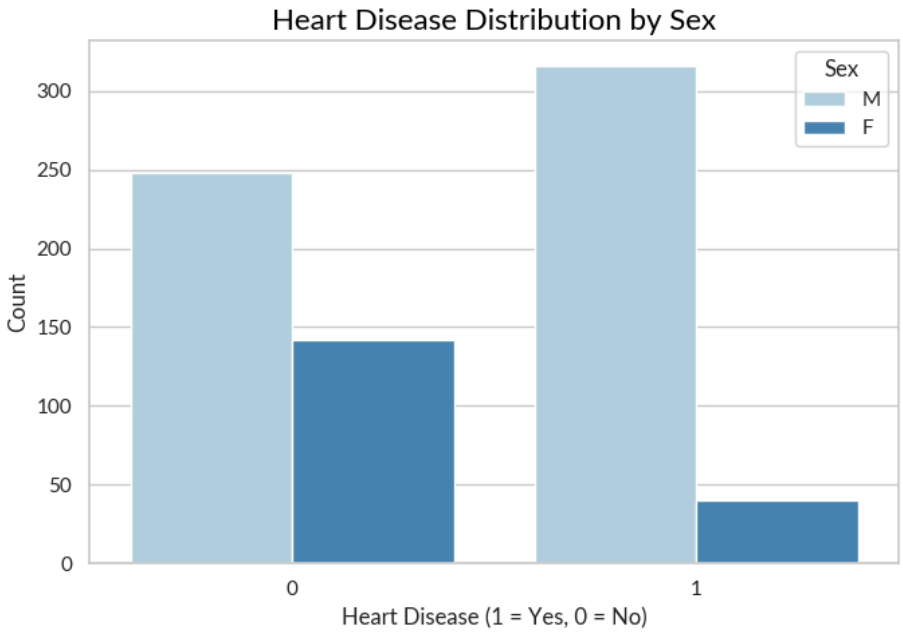

Figure 6: The heart disease distribution plot compared among biological sexes. Females are indicated by the dark blue bars, while males are indicated by the light blue bars. Despite an agreement that males have slightly higher incidences of heart disease, there is a significantly disproportionate skew in this dataset between males and females. As a result, the data used risks high selection biases and poorer predictive reliability.

Takeaways: General Directionality for Further Testing

The combination of sex-based underrepresentation and unequal disease prevalence across sexes creates an inherent risk of performance disparities within predictive models trained on this dataset. Models may inadvertently learn patterns that are more reflective of male patients, consequently marginalizing female users, especially those at either end of the age extremes (e.g., under age 30 or over age 70), where demographic representation is even lower. Since this analysis is model-agnostic, it directly informs the identification of key issues related to data imbalances and subgroup biases. These findings reinforce the need for careful performance and fairness evaluation across subgroups, thorough error analysis, and possibly, data balancing interventions in future modelling pipelines.

These insights lay the groundwork for subsequent testing phases, including **exploratory error analysis**, **subgroup testing**, and **adversarial testing**.

Exploratory Error Analysis

Exploratory error analysis involves systematically examining false-positive and false-negative groups within classification algorithms. This approach helps identify patterns and biases in model predictions, especially when handling diverse patient subgroups. Useful tools for this analysis include AI explainability methods, such as feature importance measures.

Feature importance analyses are conducted to better understand the influence of individual dataset features on heart disease prediction. A random forest model has been trained using the entire dataset. The primary goal of this model is not to optimize predictive accuracy, but to gain insights into feature relevance and impact. Random forest tree models are commonly used to understand feature importance because they provide a straightforward method for estimating the relative influence of each feature by analyzing the average decrease in impurity across all trees in the ensemble, making it an effective explainability technique.

The top fifteen input feature variations of importance are displayed in Figure 7. The three most important features identified are all measurable echocardiogram risk criteria: the ST slope up and flat features, and the old peak measure. Age is ranked as the sixth most important feature, while biological sex (e.g., of males) ranks ninth. The results therefore suggest that the age and sex features have a non-negligible influence on model outputs, and that the model itself prioritizes the male biological sex status, indicating significant potential demographic bias (Figure 7).

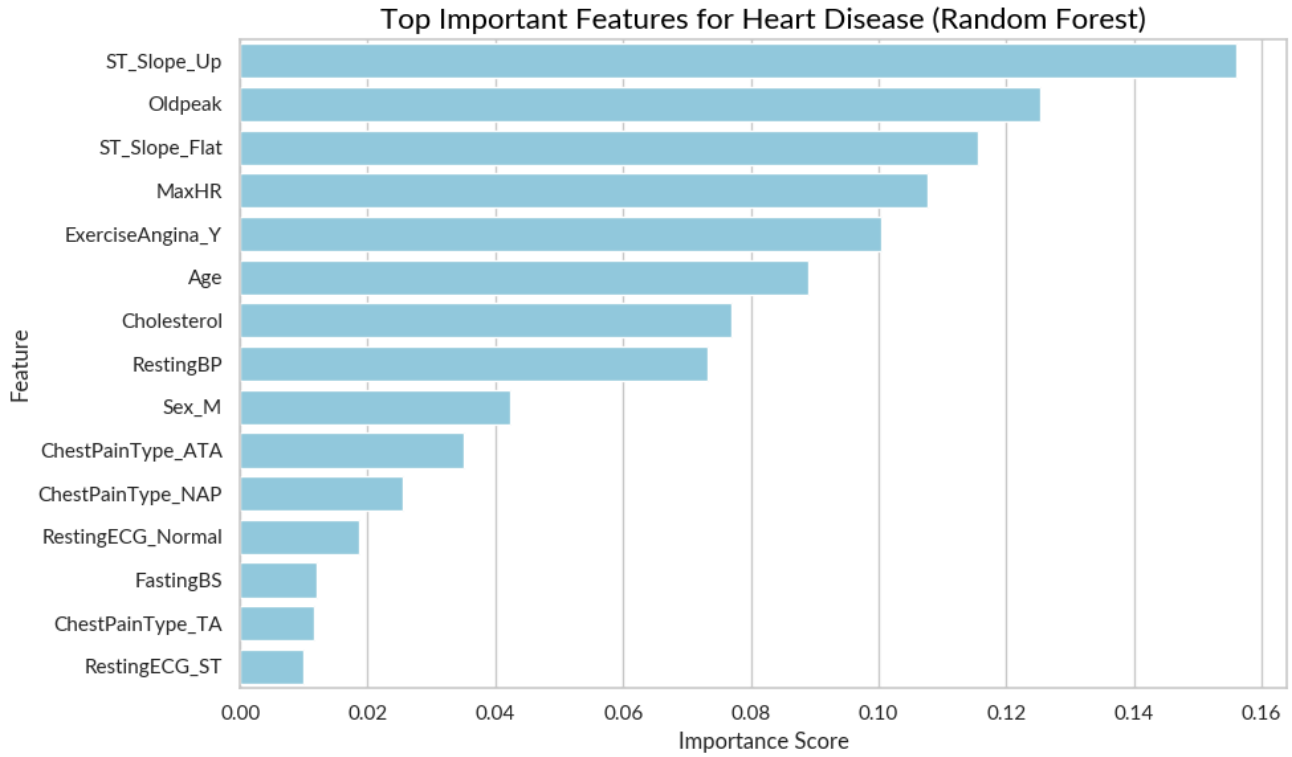

Figure 7: A ranking of the most important features for heart disease prediction within the dataset provided using a random forest model. The three most important features identified were ST Slope Up, Oldpeak, and ST Slope Flat. Age is ranked as the sixth most important feature, while male sex ranks ninth. The results suggest that age and biological sex have a non-negligible influence on model outputs, indicating significant potential demographic bias.

Shapley Additive Explanations (SHAP) Analysis

To further dissect and confirm feature influence, a Shapley Additive Explanations (SHAP) analysis has been conducted on the previously trained random forest model. SHAP values offer individual feature attributions, enabling the visualization of how each feature contributes to the model's output. Similar to Figure 7, ST slope up and ST slope flat were the most impactful features, however, the exercise angina feature became more prevalent (Figure 7).

Biological sex has a greater importance in this analysis, ranking fifth. From Figure 8, it can be observed with high granularity that when a patient is male (the pink data points on the plot), the sex feature has a stronger positive impact on predicting heart disease. In contrast, when the patient is female (the blue data points on the plot), the impact of the sex feature is negative and less significant; overall, female sex is not as strongly considered in TriageAssist's predictive modelling. This disparity suggests that the model is more likely to treat female heart disease occurrences in a lesser capacity than male occurrences, possibly due to the limited representation of female patients in the dataset.

Age is denoted as slightly less impactful in the SHAP plot, ranked as the eighth most important feature (Figure 8). As seen in the plot, the younger a patient is (blue data points), the less relevant the feature becomes to the model's output. On the contrary, older patients (pink data points) have a higher feature relevance, aligning with the general medical understanding that age is a significant determinant in risk factors for heart disease. Note the wide distributive spread and polar magnitudes of both the biological sex and age features as compared to the features listed at the bottom of the SHAP plot, which have much less impactful, centralized distributions (Figure 8).

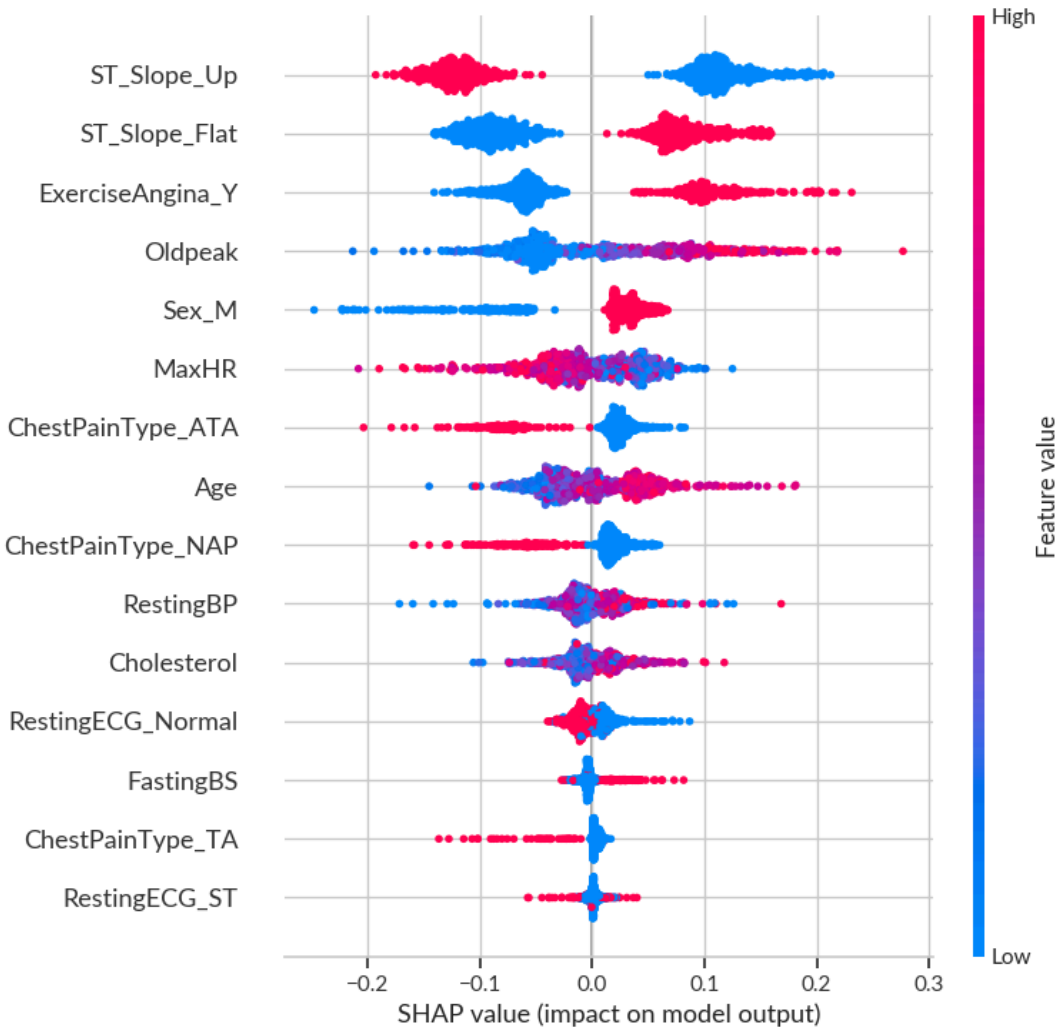

Figure 8: A Shapley Additive Explanations **(**SHAP) analysis indicating the impact of each feature used in the model (x-axis) on the overall weighting in the model (y-axis). Note that higher SHAP values indicate a greater importance within the modelling of heart disease risk. For biological sex, males are indicated by pink data points and females are indicated by blue data points. It can be concluded that data points of males are more highly prioritized over females, as they have higher SHAP values. Younger patients (blue data points) are also deemed to be less relevant in the model's outputs.

Takeaways: Implications of Exploratory Testing

While this analysis provides valuable insights into feature importance, it is important to note that these results are derived from a single random forest model, which has been chosen above due to its explainability properties. Therefore, they do not necessarily generalize to other modelling approaches. The key takeaway remains that demographic features such as biological sex and age are significant features, and the disparity between female and male impact on the system's predictions highlights the strong influences of data imbalances on the model's behaviour. The lack of adequate female data points affects the model's ability to effectively generalize to female patients, resulting in potential biases. Addressing this issue will require fairness-oriented model evaluations and data balancing strategies in subsequent testing phases.

Patient Subgroup Testing

To ensure a model-agnostic analysis, five models have been created, each representing a different methodological approach to classification. This diversity presents a more robust generalization of the findings when evaluating fairness across demographic subgroups. As described in the artifact collection section, the chosen models align with commonly used methods in similar studies:

- **Logistic Regression:** A linear, interpretable model.
- **Random Forest:** A robust, tree-based model.
- **K-Nearest Neighbors (KNN):** A distance-based model.
- **Naive Bayes:** A heavily probabilistic model.
- **Support Vector Machine (SVM):** A margin-based, flexible model.

To evaluate model performance across female and male subgroups, equal opportunity difference (EOD) was chosen as the primary fairness metric of interest. EOD measures the difference in true positive rates (TPR) between two demographic groups, aiming to assess whether both groups have similar chances of being correctly classified, in this use case, as having heart disease. A lower EOD value indicates fairer performance, as it demonstrates that the model treats both subgroups equally in identifying positive cases.

Minimizing EOD measures is critical because it addresses the risk of model bias, particularly in healthcare settings where diagnostic accuracy is vital. A high EOD implies that one group (e.g., females) has a systematically lower TPR, leading to potential disparities in the clinical decision support system being assessed. In CVD prediction, a false negative (misclassifying a patient *with* heart disease as healthy) poses a severe risk, potentially resulting in delayed or missed treatment. False positives, while inconvenient, usually result in further medical evaluation, which is safer. Therefore, achieving a low EOD ensures that both male and female patients are equally likely to receive an accurate and timely diagnosis. The selection of EOD as a fairness metric is validated by existing literature, such as Li et al. (2023), where similar considerations are discussed in the context of CVD risk prediction.

The code utilized stratified k-folds with six splits to maintain consistent class distribution across the training and testing sets. This technique ensures that performance metrics are not skewed by random variations in the data distribution, increasing the generalizability of the model performance results. The code also calculates the mean and standard deviation for each metric, offering a reliable estimate of model stability across different folds. The complete results can be found in the online code repository linked in Appendix B.

Figure 9 shows the mean accuracy, mean general TPRs, and subgroup TPRs for females and males. The disparity between biological sexes is evident, for instance, in the logistic regression model, the mean female TPR is only 69%, whereas the male TPR is much higher at 89%, highlighting the challenge of achieving balanced performance across sexes. This trend can be seen in all model types, which forecast higher accuracy and TPRs for males than females.

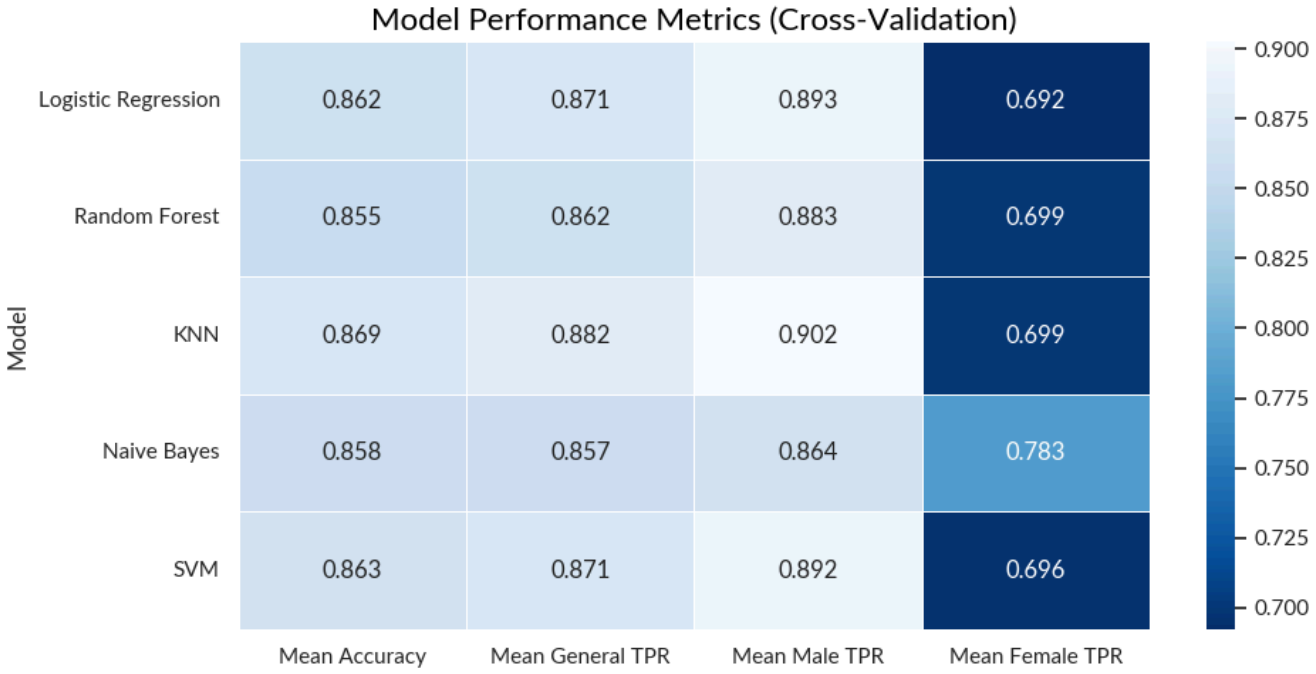

Figure 9: Comparison of various model performance metrics (scaling from 0 to 1) across different modelling algorithms on the heart disease dataset, including mean accuracy, the mean true positive rate across all populations, and the mean true positive rates across males and females. Higher values for all metrics are desired. The disparity between male and female predictions is evident, highlighting performance discrepancies across biological sex.

However, knowing the accuracy and TPR performance for different models across both sexes is not enough to measure unfairness. To address this, Figure 10 delves deeper into fairness by visualizing the mean EOD, described earlier, between male and female subgroups. The Naive Bayes model displays the lowest EOD (0.081), suggesting it handles sex-based disparities better than the other models. In contrast, the KNN model has the highest EOD (0.203), indicating a larger gap between male and female TPRs.

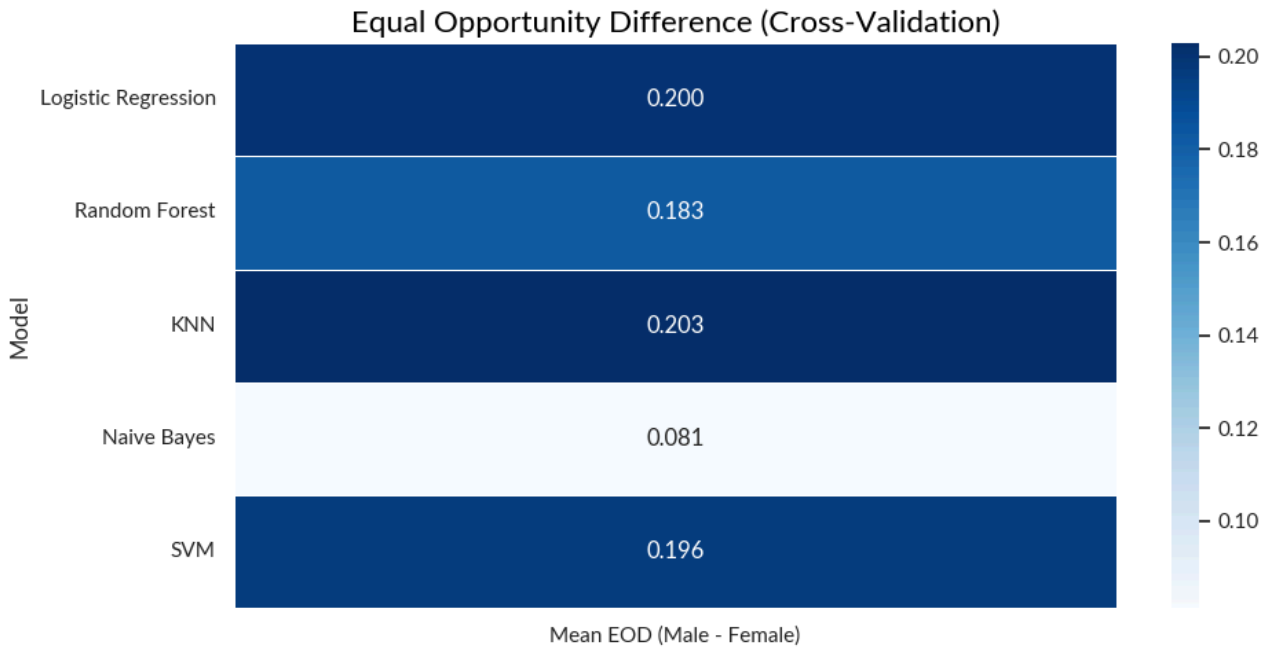

Figure 10: A gradient of the mean equal opportunity differences (EOD) is displayed across five model types. Lower EOD values indicate fairer performance. The Naive Bayes model displays the lowest EOD (0.081), suggesting it handles the sex-based disparity better than the other models. In contrast, the KNN model has the highest EOD (0.203), indicating a larger gap between male and female true positive ratios.

Takeaways: Implications of Patient Subgroup Testing

The observed discrepancies between male and female TPRs and the relatively high EOD values in several models (e.g., logistic regression, random forest, KNN, and SVM) indicate a potential risk of model bias. One commonly referenced criterion for evaluating algorithmic fairness is the 4/5th rule (or 80% rule), which suggests that a fairness ratio of less than 0.8 (or an EOD difference of 0.2) may indicate unfair treatment. Although this threshold is not universally applicable, it provides a clear and practical benchmark to assess potential bias, as highlighted by Groves et al. (2024). Since this analysis adopts a model-agnostic approach and most of the evaluated models exceed or are very close to this threshold, it is reasonable to conclude that the risk of bias is significant. This conclusion remains valid regardless of the specific model used, emphasizing the importance of addressing this issue systematically.

One of the key risks identified is that female patients may be incorrectly classified as not having heart disease, potentially leading to missed or delayed treatment. This risk is related to the high EOD values, where the disparity between male and female TPRs indicates that females are systematically less likely to be correctly identified as having heart disease compared to males. This underscores the critical need for incorporating fairness-aware techniques when developing predictive models for healthcare applications. To mitigate these biases, it is essential to incorporate fairness-aware techniques during model development. Possible methods include

debiasing approaches such as reweighting, resampling, and implementing fairness constraints during training. Additionally, gathering more female patient data, especially at either age extreme, would improve demographic imbalances and work to create more equitable models that better generalize across diverse patient populations.

Adversarial Testing

Adversarial testing involves simulating adverse conditions to evaluate model robustness, particularly focusing on appropriate edge cases. This type of testing is essential to identify how models handle underrepresented patient subgroups, which may be prone to prediction errors. To perform adversarial testing, the aim is to simulate scenarios challenging the model's predictive accuracy by focusing on intersectional edge cases. Specifically, our team created a baseline profile of the most typical heart disease patient, which follows the profile of a middle-aged male; the full set of corresponding input feature values (e.g., cholesterol and resting blood pressure) for this baseline is detailed in Appendix C. To create this baseline, the mode (most frequent value) was identified for each feature among any patient diagnosed with heart disease. Then, the age and biological sex were adjusted from this baseline to analyze biological sex and age variations:

- **Biological sex variations:** Male and female.
- **Age point variations:** 18, 58, and 80 years of age.

These ages were chosen as representative edge cases since there is limited data on younger female patients (below age 30) and older female patients (above age 70). The assumption is that as age decreases or increases, the model accuracy may decline, especially for females, due to the data scarcity described earlier. Given that all other clinical features remain indicative of heart disease, the models are expected to predict a high likelihood of heart disease in these scenarios. The same models used in the *"Subgroup Testing"* section were utilized for consistency; however, in this adversarial testing, all models were trained on the entire dataset, rather than performing k-fold cross-validation, to maintain a realistic yet consistent comparison.

Figure 11 shows that for ages 58 and 80, the predicted probability of having heart disease is similar for both males and females across all models, indicating relative model stability for older patients. However, for age 18, significant differences emerged. The predicted probability for male versus female patients was markedly different, highlighting a potential source of bias. *In this particular sample, the logistic regression model classified the 18-year-old female as not having a heart disease, despite all other indicators pointing to a high likelihood.* This discrepancy can be extremely dangerous, as it may result in missed critical diagnoses for young females.

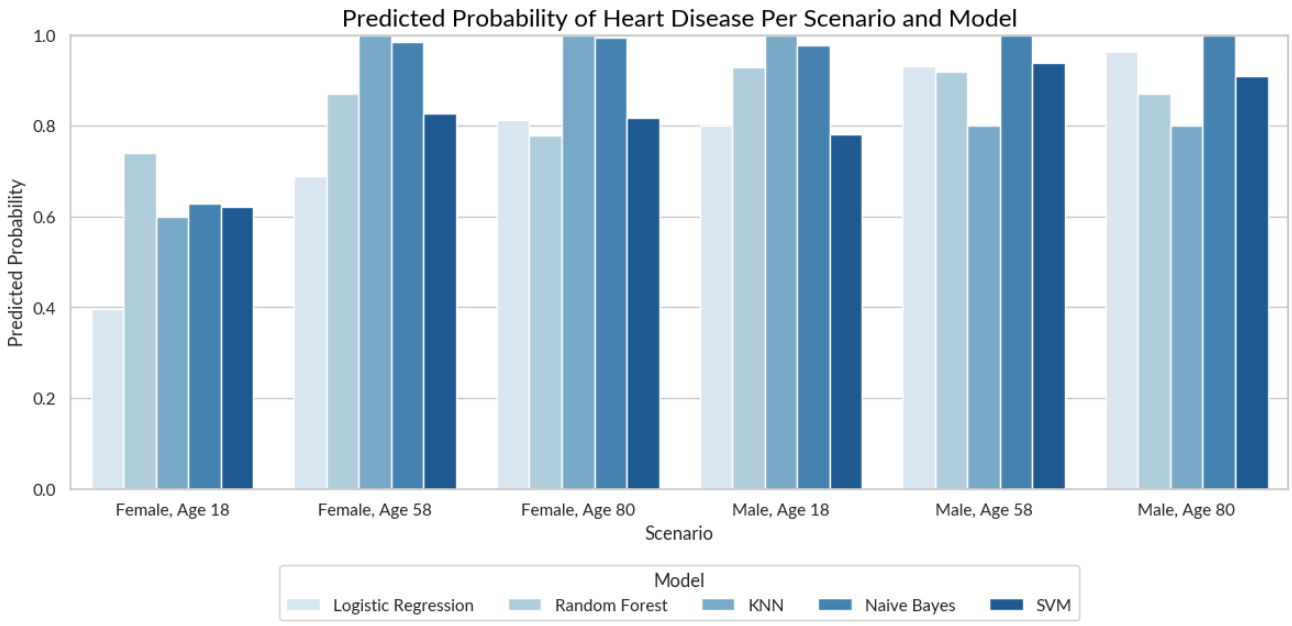

Figure 11: The predicted probability of heart disease across six different testing scenarios (listed along the x-axis). Higher probabilities are expected since the scenarios represent the most typical heart disease symptoms while only modifying age and biological sex. If the predicted probability exceeds 0.5, the model predicts that the patient has heart disease. For ages 58 and 80, the predicted probability of having a heart attack was similar for both males and females across all models, indicating relative model stability for older patients. However, for age 18, significant differences emerged, particularly for the female sex.

Takeaways: Implications of Adversarial Testing

The findings indicate that the intersection of biological sex and age is a critical factor when analyzing model performance; addressing bias requires focusing not only on biological sex distributions but also on age distributions within the data. A failure to account for this intersectionality may lead to systematic errors, particularly for younger females, where the model's predictions are less reliable. Unfortunately, as seen in the testing results, TriageAssist is at high risk for these errors. Therefore, data collection efforts should prioritize enhancing representation not just by biological sex but also by age, especially at the extremes of the spectrum. This dual consideration will help create fairer and more reliable predictive models for CVD risk assessment.

Stage 5: Synthesis of Key Audit Matters

Three parties have been identified within the synthesis of key audit matters: developers, clinicians, and company-wide management. The demarcation of each party's role and capabilities in specific development and deployment tasks and responsibilities requires a granular set of actions for each; however, executive management will ultimately steer, guide, and prioritize the decisions behind the assignments of any recommendations.

Developer Actions

Based on the findings from the different testing stages, five technical actions are recommended to improve the fairness and robustness of TriageAssist. The following key areas of improvement are identified specifically for developers, denoted by D.

Recommendation D1: Address Data Imbalance Through Data Preprocessing Techniques

One of the primary challenges identified during the exploratory error analysis and subgroup testing is the disparity between male and female TPRs and EODs, which indicates potential bias in model predictions. To address this, it is critical to adopt fairness-aware techniques throughout the model development pipeline. The most promising approaches include debiasing methods such as reweighting, resampling, and applying fairness constraints during training. An experiment was conducted to assess the impact of reweighting on model performance and fairness, as shown in Figure 12. Stratified k-folds with six splits have been used to maintain consistent class distribution across the training and testing sets. *The results of reweighting are promising.* The logistic regression and SVM models show a negligible accuracy decrease, while the EOD decreases significantly. The random forest models experience a substantial accuracy drop (over 20%), although the EOD improved notably, from 0.18 to -0.01. Detailed metrics and a visual representation of these improvements are shown in Figure 12 (impact on predictive accuracy) and Figure 13 (EOD reduction).

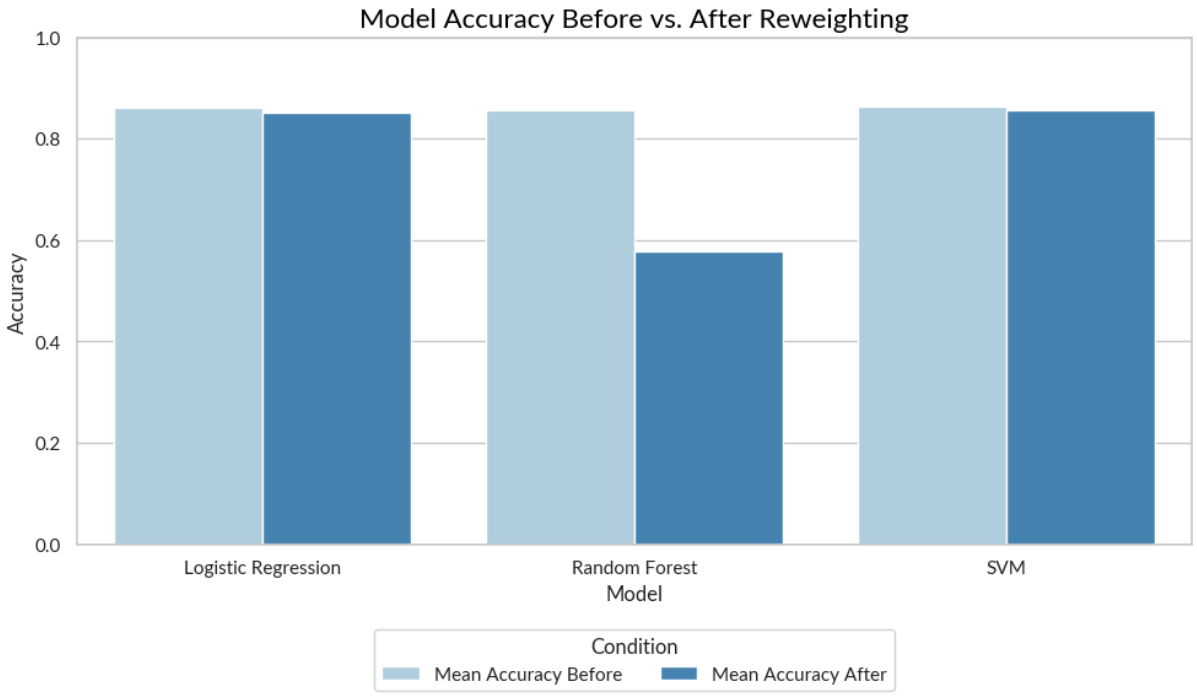

Figure 12: A comparison of model accuracy before and after reweighting techniques have been applied across three models: logistic regression, random forest, and support vector machines. The logistic regression and support vector machine models show negligible decreases in accuracy.

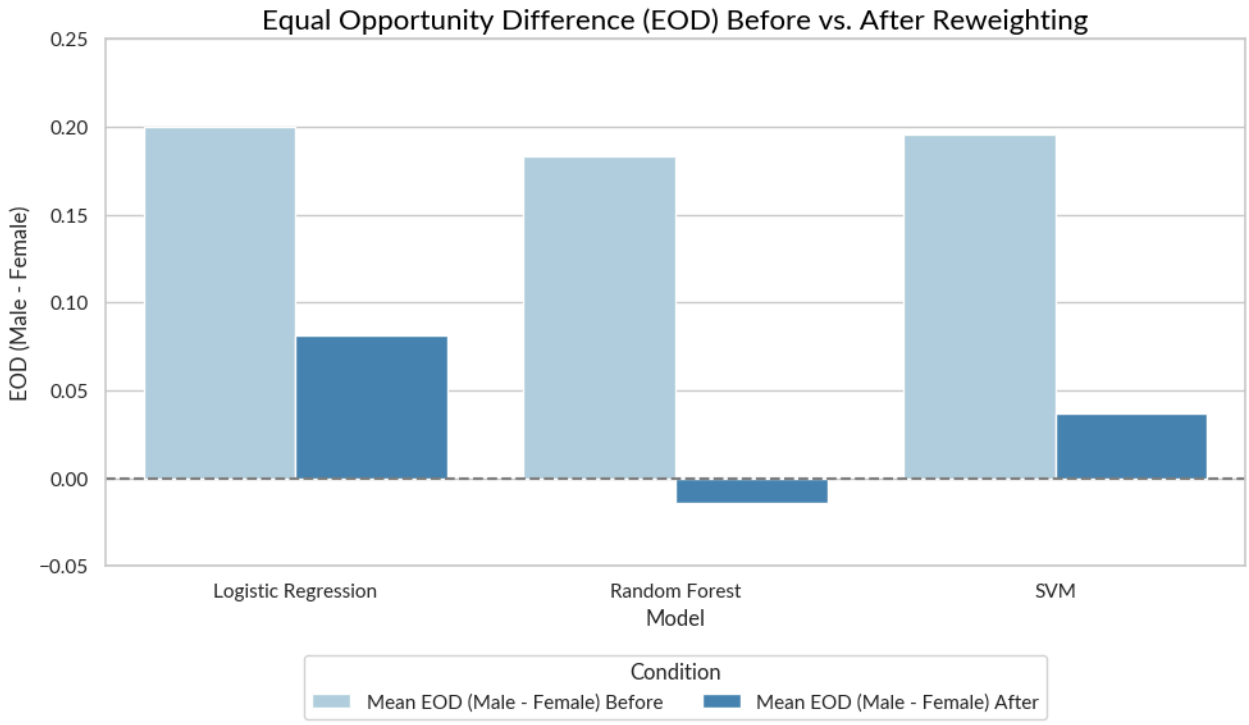

Figure 13: A comparison of equal opportunity difference (EOD) measures before and after reweighting techniques have been applied across three models: logistic regression, random forest, and support vector machines. The EOD decreases significantly in all cases.

While reweighting proves effective for some models, it is important to recognize its limitations. While some models, like SVM, handle reweighting well in terms of accuracy, others, particularly the random forest model, exhibit a significant accuracy drop. Not all models are suitable for reweighting, particularly KNN and Naive Bayes, which are sensitive to changes in the data distribution. In light of these results, developers should carefully assess which model architecture and preprocessing techniques align best with the goals of accuracy and fairness. Models that naturally accommodate fairness adjustments are recommended.

Recommendation D2: Address Data Imbalances Through Expanded Data Collection

The descriptive data analysis and adversarial testing reveal a clear imbalance in the dataset, displaying an underrepresentation of female patients, particularly those at either age extreme. This imbalance directly affects model performance, as it limits the algorithm's ability to generalize to either demographic subgroup. To mitigate this risk, data collection efforts should focus on enhancing representation not only by sex but also by age. Prioritizing data from younger females (below age 30) and older females (above age 70) is essential in ensuring balanced model performance across subgroups. Further, it is essential to conduct more systematic research on CVD, especially in females, as existing clinical decision support systems lack diversity in causal structures (Li et al., 2023). **If acquiring more data is not feasible, the existing dataset should be**

carefully evaluated using remediation techniques, as suggested in Recommendation D1, to reduce the representation gap.

Recommendation D3: Conduct More Comprehensive Fairness Testing Protocols

Addressing fairness in TriageAssist requires a broader perspective that includes intersectionality and further social determinants of health to optimize demographic enrichment data. Beyond biological sex and age, additional attributes like race, socioeconomic status, and geographical data should also be sought and considered within TriageAssist. Including these dimensions in fairness evaluations can help identify more nuanced biases that may otherwise be overlooked. Moreover, intersectionality testing should be expanded to cover several combinations of demographic variables, as the interplay between biological sex and age has already demonstrated significant discrepancies in model predictions. By understanding how multiple demographic factors intersect in predictive measures, TriageAssist can be adjusted accordingly to perform equitably across a wide range of patients. Reducing an overdependence on specific features that may lead to inaccurate assumptions is essential, as highlighted by Nolin-Lapalme et al. (2024). By refining the data pipeline to address these concerns, TriageAssist can better accommodate the diversity inherent in clinical settings.

Recommendation D4: Enhancing Transparency and Trust Through Confidence Scores, Warnings, and Explainable (XAI) Techniques

To increase patient and clinician confidence in TriageAssist, confidence scores should be displayed for each prediction. Providing model performance indicators, such as accuracy and prediction confidence, has been shown to improve users' trust and acceptance of AI-generated recommendations. As Yin et al. (2024) share, confidence metrics help users better interpret AI outputs and make more well-informed decisions. Furthermore, if known issues are present among TriageAssist, particularly those that do not have simple or accessible fixes (e.g., data representation), warnings and limitations of use, encouraging additional caution from users should be presented in the output. Additionally, incorporating XAI techniques, such as Shapley values, can further enhance transparency by showing feature contributions to model decisions. This not only makes the system more interpretable but also builds evidence-based credibility among clinicians who rely on these insights for critical decision-making.

Recommendation D5: Consider SVM as a Suitable Model For Implementation

Given that the logistic regression and SVM models maintained stable accuracy while reducing EOD, SVM stands out as one of the most suitable choices when evaluating the trade-offs between models. The SVM model's robustness to edge cases, high accuracy before and after reweighting, and reduced EOD make it a strong candidate for Prioritize™'s expanded deployment plans.

Clinical Actions

Three actions, denoted by a C, are recommended for nurses and doctors within the direct usage scope of TriageAssist.

Recommendation C1: Proper Human-in-the-Loop Placement of TriageAssist

TriageAssist should only be used after the initial nurse assessment has been completed to ensure that it is only enhancing clinical decisions rather than directly making them. This recommended workflow is akin to the use case presented in Figure 2. This ensures that clinicians maintain the ultimate decision-making power in triaging and ultimately reduces the automation bias that is likely to occur among clinical decision support systems. If the system is used before a nurse's initial assessment, it is likely to bias their own triage diagnoses, reduce the nuanced symptoms that the patient is experiencing, and ultimately risk an overreliance on potentially harmful diagnoses guidance, resulting in higher liability and reputational risks for stakeholders and Prioritize™.

Recommendation C2: Additional Allocation of Time and Effort in Triaging Underrepresented Demographics

Based on the discrepancies between biological sex and age attributes within the modelling and output stages, we recommend that clinicians are especially careful in triaging female patients, especially those at either extreme of the age spectrum. Any additional time available during emergency room triage should be carefully prioritized for patient subgroups who currently face, and have historically faced, the highest risk of misclassification by clinicians, clinical decision support tools such as TriageAssist, and the broader healthcare system. Spending more time to confirm nuanced differences in symptom presentation is critical to patient well-being for *all* populations.

Recommendation C3: Factoring Confidence Scores, Explanatory Warnings, and Usage Limitations Within Triaging Decisions

Building on the developer action to add confidence scores for each prediction to calibrate trust and provide more appropriate warnings and limitations of use, we recommend that clinicians take the time within the triaging process to consider these additional points of uncertainty in their assessment. For instance, if the TriageAssist system outputs low confidence scores or limitations of use within a particular edge case, clinicians should be more cautious in taking the corresponding CVD prediction of that patient into account. Rather, they should take a more dominant human-in-the-loop positioning. Similarly, if warnings are apparent for various demographic subgroups, clinicians should factor this into their system usage (e.g., be less likely to use the system outputs in cases deemed to have severe warnings).

Company-Wide Management Risk Mitigation Measures

There are five more generalized, company-wide, risk mitigation measures (denoted by an M), particularly for executive-level management members to implement in a larger scope across the company. These measures are beneficial for all stakeholders.

Recommendation M1: Delay Further Downstream Use of the Model Until Course Correction Implementation

We believe that each of the recommendations provided above are essential to the safe operation of TriageAssist, ultimately only providing benefit to Prioritize™. Hence, management should proactively support and work toward the implementation of these course corrections, particularly within the development pipeline, before transferring the modelling and algorithm within the pilot system to even more products, services, and settings, which will only further perpetuate these biases and failures of care. Ensuring TriageAssist covers more diverse populations also provides a larger client base with stronger market positioning and a more agile response when expanding to more challenging regulatory markets requiring more extensive and granular patient performance testing, such as the Food and Drug Administration in the United States (Digital Health Center of Excellence, 2022). Additionally, implementing the recommended course corrections reduces the likelihood of potential liability or reputational risks from the company's perspective and those using TriageAssist.

Recommendation M2: Provide Accessible and Continual Feedback Mechanisms

Due to the rapid changes that can occur within the real-time use of clinical decision support systems, continuous monitoring is critical in ensuring reliability, especially from those working directly with it daily. Prioritize™ therefore has to ensure there are ample feedback mechanisms for direct input from those on the frontline– developers, nurses, doctors, patients, hospital staff, and members of the community, to raise any questions, suggestions, or concerns they have with TriageAssist. Equipping users with these pathways fosters user engagement and ownership, making it known that they have a voice in its operation. Ultimately, even outside of these audits, a well-performing algorithmic system can fail in unexpected ways when faced with edge cases or adversarial inputs. Therefore, accessible feedback is essential in providing high levels of visibility for management to consider as issues arise, helping to proactively catch design flaws or misclassifications, usage and workflow problems, or safety concerns in diverse, contextual environments before they escalate.

Recommendation M3: Provide Appropriate Usage Resources for Nurses and Doctors

Based on the numerous human-in-the-loop pathways that clinicians can take when using TriageAssist, either as a pre-screening tool or as a cross-validation check, Prioritize™ must ensure nurses and doctors are aligned on the appropriate implementation of TriageAssist. To do so, we recommend the open access and frequent distribution of training resources such as detailed

scenario-based demonstrations, interactive training sessions, tutorials, and workshops, in addition to pamphlets. This training should be mandatory for all clinicians during TriageAssist's initial deployment, and continuously as new features are introduced, new data sources are integrated, or usage patterns are modified– any of which may affect the product's scope, limitations, or associated warnings. In doing so, we anticipate the reduction of potential malpractice incidents and the development of mistrust of the system.

Recommendation M4: Encourage a Generalized Awareness and Usage of Reflexive Responsible AI Guidelines, Tools, and Frameworks

Building on the prior recommendation, aside from specific usage training for clinicians, Prioritize™ should mandate more generalized training on algorithmic fairness practices to ensure those internal at the company understand their own positionality, presumptions, and dispositions surrounding user, interpretation, control, and cause-effect biases within their specific roles and impacts. Several benefits arise from providing stakeholders (e.g., developers, hospital providers, clinicians, management) with responsible AI guidelines, tools, and frameworks. For instance, Nolin-Lapalme et al. (2024) note increased consistency and transparency when using the Machine Learning Reproducibility checklist. Additionally, the open-source FairMLHealth platform can be used on clinical decision support systems to diversify and explore various data and model selection options across numerous fairness metrics more efficiently than traditional evaluation processes. Furthermore, more interactive programs, including Mila's Trustworthy and Responsible AI Learning (TRAIL) Certificate for Professionals, exist to work through independent hazard identification and mitigation across various algorithmic contexts. In general, ensuring stakeholders are made aware that these instruments exist– and that the company supports their exploration, use, and participation– fosters a culture of reflexive growth.

Recommendation M5: Only Consider the Implementation or Partnership of TriageAssist Among Hospitals and Organizations Supporting Equity, Diversity, and Inclusion Measures

Our last recommendation is to insist that Prioritize™ only implements TriageAssist within hospitals or deployment environments where the stakeholders involved advocate and believe in equity, diversity, and inclusion for the entire patient population, across *all* patient subgroups. The findings noted in this audit require a nuanced level of care and understanding as to why it is so important to carefully consider and implement course corrections, and providers and users need to be aligned on these choices, and more generally, the ethics behind them. Likewise, as Prioritize™ expands, especially within the particular sociotechnical political climate at hand, our team encourages partnerships only with organizations that are aligned with the same ethos presented in this report. Aside from the utmost priority of patient well-being, if Prioritize™ chooses to move forward within this framing, it will only reinforce a positive brand image and open more opportunities for EDI-related grants, funding, and awards.

Limitations of Audit

This audit has a significant time constraint of five weeks. Despite its comprehensive findings, several actions were not feasible in this timeframe. For instance, a limited number of fairness metrics, such as EOD, were considered within the patient subgroup testing, and this audit acknowledges that more measures could have been used to consider various qualitative, self-reported, or behavioural measurement methods. Other ethical principles also could have been investigated further, such as privacy or sustainability. In future audits, more comprehensive information should be provided to look at the product's ecosystem from a broader lens. Due to Vantamed+'s sociotechnical expertise and capabilities, we performed an audit that best suited our strengths, however, we acknowledge that we did not perform a specific legal analysis across any relevant laws or regulations (e.g., *Canada's Directive on Automated Decision-Making* or *Voluntary Code of Conduct on the Responsible Development and Management of Advanced Generative AI System*s) that could better target specific compliance concerns. We recommend obtaining additional legal consultation or commissioning a regulatory compliance audit to ensure that TriageAssist aligns with current and emerging legal frameworks, particularly as Canadian AI governance continues to evolve.

There were also resource constraints on the dataset and modelling information provided, which biased the scope of the audit. More broadly, our team lacked access to the specific model type, weightings, and features that TriageAssist specifically used. Without this information, we were unable to completely validate the impact of their specific algorithmic selections. Any data or results from the initial pilot deployment were also lacking, other than four stakeholder interviews. Due to the limited accessibility of current performance data and information on how the technology is being used in practice, our analysis required several technical and usage-related assumptions.

At a more granular scale, the dataset used also only contained two protected attributes, biological sex and age. Demographic attributes, including ethnicity, sociotechnical status, or various geographic populations, were not present in the dataset, and therefore, certain patient subgroups and intersectionalities could not be investigated. It is also important to note that the biological sex feature we focused on only included female and male subcategories; it did not look at intersex individuals, who were therefore excluded from any testing within TriageAssist.

Takeaway: Limitations

This audit was conducted under a five-week time constraint with limited access to model architecture and demographic data, which restricted the scope of possible technical and ethical evaluations. While the audit leveraged Vantamed+'s sociotechnical strengths, future audits should include legal analysis, broader demographic representation, and access to more comprehensive system and deployment data for a more complete assessment.

Stage 6: Post-Audit Measures

Re-Auditing TriageAssist

This initial audit provides a strong baseline with thirteen tangible key audit matters, however, continuous monitoring is necessary to ensure the recommended changes are addressed and, more pointedly, implemented. Liu et al.'s (2022) framework does not recommend a particular frequency of re-auditing; however, we recommend a re-audit at several key checkpoints, including:

- Once the TriageAssist platform's pilot deployment has been completed to factor in real-time user feedback, as well as in any subsequent changes within its operational environments, contextual settings, or usage workflows.
- Within any pivotal change in the ML pipeline development of TriageAssist, including the implementation of a new dataset(s), re-training, the selection of a new algorithmic model(s), or new evaluation measures.
- At any period when TriageAssist has identified that their patient demographics have changed significantly.
- New regulatory or compliance mandates and precedents have been identified or set internally within Prioritize™, or more generally, released within the Canadian or global medical device industries.
- Following any incident or catastrophic failure that is suspected to occur, or has occurred, within the TriageAssist system or similar clinical decision support systems in the market.
- After any public complaints and/or legal actions have been found, suspected, or taken against the TriageAssist product or Prioritize™.

Takeaway: Re-Auditing TriageAssist

There are at least six checkpoints for which an audit should be re-conducted for areas of concern.

Concluding Statement

This algorithmic audit was conducted by Vantamed+ to evaluate TriageAssist, a clinical decision support system aiming to help cardiovascular disease triaging in emergency room settings. The system was developed by Prioritize™. This audit follows Liu et al.'s (2022) medical algorithmic audit framework, chosen for its contextual relevance in this particular use case, academic and industry credibility, and holistic, end-to-end, third-party approach within algorithmic evaluation. Each of the framework's six stages has been analyzed in this report: scoping, mapping, artifact collection, testing, synthesis of key audit matters, and post-audit measures.

Based on an initial review of the documentation provided and a detailed literary analysis of in-market competition, our team explored the most prevalent concern found within the documentation provided by Prioritize™: ensuring that TriageAssist performs equitably across

diverse patient subgroups, particularly with respect to biological sex and among its intersections with age. Given the known disparities in CVD predispositions, symptom presentations, treatment efficiencies between males and females, and growing mortality risk factors within Quebec, this report covered two objectives: first, it evaluated the current landscape of how TriageAssist is currently performing amongst different patient demographic subgroups, and second, it informed Prioritize™ of improvement areas regarding any discrepancies found, to allow their company to engage in actionable course corrections derived from evidence-based key audit matters. Ultimately, in addressing these factors, TriageAssist could better ensure fairness, safety, and performance among its product's deployment to improve its trust and adoption among the system's positioning in human-in-the-loop workflows, reduce liability and reputational risk, and better prepare for market expansion among funding and strategic partnership avenues.

Takeaway: Critical Issues

The audit identifies four critical issues through a comprehensive multi-step process, including stakeholder value mapping, usage process workflows, descriptive data analysis, exploratory error analysis, patient subgroup testing, and adversarial testing. In order of highest to lowest priority, the issues are as follows:

- **Data Imbalances:** The most concerning finding in this audit is the significant underrepresentation and prevalence of female patients in the dataset, particularly in the younger and older extremes of age groups. As the data used in the system bounds the modelling and evaluation stages, this negatively impacts the entire machine learning pipeline.
- **Usage Workflow Challenges:** There is a potential over-reliance on the clinical decision support system's algorithmic outputs. This is particularly true if the system is used as a pre-screening tool by nurses and doctors rather than for cross-validation checks, which can overshadow critical human verification steps and lead to strong automation biases with increased liability and reputational risks.
- **Model Biases:** Given the imbalanced dataset, various modelling techniques tested were severely impacted, presenting inconsistent predictive performance between sexes. Compared to the most common patient baseline profile in the data, a middle-aged male, younger females (< age 30) are at the highest risk of being misclassified by TriageAssist, followed by older females (> age 70).
- **Lack of Transparency:** The absence of confidence scores, warnings, limitations of use, and explainable outputs in TriageAssist can make it difficult for clinicians to interpret and use model predictions appropriately within the system's technical capacities and boundaries.

To further remedy or mitigate these issues, all key audit matter recommendations have been summarized. Company-wide risk mitigation measures are listed first, requiring the most imperative attention. These are followed by the second highest priority, developer actions, as they carry the most downstream, cascading effects, followed by clinical actions, deemed to be more

indirect precautionary measures. Each measure or action has been ranked from highest to lowest priority to implement.

Takeaways: Recommended Remedies or Mitigations for Areas of Concern

---

Company-Wide Risk Mitigation Measures

- **Priority 01:** Implement recommended course corrections of TriageAssist before expanded downstream use of the model across diverse clinical settings.
- **Priority 02:** Construct accessible and continual feedback mechanisms for stakeholders, including developers, nurses and doctors, and patients, surfacing further problem analyses within TriageAssist's development and deployment.
- **Priority 03:** Provide appropriate training resources and encourage the awareness and use of generalized responsible AI guidelines to ensure appropriate reflexivity within each stakeholder's roles and tasks.
- **Priority 04:** Only partner with hospitals, funding agencies, research institutions, and non-profit organizations that support equity, diversity, and inclusion measures at the forefront of their mission.

Developer Actions

- **Priority 01:** Address data imbalances through specialized data preprocessing techniques and more extensive data collection for various patient demographic subgroups, particularly young female patients.
- **Priority 02:** Conduct more comprehensive fairness testing protocols, including more comprehensive intersectionality testing among various cross-sections of patient demographics.
- **Priority 03:** Enhance transparency and trust through the addition of confidence measures, warnings, limitations of use, and the further application of explainable (XAI) techniques.
- **Priority 04:** Consider support vector machine (SVM) models to be among the most suitable for implementation in TriageAssist.

Clinical Actions

- **Priority 01:** Promote the specific human-in-the-loop placement of TriageAssist to avoid automation bias, particularly its use as a cross-validation tool after the initial nurse assessment of the patient has been completed.
- **Priority 02:** Allocate more time, effort, and resources to emergency room triaging for females and those at either age extreme.
- **Priority 03:** Appropriately consider and factor in confidence scores, explanatory warnings, and usage limitations within the model outputs when making triaging decisions.

Should you have any questions, comments, or concerns regarding the information shared in this report, please do not hesitate to contact our team. As a third-party auditor, Vantamed+ is committed to supporting responsible, evidence-based AI integration in healthcare. We look forward to assisting with the implementation of these recommendations and collaborating on future audits to ensure continued improvement and accountability in TriageAssist.

References

Campbell, N. R., Ordunez, P., Giraldo, G., Morales, Y. A. R., Lombardi, C., Khan, T., ... & Varghese, C. (2021). WHO HEARTS: a global program to reduce cardiovascular disease burden: experience implementing in the Americas and opportunities in Canada. Canadian Journal of Cardiology, 37(5), 744-755.

Digital Health Center for Excellence. (2024, December 20). Clinical Decision Support Software Frequently Asked Questions (FAQs). U.S. Food and Drug Administration. https://www.fda.gov/medical-devices/software-medical-device-samd/clinical-decision-support-software-frequently-asked-questions-faqs

Federico Soriano. (2021). Heart failure prediction (Version 1) [Dataset]. Kaggle. Retrieved from https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction

Groves, L., Metcalf, J., Kennedy, A., Vecchione, B., & Strait, A. (2024). AuditingWork: Exploring the New York City algorithmic bias audit regime. Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), 1-14.

Jaffer, S., Foulds, H. J., Parry, M., Gonsalves, C. A., Pacheco, C., Clavel, M. A., ... & Norris, C. M. (2021). The Canadian Women's Heart Health Alliance ATLAS on the epidemiology, diagnosis, and management of cardiovascular disease in women—chapter 2: scope of the problem. CJC Open, 3(1), 1-11.

Lapalme, A. N., Corbin, D., Tastet, O., Avram, R., & Hussin, J. G. (2024). Advancing fairness in cardiac care: Strategies for mitigating bias in artificial intelligence models within cardiology. Canadian Journal of Cardiology, 40(*10*), 1907-1921.

Li, F., Wu, P., Ong, H. H., Peterson, J. F., Wei, W. Q., & Zhao, J. (2023). Evaluating and mitigating bias in machine learning models for cardiovascular disease prediction. Journal of Biomedical Informatics, 138, 104294.

Liu, X., Glocker, B., McCradden, M. M., Ghassemi, M., Denniston, A. K., & Oakden-Rayner, L. (2022). The medical algorithmic audit. The Lancet Digital Health, 4(5), e384-e397.

Ming Yin, Jennifer Wortman Vaughan, and Hanna M. Wallach. (2019). Understanding the Effect of Accuracy on Trust in Machine Learning Models. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 1-12. ACM.

Mulvagh, S. L., Colella, T. J., Gulati, M., Crosier, R., Allana, S., Randhawa, V. K., ... & Norris, C. M. (2024). The Canadian Women's Heart Health Alliance ATLAS on the epidemiology, diagnosis, and management of cardiovascular disease in women—chapter 9: summary of current status, challenges, opportunities, and recommendations. *CJC Open*, *6*(2), 258-278.

National Center for Advancing Translational Sciences (NCATS). (n.d.). “Subgroup” Glossary Index. U.S. Department of Health and Human Services, National Institutes of Health. https://toolkit.ncats.nih.gov/glossary/subgroup/#:~:text=Subgroup%20refers%20to%20a%20subset,Patient%2DFocused%20Drug%20Development%20Glossary

Paulus, J. K., & Kent, D. M. (2020). Predictably unequal: understanding and addressing concerns that algorithmic clinical prediction may increase health disparities. NPJ Digital Medicine, 3(1), 99.

Public Health Agency of Canada. (2018, November 16). Surveillance of heart diseases and conditions. The Government of Canada. https://www.canada.ca/en/public-health/services/diseases/heart-health/heart-diseases-conditions/surveillance-heart-diseases-conditions.html

Rawat, S. (2019). Heart Disease Prediction. GitHub. Retrieved from https://github.com/ShubhankarRawat/Heart-Disease-Prediction

Rawat, S. (2019, August 10). Heart Disease Prediction. Towards Data Science. Retrieved from https://towardsdatascience.com/heart-disease-prediction-73468d630cfc/

Sax, D. R., Mark, D. G., Rana, J. S., Huang, J., Casey, S. D., Norris, R. P., ... & Reed, M. E. (2024). Pilot trial of an electronic decision support to improve care for emergency department patients with acute heart failure. ESC Heart Failure, 11(6), 4432-4436.

Vasey, B., Ursprung, S., Beddoe, B., Taylor, E. H., Marlow, N., Bilbro, N., ... & McCulloch, P. (2021). Association of clinician diagnostic performance with machine learning–based decision support systems: a systematic review. JAMA Network Open, 4(3), e211276-e211276.

Appendices

Appendix A: Key Stakeholder Value Mapping

Table A1 notes all key stakeholders identified in this audit, in addition to the values each has, and how they further align with

Table A1: Key stakeholders, their values, and aligned audit areas related to the implementation and use of TriageAssist. The table highlights the strategic alignment between stakeholder priorities and audit areas to ensure ethical, transparent, and efficient system deployment.

Stakeholders and Roles	Values	Aligned Audit Areas & Related Impacts
Provider: Prioritize™ Roles: Developer and Deployer	A. Business Success and Profitability B. Compliance and Safety in Clinical Deployment C. Workflow Integration and System Optimization D. Public Trust and Responsible AI Leadership	Since the audit is focused on addressing Prioritize™'s concerns and areas of opportunity, all of the audit areas are aligned with their values. These are some examples: A. **Strategic Partnerships & Funding:** Facilitate effective collaboration and implementation across hospital networks, and increase eligibility for funding, awards, and grants. B. **Ethical Safeguards & Regulatory Preparedness:** Ensure the well-being of all patient demographics, eliminate liability risks, and prepare for market expansion requiring regulatory compliance. C. **Product Usage & Deployment Practices:** Identify appropriate product usage patterns and deliver effective instructional materials and training supports. D. **Reputational Risk Management:** Address current market mistrust and prevent negative publicity and reputational damage.
Client: Hospitals Role: Institutional Users of Product	A. Reducing Wait Times for High-Risk Patients B. Optimizing Clinical Resources C. Improving Health Outcomes Equitably D. Supporting Staff with Augmentative Tools	A. **High-Risk Identification:** Ensure the well-being of high-risk patient demographics. B. **Product Deployment & Workflow Logistics:** Facilitate effective collaboration and implementation across hospital networks. C. **Ethical and Legal Safeguards:** Eliminate liability risks and enhance trust. D. **Training and Trust:** Deliver effective instructional materials and training supports.

Users: Nurses and Doctors **Role: Frontline Users of Product**	A. Rapid and Accurate Assessments B. Patient Safety and Professional Integrity C. Clear, Interpretable Recommendations D. Workflow Compatibility in High-Stress Environments E. Fairness and Empathy in Care Delivery	A. **Output Interpretation & Usage Guidance:** Provide clear and accurate usage instructions. B. **Ethical Responsibility & Harm Minimization:** Ensure ethical usage and minimize harm. C. **Transparency and Trust:** Enhance patient and clinician understanding and confidence. D. **Usage Logistics:** Streamline deployment and integration processes. E. **Equity Messaging:** Promote fairness and transparency in healthcare delivery.
Patients **Role: End Beneficiaries**	A. Timely and accurate care B. Human touch C. Clear communication	A. **Ensure the well-being of all patient demographics:** Deliver timely and accurate care. B. **Improve patient experience:** Maintain a human-centered approach. C. **Enhance transparency:** Facilitate clear and understandable communication.
Government/Health Regulators **Role: Oversight Authorities**	A. Public health protection B. Data privacy C. Fairness	A. **Compliance:** Adhere to public health standards and regulations. B. **Safety:** Protect patient data and ensure secure information handling. C. **Equity:** Promote fairness in healthcare decision-making and access.

Appendix B: Code Repository Access Information

This repository presents the results of an AI ethics audit of TriageAssist, a clinical decision support system developed by Prioritize™, designed to assist emergency room clinicians in triaging cardiovascular disease (CVD) cases. It includes the audit code file, the dataset used for all dataset analyses and testing, the use case brief and documentation, and this audit report.

The audit was conducted by Vantamed+, an independent AI ethics consultancy. Please access the Readme.me file provided for an explanatory overview of all repository contents. For accessibility purposes, the code files are open-access and public.

Online GitHub Repository Link: https://github.com/ma-carrillo/FinalProject_ECSE681

Appendix C: Baseline Heart Disease Patient Profile

To compare patient values within the *"Adversarial Testing"* section, a baseline profile of the most typical heart disease patient was created. It follows the profile of a middle-aged male, with the full set of corresponding input feature values listed below in Table C1. To create this baseline, the mode (most frequent value) was identified for each feature among any patient diagnosed with heart disease (HeartDisease = 1).

Table C1: This profile represents the most typical heart disease patient, constructed using the mode of each input feature among patients diagnosed with heart disease (HeartDisease = 1), used for adversarial testing purposes.

Feature	Value
Age	58
Sex	Male
Chest Pain Type	Asymptomatic
Resting Blood Pressure	140.0
Cholesterol	282.0
Fasting Blood Sugar	0.0
Resting ECG	Normal
Max Heart Rate	140
Exercise Angina Present	Yes
Old Peak Value	0.0
ST Slope Profile	Flat