Machine Learning and Public Health: Identifying and Mitigating Algorithmic Bias through a Systematic Review

Sara Altamirano, Arjan Vreeken, Sennay Ghebreab
Informatics Institute, University of Amsterdam, The Netherlands
{s.e.altamirano, a.vreeken, s.ghebreab}@uva.nl

Abstract

Machine learning (ML) promises to revolutionize public health through improved surveillance, risk stratification, and resource allocation. However, without systematic attention to algorithmic bias, ML may inadvertently reinforce existing health disparities. We present a systematic literature review of algorithmic bias identification, discussion, and reporting in Dutch public health ML research from 2021 to 2025. To this end, we developed the Risk of Algorithmic Bias Assessment Tool (RABAT) by integrating elements from established frameworks (Cochrane Risk of Bias, PROBAST, Microsoft Responsible AI checklist) and applied it to 35 peer-reviewed studies. Our analysis reveals pervasive gaps: although data sampling and missing data practices are well documented, most studies omit explicit fairness framing, subgroup analyses, and transparent discussion of potential harms. In response, we introduce a four-stage fairness-oriented framework called ACAR (Awareness, Conceptualization, Application, Reporting), with guiding questions derived from our systematic literature review to help researchers address fairness across the ML lifecycle. We conclude with actionable recommendations for public health ML practitioners to consistently consider algorithmic bias and foster transparency, ensuring that algorithmic innovations advance health equity rather than undermine it.

1 Introduction

Machine Learning (ML) is transforming Public Health (PH) through accurate prediction, real-time monitoring, and data-driven decision-making (Mhasawade, Zhao, and Chunara 2021; Rajkomar et al. 2018; Wiemken and Kelley 2020; Benke and Benke 2018; Jiang et al. 2017). As PH systems adopt ML, these technologies become essential. Yet ethical governance lags behind. Despite progress in ML fairness, its application in PH—especially for vulnerable populations—remains limited. Without intervention, algorithmic bias (AB) can exacerbate health disparities, misallocate resources, and reinforce care barriers (Fletcher, Nakeshimana, and Olubeko 2021; Flores, Kim, and Young 2024; Giancesco et al. 2018). These risks place PH+ML research within the broader domain of Responsible Artificial Intelligence (AI). The challenge is sociotechnical: integrating algorithmic decision-making (ADM) into real-world settings where equity, transparency, and accountability are critical. In high-stakes fields like PH, where decisions affect entire populations, unreported or misunderstood bias carries significant risks. As ML adoption grows, it becomes urgent to assess how AB is embedded in PH+ML research and whether it is adequately addressed by the research team (Xu et al. 2022; Thomasian, Eickhoff, and Adashi 2021).

This study systematically investigates how AB is identified, discussed, and reported in Dutch PH+ML research. While ML increasingly informs PH, it remains unclear whether researchers recognize and communicate AB risks. We highlight reporting gaps and advocate for transparent, standardized, fairness-aware practices in a domain with direct population-level impact.

To this end, we conduct a systematic literature review (SLR) of Dutch PH+ML studies. We extract metadata, including ML tasks, datasets, and performance metrics. Using our Risk of Algorithmic Bias Assessment Tool (RABAT), we code how AB risks, subgroup vulnerabilities, and fairness considerations are addressed—or not. Alongside this retrospective assessment, we introduce the ACAR framework (Awareness, Conceptualization, Application, Reporting): a forward-looking guide to help PH+ML researchers assess and address AB across the research lifecycle. ACAR translates SLR insights into guiding questions that embed fairness and accountability from conception to reporting, enabling ethical, context-sensitive system design.

The Netherlands is a highly relevant setting, with strong PH and ML innovation supported by advanced health infrastructure and digitalization. Despite universal healthcare, disparities persist—especially among ethnic minorities and socioeconomically disadvantaged groups, who face care barriers and higher burdens of chronic and psychiatric conditions (Ikram et al. 2014; Ilozumba et al. 2022; Kroneman et al. 2016). While focused on Dutch PH+ML, this study offers globally relevant insights for countries seeking to ensure ML promotes, rather than undermines, health fairness and ethical accountability.

In summary, this study advances Responsible AI by (1) conducting an SLR of AB reporting in Dutch PH+ML research, (2) developing and applying RABAT to identify key gaps, and (3) introducing ACAR as a practical fairness framework. These contributions show that addressing ethical concerns in PH+ML requires more than technical fixes;

it demands rethinking how we evaluate, design, and report algorithms for the public good.

2 Background

2.1 Key Definitions

ADM refers to computational systems, often ML-powered, that support or automate decisions with real-world consequences (Burrell 2016; Mittelstadt et al. 2016). These systems are increasingly used in high-stakes domains such as credit scoring (Altman and Saunders 1998), stock market prediction (Patel et al. 2015), educational risk alerts (Arnold and Pistilli 2012), recidivism prediction (Angwin et al. 2016), digital public services (Margetts and Naumann 2017), and PH surveillance (Ginsberg et al. 2009; Hillebrand et al. 2020). While enabling large-scale pattern detection and efficiency gains, ADM also risks amplifying structural disparities. For instance, models may rely on proxy variables like ZIP codes that correlate with race, producing biased outcomes, and their opacity can obscure accountability (Barocas and Selbst 2016; Burrell 2016). These risks are well documented, particularly in public sector ADM, where opaque systems have exacerbated existing inequalities (Eubanks 2018).

One of the main risks of ADM systems is AB: systematic patterns in data, model design, or deployment that yield unfair or discriminatory outcomes, often disadvantaging particular individuals or groups. It arises from sources such as sampling gaps, flawed feature selection, and inherited historical inequities (Mehrabi et al. 2021; Suresh and Guttag 2019). Documented cases include racial bias in facial recognition systems (Buolamwini and Gebru 2018), recidivism risk scores (Angwin et al. 2016), clinical prediction tools based on healthcare costs (Obermeyer et al. 2019), and stereotyping in GenAI image generation (Ferrara 2024). AB may be passive—inherited from data—or active, introduced during design. Detecting it can reveal structural inequities and support mitigation. Risks include misallocation of resources, flawed predictions, and harm to marginalized groups. AB reflects broader structural injustices and calls for responses that combine technical mitigation with societal interventions.

In contrast, algorithmic fairness (AF) refers to the principles, definitions, and interventions aimed at preventing or mitigating the harms of AB. AF encompasses both individual and group level objectives grounded in social and ethical judgments about justice and discrimination (Barocas, Hardt, and Narayanan 2023; Mehrabi et al. 2021). AF interventions range from reactive audits to proactive fairness-aware optimization, and their application must be context-sensitive to align with domain-specific values (Caton and Haas 2024; Binns 2018; Selbst et al. 2019). Formal AF definitions include individual fairness (Dwork et al. 2012), statistical parity (Feldman et al. 2015), equalized odds (Hardt, Price, and Srebro 2016), predictive parity (Chouldechova 2017), and counterfactual fairness (Kusner et al. 2017). These metrics often involve trade-offs and require careful consideration within each application context (Verma and Rubin 2018). Moreover, effective AF can enhance transparency, reduce disparities, and foster public trust (Holstein et al. 2019). However, over-reliance on a single metric risks obscuring deeper systemic inequities (Corbett-Davies et al. 2023) or enabling "fairwashing"—superficial fairness claims without substantive change (Aïvodji et al. 2019). Meaningful AF demands stakeholder engagement, institutional commitment, and ongoing evaluation, especially in PH, where fairness challenges are complex and multifaceted.

2.2 Reporting Bias in PH+ML Research

Guidelines for reporting AB in PH+ML remain limited. While Thomasian, Eickhoff, and Adashi (2021) outline pipeline bias strategies specific to PH data systems, most available frameworks are adapted from clinical or general ML contexts. Vollmer et al. (2020) propose 20 critical questions via the TREE checklist for clinical ML, and Rajkomar et al. (2018) offer a widely cited fairness checklist rooted in distributive justice principles. Fletcher, Nakeshimana, and Olubeko (2021) contribute three structured criteria—appropriateness, fairness, and bias—for evaluating ML systems in global health.

Documentation tools such as Model Cards and Datasheets promote transparency but lack PH+ML-specific focus (Mitchell et al. 2019; Gebru et al. 2021). Healthsheets (Rostamzadeh et al. 2022) adapt datasheets to healthcare, using expert input to foreground dataset-level bias. BEAMRAD (Galanty et al. 2024) evaluates transparency in medical imaging and signal datasets, linking poor documentation to downstream risks. However, these tools rarely address subgroup harms or equity challenges specific to PH contexts.

Several EQUATOR Network frameworks aim to improve reporting quality in biomedical research. TRIPOD+AI and PROBAST+AI offer detailed guidance on reporting and assessing risk of bias and applicability in clinical prediction models. However, they provide only high-level recommendations regarding subgroup analyses and do not operationalize fairness evaluations or subgroup-specific metrics (Collins et al. 2024; Moons et al. 2025). STARD-AI, CLAIM, CONSORT-AI, and SPIRIT-AI target diagnostic or interventional studies but address AB in population-level ML (Ibrahim et al. 2021; Mongan, Moy, and Kahn 2020; Liu et al. 2020). Other EQUATOR tools—PRISMA (Page et al. 2021), SPIRIT (Chan et al. 2013), STROBE (von Elm et al. 2007), CARE (Gagnier et al. 2013), COREQ (Tong, Sainsbury, and Craig 2007), SRQR (O'Brien et al. 2014), STARD (Bossuyt et al. 2015), TRIPOD (Collins et al. 2015), CHEERS (Husereau et al. 2013), and ARRIVE (Kilkenny et al. 2010)—support transparency but overlook fairness considerations, subgroup calibration, or bias mitigation.

Beyond clinical guidelines, frameworks like FUTURE-AI (Lekadir et al. 2025), IEEE 7003 (IEEE 2025), and WHO/UNESCO ethics guidance (World Health Organization 2021) emphasize responsible AI development, but remain high-level and lack actionable tools for AB reporting in PH+ML research.

Open-source toolkits such as AIF360 (Bellamy et al. 2019), Fairlearn (Bird et al. 2020), Themis (Galhotra, Brun, and Meliou 2017), fairmodels (Wiśniewski and Biecek 2022), the What-If Tool (Wexler et al. 2020), FairTest

(Tramèr et al. 2017), and Aequitas (Saleiro et al. 2018) provide general fairness metrics and mitigation techniques. While Aequitas includes a "fairness tree" adapted for population-level settings, none of these tools support PH-specific bias detection or subgroup-sensitive calibration.

In sum, existing reporting frameworks promote general transparency but lack mechanisms for identifying, measuring, or mitigating AB in PH+ML research.

2.3 Dutch Context and Governance Landscape

The Netherlands is a highly relevant setting for studying AB in PH+ML research, with advanced data systems, high Electronic Health Records (EHR) coverage, and a diverse population (Kroneman et al. 2016). ML is increasingly used in Dutch public institutions, including PH, supporting ADM across policy domains. However, structural disparities persist. Ethnic minorities and disadvantaged groups face higher burdens of chronic disease and barriers to mental health care, while migrant populations report unmet cultural and linguistic needs (Ikram et al. 2014; Ilozumba et al. 2022; Teunissen et al. 2015). Without explicit AB attention, ML risks amplifying these inequities.

Recent governance reforms reflect growing awareness, regulatory action, and progress in technical tools for responsible AI. In 2022, Statistics Netherlands (CBS) replaced its binary "Western/non-Western" classification with more nuanced categories, reshaping how sensitive attributes are defined and operationalized (Centraal Bureau voor de Statistiek (CBS) 2022). The Netherlands also maintains a national Algorithm Register, which documents deployed ADM systems in domains such as youth care, benefits, and PH surveillance (Government of the Netherlands 2024). While ML increasingly informs PH decision-making, systematic AB reporting remains rare. One exception is Holstege et al. (2025), who audited bias in an administrative risk profiling system. Although Dutch and EU governance frameworks emphasize transparency, they do not mandate AB reporting in PH+ML research. The Dutch AI Impact Assessment and EU AI Act prioritize high-risk domains such as law enforcement, employment, and healthcare, with limited attention to PH or AB reporting (Ministry of Infrastructure and Water Management 2024; European Parliament and Council of the European Union 2024). The Dutch Vision on Generative AI broadly mentions health and fundamental rights but does not address PH or AB directly (Ministry of the Interior and Kingdom Relations 2024).

2.4 Positioning This Study

A growing body of literature has explored AB and AF in PH. Mhasawade, Zhao, and Chunara (2021) identified gaps in fairness-aware ML and emphasized the need to incorporate structural determinants alongside individual factors. Delgado et al. (2022) reviewed COVID-19 AI systems, highlighting recurring biases in data representativeness, demographic omission, and validation practices. Other reviews propose PH-relevant AF frameworks: Sikstrom et al. (2022) conceptualize fairness as a multidimensional construct, while Chin et al. (2023) centers health and health care equity for patients and communities as the goal, specifically within the wider context of structural racism and discrimination. Char, Abràmoff, and Feudtner (2020) offer a stage-based ethical reflection framework for healthcare ML, outlining questions across development, implementation, and oversight—though not tailored to PH concerns. Morgenstern et al. (2020) conducted a scoping review of 231 ML prediction studies in PH, noting transparency and calibration gaps but not assessing AB or subgroup fairness.

Empirical studies further show how design and deployment choices can reinforce inequities. Obermeyer et al. (2019) found that using healthcare cost as a proxy for illness systematically underestimated patient needs. Tsai et al. (2022) and Flores, Kim, and Young (2024) identified structural and data-related fairness gaps in PH forecasting and surveillance.

Unlike conceptual or secondary reviews, this study systematically examines how AB risks are identified, discussed, and reported in Dutch PH+ML research. Based on the findings, we present the ACAR framework to translate reporting gaps into actionable fairness-aware guidance for PH+ML.

3 Methods

3.1 Study Design

This study addresses the growing use of ML in PH research, the risks of AB to vulnerable populations, the limited integration of AF, and the Netherlands' position as a scientifically advanced yet unequal health system. We conducted an SLR to analyze trends in AB reporting and AF consideration in Dutch PH+ML studies. Following interdisciplinary SLR guidelines (Carrera-Rivera et al. 2022), the review was structured for rigor, replicability, and comprehensiveness. Our goal was to examine how AB risk is addressed across the research process, focusing on impacts on minorities and disadvantaged groups. Our main research question was: To what extent is AB risk identified, discussed, and reported in Dutch PH+ML research from 2021–2025?

The review targeted three categories—*extent of bias discussion*, *identification of subgroups at risk*, and *reporting transparency*—assessed via RABAT. Scope was defined using the PICOC framework (Carrera-Rivera et al. 2022): Population = Dutch PH research; Intervention = ML methods; Comparison = not applicable; Outcome = AB identification, discussion, and reporting across the three categories; Context = peer-reviewed Dutch PH+ML studies published between 2021–2025. Beyond the SLR, we introduce the ACAR framework to guide AF integration across the PH+ML research lifecycle.

3.2 Inclusion and Exclusion Criteria

Studies were included if they met all of the following:

- **Topic Relevance**: Addressed PH topics such as population health, disease prevention, epidemiology, surveillance, or population-level interventions.
- **ML Application**: Trained ML models for prediction, classification, or related tasks.
- **Dutch Context**: Used Dutch data or included at least one author (first, second, or last) affiliated with a Netherlands-based institution.

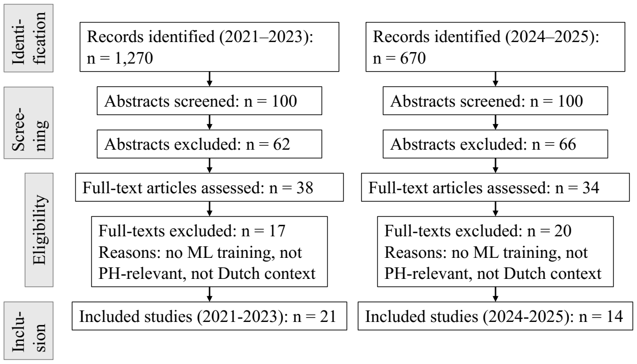

Figure 1: PRISMA flow diagram of study selection process.

- **Publication Type**: Peer-reviewed articles only.
- **Language**: English.
- **Publication Date**: January 2021 — February 2025.

Studies were excluded if any of the following applied:

- **Clinical/Operational Focus**: Centered on clinical decision-making, biomedical mechanisms, or healthcare management (e.g., hospital operations, cost optimization) without clear PH relevance.
- **No Valid ML Use**: Misused ML terms (e.g., called regression or feature selection ML) or did not train models.
- **Technology-Only Scope**: Focused solely on technical development unrelated to PH outcomes (e.g., signal processing, device engineering).

When PH relevance was unclear, full-text screening assessed if the study's objectives, population, outcomes, or implications supported population-level insights, policy decisions, or PH service delivery. Studies were included if they met our operational definition of PH+ML: ML methods relevant to population-level outcomes or potential to inform PH policy, surveillance, or interventions. The inclusion process is summarized in the PRISMA diagram (Figure 1).

3.3 Search Strategy and Screening Process

This review initially focused on AB in PH+ML research by GGD Amsterdam, the city's municipal health service, given its policy role, robust data infrastructure, and diverse population. By 2030, an estimated 55.7% of Amsterdam's residents will have a non-Dutch background, including sizable Surinamese, Turkish, Moroccan, and Indonesian communities. These groups bear disproportionately high burdens of chronic diseases—especially diabetes, cardiovascular conditions, and mental health disorders—compared to ethnic Dutch populations (Ikram et al. 2014). Combined with suboptimal cultural responsiveness in mental healthcare, Amsterdam is a salient case for AB in ADM (Ilozumba et al. 2022). Its diversity and PH infrastructure provide a strong context to examine how ML interacts with population heterogeneity and equity outcomes (Essink-Bot et al. 2013).

A manual review of GGD Amsterdam's archive in December 2023 for peer-reviewed studies (2021–2023) mentioning ML found no eligible publications. Follow-up keyword searches confirmed the absence of ML or AI usage, prompting a shift to a national-level review of Dutch PH+ML research.

Given the topic's interdisciplinary nature, with studies scattered across medical, epidemiological, and technical venues not consistently indexed by PubMed or Scopus, Google Scholar was selected. Its broad coverage and relevance-based sorting were well suited to capture cross-disciplinary work. Since ML and PH are defined inconsistently across fields, this approach helped capture studies otherwise missed due to terminological or disciplinary silos.

2021–2023 Search and Screening The structured search was conducted in December 2023 using the query:

```
("machine learning" AND "Public
Health" AND "Netherlands") AND
("Amsterdam UMC" OR RIVM OR "Public
Health Service")
```

The search returned 1,270 results, sorted by relevance. Abstract-level screening was extended to the first 200 records; however, no additional eligible studies were identified in abstracts 101–200, limiting our review to the top 100 abstracts. Screening followed a two-stage protocol: (1) abstract-level screening against inclusion criteria, and (2) full-text review for shortlisted studies. Ultimately, 21 studies published between 2021 and 2023 met all inclusion criteria.

2024–2025 Updated Search and Screening To assess whether AF and AB considerations in Dutch PH+ML research have progressed, particularly amid the rise of generative AI tools such as large language models (LLMs), a follow-up search was conducted in February 2025. The same query and screening criteria were used to ensure consistency with the 2021–2023 review.

The search returned 670 results. Again, the first 100 abstracts were screened, with shortlisted studies reviewed in full. Fourteen studies met all inclusion criteria and were added to the sample.

In both search phases, studies were excluded if they lacked model training—such as those using ML methods only for feature selection, post hoc analyses, or variable importance without testing on unseen data. Also excluded were studies proposing ML methods without implementation, protocol studies with unrealized ML components, gray literature, theses, reviews, and non-peer-reviewed documents to ensure inclusion of genuine PH+ML applications.

All records were tracked in a structured spreadsheet, with detailed documentation of inclusion decisions.

3.4 RABAT: Risk of Algorithmic Bias Assessment Tool

To assess how AB is identified, discussed, and reported in Dutch PH+ML research, we developed RABAT, which selectively adapts elements from three frameworks. These were chosen for their credibility and complementary focus on people, processes, data, and methods, all of which are critical for an integral assessment of AB in PH+ML research (see Appendix Table A.5 in the extended/supplementary version). RABAT adds novelty through its focus

Q#	Condensed RABAT Questionnaire
Q1	Is data bias discussed (e.g., representativeness)?
Q2	Is model bias discussed (e.g., performance disparities)?
Q3	Is bias framed in ML fairness terms?
Q4	Are potential societal impacts discussed?
Q5	Are at-risk subgroups identified?
Q6	Are data sampling, inclusions, and exclusions described?
Q7	Are sensitive attributes reported?
Q8	Is bias described in sufficient length and structure?
Q9	Is mitigation of fairness-related harms reported?
Q10	Is meaningful informed consent discussed?

Table 1: Condensed version of the RABAT questionnaire.

on PH+ML, structured scoring logic, and emphasis on AB and subgroup risks. First, the Cochrane Risk of Bias Tool identifies systematic errors in randomized trials, assessing domains like selection, performance, detection, and reporting bias—essential for evaluating study validity and transparency in health research (Higgins et al. 2011). Second, PROBAST (2019) provides structured criteria for assessing bias in prediction model studies, focusing on how predictors are defined, selected, and handled, which are key considerations in ML-based health modeling (Moons et al. 2019; Wolff et al. 2019). Third, Microsoft's Responsible AI (MS RAI) checklist emphasizes reflecting on societal impacts, fairness framing, and stakeholder harms throughout the ML lifecycle (Madaio et al. 2020). We derived three categories from this integration: *extent of bias discussion*, *identification of subgroups at risk*, and *reporting transparency*. These categories reflect a PH perspective by placing people at the center of bias discussion, subgroup risk, and research reporting. For example, the societal impact question (Q4) draws from MS RAI checklist's prompts about the system's role, affected stakeholders, and benefit–harm trade-offs. The ten resulting questions reflect our guiding research question. A condensed version of RABAT appears in Table 1; full item wording and grading criteria are available in the extended version (Appendix Tables A.1–A.4).

Scoring and Evaluation Process. Two reviewers independently applied RABAT to all studies from the 2021–2023 and 2024–2025 review phases. Prior to full review, the rubric was pilot-tested on a sample of papers to refine definitions and eliminate ambiguity. Discrepancies were discussed to finalize the scoring guide. Each RABAT item was scored on a four-point scale: 0 = absent, 1 = minimal, 2 = moderate, and 3 = extensive; higher scores indicate stronger performance on AB-related reporting. In addition, mean RABAT scores were classified into *Low Risk*, *Some Concerns*, or *High Risk* using calibrated thresholds (<0.75 = High Risk; 0.75–1.5 = Some Concerns; >1.5 = Low Risk), adapted from PROBAST principles and adjusted for the observed left-skewed score distribution. Percentage agreement between reviewers was calculated per RABAT item (excluding `NA`). Agreement ranged from 71.4% to 88.6%. Highest rates were for *ML fairness* (Q3, 88.6%), *sensitive attributes* (Q7), and *harm transparency* (Q9), both 85.7%. *Societal impact* (Q4) reached 80.0%, while *model bias* (Q2), *subgroups at risk* (Q5), and *informed consent* (Q10) were near 77%. *Sampling and missing data* (Q6) had 74.3%; the lowest agreement was for *data bias* (Q1) and *bias articulation* (Q8), both 71.4%. These results reflect moderate to substantial agreement, strongest for clearly defined items. RABAT also served as a quality appraisal tool, replacing standard SLR checklists (Carrera-Rivera et al. 2022), assessing presence and depth of AB reporting to evaluate transparency and rigor. Scores were logged in a spreadsheet; full rubric and examples are available on request.

3.5 ACAR: Awareness, Conceptualization, Application, and Reporting

To synthesize our findings and support practical translation, we developed a four-stage framework adapted from Design Thinking. Design Thinking offers a flexible, problem-oriented approach widely used in applied fields to guide iterative learning, user-centered design, and structured reflection (Brown 2009). Though rooted in design and engineering, its focus on awareness, iterative conceptualization, stakeholder relevance, and transparent problem-solving aligns with the needs of AB-aware ML. While general frameworks offer lifecycle guidance, ACAR adapts these principles to the PH+ML domain, tailoring its stages to domain-specific concerns such as subgroup relevance, societal impact, and context-aware reporting. It emphasizes AB identification, fairness conceptualization, method application, and transparent reporting as sequential stages in the research workflow (see extended version, Appendix Table A.6). We define each of the four ACAR stages as follows:

1. **Awareness:** Recognize that AB may emerge from data, model design, or social context. Reflect on whether fairness is relevant, who might be affected, and how broader societal impacts could arise.
2. **Conceptualization:** Define AB, fairness, and subgroup risks in relation to research objectives and methodology. Frame these concepts early to establish a clear fairness lens.
3. **Application:** Implement strategies to address AB in data and modeling workflows, including sampling decisions, subgroup testing, and bias mitigation techniques, whether experimental or in production.
4. **Reporting:** Clearly communicate how AB risks and fairness considerations were addressed, including structured bias discussions, subgroup findings, transparency about limitations, and ethical elements like consent.

4 Results

4.1 Overview of Included Studies

This review includes 35 peer-reviewed PH+ML articles published between 2021 and 2025 that met all inclusion criteria (see Appendix Table A.7, extended version). The sample comprised nine studies from 2021, seven from 2022, five from 2023, twelve from 2024, and two from early 2025, reflecting a sharp increase in publications from 2024. Covered domains included infectious disease surveillance, mental health, behavioral prediction, environmental exposure, and chronic disease modeling.

Most studies applied classification (n = 21) or regression (n = 9); a few used time-series forecasting (n = 3) or survival analysis (n = 1). Common methods included Random Forest, Logistic Regression, and XGBoost, often combined with Support Vector Machines, LASSO, or deep learning models. Ensemble and hybrid approaches were also used. Performance metrics were consistently reported, with frequent use of area under the curve (AUC), sensitivity, specificity, accuracy, precision, and F1 score; fewer studies included calibration, confidence intervals, or explainability metrics such as SHapley Additive exPlanations (SHAP).

Dataset sizes ranged from small (e.g., 134 proteomics patients, 253 helpline transcripts) to large cohorts such as the Public Health Monitor (PHM) (244,557 individuals), national emergency department (ED) admissions for acute coronary syndrome (214,953 patients, 2010–2017), and the STIZON registry (over one million primary care records curated for research use). All studies used Dutch data or involved Dutch-affiliated authors, often through cross-sector collaborations spanning public health, clinical, and academic institutions.

Most studies addressed ethical or regulatory compliance, citing approvals from Dutch Medical Ethics Committees, GDPR waivers, or the Declaration of Helsinki. A minority lacked formal ethics or consent statements. Human oversight was typically reported through expert input or annotation. Real-world deployment was rare: most models were retrospective, with only two used in clinical or PH systems. Several proposed deployment but remained in development. The split sample design showed fairness practices were largely absent across both periods reviewed.

As described in Section 3, all studies were evaluated using the ten-item RABAT framework, with each item scored from 0 (absent) to 3 (extensive). The following subsections present findings by category.

RABAT Question	Mean Score	Risk Level
Q1. Data bias	0.74	High Risk
Q2. Model bias	0.54	High Risk
Q3. ML fairness	0.06	High Risk
Q4. Societal impact	0.59	High Risk
Q5. Subgroups at risk	0.37	High Risk
Q6. Sampling & missing data	1.64	Low Risk
Q7. Sensitive attributes	0.07	High Risk
Q8. Bias articulation	0.60	High Risk
Q9. Harm transparency	0.13	High Risk
Q10. Informed consent	0.96	Some Concerns

Table 2: Mean RABAT scores classified as Low Risk, Some Concerns, or High Risk using calibrated thresholds (<0.75, 0.75–1.5, >1.5) adapted for left-skewed distributions.

4.2 Category 1: Extent of Bias Discussion

Q1: Data Bias The median score was 1.0 (IQR: 0.5–1.0), with a mean of 0.74 (SD = 0.46), classified as *High Risk*. Six studies (17%) did not mention data bias; 27 (77%) engaged only minimally. Two studies (6%) reached a moderate level; none were rated extensive. Most discussions were indirect—referring to class imbalance, missing data, or non-representativeness—without explicitly framing these as sources of bias. Common issues included subgroup overrepresentation, incomplete linkage, or limited access to key variables (e.g., age, sex, socioeconomic status). Some papers addressed data quality, stratification, or exclusions, but rarely linked these to biased outcomes. Only one study described a mitigation strategy; none conducted structured assessments. Mentions of provenance, measurement error, or distributional skew were largely absent.

Q2: Model Bias The median score was 0.5 (IQR: 0.0–1.0), with a mean of 0.54 (SD = 0.48), classified as *High Risk*. Thirteen studies (37%) did not address model bias; 21 (60%) engaged only minimally, typically noting overfitting, performance variation, or majority-class favoring. One study (3%) scored moderate; none were extensive. Discussions were brief and framed as technical concerns. Some cited uncertainty from small datasets, regional misclassification, or modest predictive performance, but without linking these to structural risks. A few mentioned subgroup underperformance or the need for validation, but none reported disaggregated error, architectural bias, or AB mitigation efforts. No study conducted formal audits or addressed AF risks in ADM.

Q3: ML Fairness The median score was 0.0 (IQR: 0.0–0.0), with a mean of 0.06 (SD = 0.16), classified as *High Risk*. Thirty-one studies (89%) did not mention fairness; the remaining four (11%) engaged only minimally. None were rated moderate or extensive. Fairness was largely absent as an analytic or ethical concept. Indirect mentions referred to generalizability, overfitting, or geographic variation, without linking to protected attributes or subgroup harm. No study defined fairness, used fairness metrics, or applied fairness-aware techniques. Even when model bias was noted, its potential impact on specific populations was not explored. ML fairness remained outside the conceptual and methodological scope of nearly all studies.

Q4: Societal Impact The median score was 0.5 (IQR: 0.0–1.0), with a mean of 0.59 (SD = 0.51), classified as *High Risk*. Twelve studies (34%) did not mention societal impacts; 22 (63%) engaged only minimally, typically referencing clinical utility, PH relevance, or policy applications. One study (3%) scored moderate; none were extensive. When discussed, societal impacts centered on potential benefits (e.g., screening, efficiency, decision support), while harms, trade-offs, and stakeholder-specific effects were rarely addressed. Some noted implications for workflows or planning, but equity, access, and unintended consequences were seldom considered. No study examined differential impacts on vulnerable groups or broader ethical dimensions of ADM.

4.3 Category 2: Identification of Subgroups at Risk

Q5: Subgroups at Risk The median score was 0.0 (IQR: 0.0–0.5), with a mean of 0.37 (SD = 0.52), classified as *High Risk*. Twenty studies (57%) did not identify any at-risk subgroups; thirteen (37%) engaged only minimally, typically referencing broad demographic categories (e.g., age, sex, comorbidities) without analyzing differential model performance. Two studies (6%) scored moderate; none were extensive. Mentions of subgroup risks were often indirect or speculative, such as noting exclusion of severely affected patients, limited data for specific groups (e.g., immunocompromised individuals), or demographic imbalances. Several studies listed protected attributes or used stratified analyses, but without linking these to fairness. No study systematically assessed whether ML models produced disparate outcomes across subpopulations or investigated sources of differential error or harm.

Q6: Sampling and Missing Data This item received the highest scores across all RABAT questions, with a median of 2.0 (IQR: 1.0–2.0) and a mean of 1.64 (SD = 0.80), classified as *Low Risk*. Three studies (9%) included no relevant content; nine (26%) provided only minimal information. The remaining 23 (66%) described sampling and missing data handling at moderate or extensive levels. Most clearly identified datasets and eligibility criteria, often reporting variable distributions by sex, age, or location. Several justified exclusions or discussed implications of missing data. Common imputation methods included median substitution and multiple imputation (e.g., MICE), though few assessed their impact on model performance. Fairness considerations, such as whether missingness or exclusions disproportionately affected specific groups, were absent. Even in detailed methodological reporting, representation bias and subgroup exclusions were not discussed in fairness terms.

Q7: Sensitive Attributes The median score was 0.0 (IQR: 0.0–0.0), with a mean of 0.07 (SD = 0.18), classified as *High Risk*. Thirty studies (86%) did not mention sensitive or fairness-relevant attributes such as race, ethnicity, disability, or migration background. Five (14%) engaged only minimally, typically listing variables like sex or age in tables or covariates, without analyzing model output variation across groups. No studies scored moderate or extensive. While many studies collected sociodemographic data or stratified by sex, none assessed disproportionate model errors, subgroup representation, or selection risks. As with Q6, no study problematized links between sensitive attributes and potential harms from ADM.

4.4 Category 3: Reporting Transparency

Q8: Bias Articulation The median score was 0.5 (IQR: 0.0–1.0), with a mean of 0.60 (SD = 0.48), classified as *High Risk*. Eleven studies (31%) did not discuss bias risks; twenty-two (63%) engaged only minimally, typically referencing imbalanced data, generalizability, or limited data quality. Two studies (6%) provided a moderate discussion, though engagement was often scattered or implicit. Most framed bias in technical terms, such as overfitting, missing data, or performance degradation across sites. Few treated bias as a systemic risk requiring mitigation, and none offered a dedicated section or structured analysis. Limitations were often described as general uncertainty or data quality issues, not as fairness-relevant concerns. No study examined how AB might propagate through the pipeline or affect subgroups differentially.

Q9: Harm Transparency The median score was 0.0 (IQR: 0.0–0.0), with a mean of 0.13 (SD = 0.28), classified as *High Risk*. Twenty-eight studies (80%) made no mention of fairness-related harms or mitigation. Seven (20%) acknowledged bias mitigation techniques—such as SMOTE, class weighting, or exclusion of correlated variables—but did not assess whether these reduced fairness risks. No study reported how harms were identified, assessed, or mitigated across the ML lifecycle. Bias was typically framed as a technical challenge affecting performance, not a source of downstream impact on populations. Even in PH or clinical settings, subgroup harms were not considered. None included audits, fairness checks, or mitigation justifications grounded in ethical or equity frameworks.

Q10: Informed Consent The median score was 1.0 (IQR: 0.0–1.875), with a mean of 0.96 (SD = 0.86), classified as *Some Concerns*. Eight studies (23%) did not mention informed consent. Ten studies (29%) met the minimal threshold, typically through brief statements of ethics approval or consent, or by citing waivers without elaboration on scope, process, or participant understanding. Six studies (17%) provided moderate reporting, including ethics approvals and confirmation of written or institutional consent, though none addressed risks of re-identification or model reuse. Only two studies (6%) scored extensive. Across the sample, informed consent was generally treated as a procedural requirement rather than an ethical concern, and no study engaged with the challenges of meaningful consent in ADM contexts.

4.5 Overall Scoring Summary

Of the ten RABAT items, only one—*Sampling and Missing Data* (Q6)—had a mean score above 1.0 and a median of 2.0 (IQR: 1.0–2.0), indicating generally moderate, occasionally extensive, reporting. In contrast, eight items had mean scores below 0.75 and were classified as *High Risk*, including three—*ML Fairness* (Q3), *Sensitive Attributes* (Q7), and *Harm Transparency* (Q9)—with means under 0.15, reflecting near-total absence of fairness-specific reporting. Figure 2 shows the score distributions. Items tied to conventional epidemiologic reporting, such as Q6 and *Informed Consent* (Q10), had higher medians and visibly wider score ranges. By contrast, fairness-oriented items like Q3, Q7, and Q9 were heavily skewed, with most studies scoring zero and minimal spread. *ML Fairness*, for example, received no scores above 1 (minimal level). These contrasts highlight a clear asymmetry between methodological rigor and fairness-aware reporting in PH+ML research.

Table 2 summarizes score distributions by question. Only Q6 was classified as *Low Risk*, with two-thirds of studies providing moderate-level reporting. Q10 fell in the *Some*

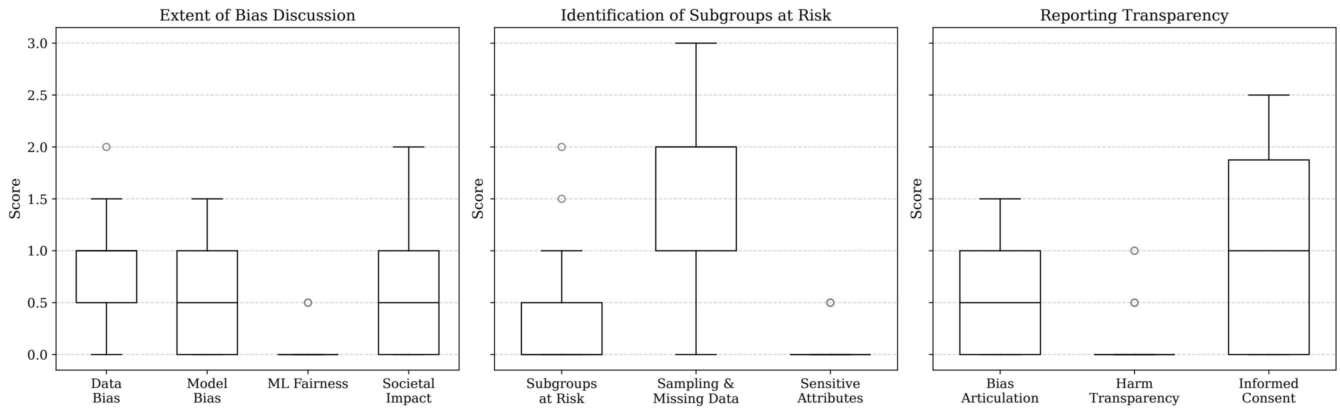

Figure 2: RABAT score distribution across ten questions, grouped by category. While some AB reporting aspects (e.g., *sampling and missing data*, *informed consent*) show moderate coverage, fairness-specific items (e.g., *ML fairness*, *sensitive attributes*, *harm transparency*) score consistently low, revealing critical gaps in Dutch PH+ML research.

Concerns range, with several studies reporting moderately but many offering limited detail. The remaining eight items were classified as *High Risk*, reflecting low mean scores and frequent absence of relevant content. Notably, the lowest mean scores were for fairness-specific items, indicating systemic neglect in this area.

4.6 ACAR Framework Adaptation and Guidance

To align evidence with practice, we mapped RABAT questions to the ACAR stages and analyzed associated patterns. In the *Awareness* stage, most studies mentioned data (Q1) or model limitations (Q2), but rarely linked these to structural bias or societal risks (Q4). Only one study discussed a mitigation strategy for data bias, and just one reached moderate-level discussion of societal impact, with none rated extensive. For *Conceptualization*, fairness was absent in 89% of studies (Q3), and subgroup vulnerabilities (Q5) were largely unaddressed, with only two studies scoring moderate. In the *Application* stage, while sampling and missing data (Q6) were consistently reported (66% moderate/extensive), sensitive attributes (Q7) and subgroup testing or mitigation (Q9) were almost entirely absent—both had mean scores below 0.15. Finally, in *Reporting*, some studies articulated bias (Q8) or mentioned consent (Q10), but few addressed fairness-related harms (Q9), and no study included structured audits or mitigation justifications. These gaps motivate our proposed guiding questions to support fairness-aware practice across the PH+ML workflow.

Key Gaps and Weaknesses The most pronounced deficiencies appeared at the *Conceptualization*, *Application*, and *Reporting* stages. Fairness was rarely defined (Q3), subgroup risks were largely unexamined (Q5), and no study applied fairness metrics (Q3) or conducted structured mitigation and transparent harm reporting (Q9). Sensitive attributes (Q7) were routinely omitted, and structured harm assessments were absent or superficial. Reporting focused primarily on technical limitations, with minimal articulation of AB (Q8) and limited detail on consent practices (Q10). These patterns reflect a pervasive lack of fairness-oriented engagement across the PH+ML workflow.

Guiding Questions for PH+ML Researchers To support structured reflection and fairness integration across the PH+ML research lifecycle, we propose a set of guiding questions organized by ACAR stage. Designed to complement the RABAT assessment, these prompts help translate fairness considerations into concrete actions across study design, model development, and reporting. The questions are general by design and intentionally crafted to be accessible to interdisciplinary teams, reflecting our finding that PH+ML research is often conducted by collaborators with diverse disciplinary backgrounds, and varying levels of technical expertise. Table 3 presents guiding questions to help researchers operationalize fairness in PH+ML.

5 Discussion

Understanding how AB risk is discussed and reported in PH+ML is essential for evaluating real-world harms and guiding equitable system design. Although ML adoption in PH is accelerating, the field lacks consistent standards for identifying which AB risks matter, to whom, and how they should be reported. This review addresses that gap by applying a structured lens to assess AB-related transparency and synthesizing reporting patterns in recent Dutch PH+ML studies. Gaps such as the lack of fairness metrics or subgroup error analysis may reflect limited methodological guidance, a focus on technical performance, or unfamiliarity with fairness-aware tools. To examine these patterns, we systematically evaluated 35 studies using RABAT, a ten-item tool assessing how AB risk is addressed in PH+ML research. Questions covered three areas: extent of bias discussion, identification of subgroups at risk, and reporting transparency, scored from 0 (absent) to 3 (extensive). This revealed consistent gaps—particularly in fairness framing, subgroup focus, and harm reporting—alongside variation in overall reporting quality. In contrast, data sampling and informed consent were more commonly addressed, likely

ACAR Stage	Guiding Questions
Awareness	**Bias Sources:** Have you identified potential bias from data, models, or population differences? **Subgroup Impact:** Could your model perform differently across population subgroups? **Equity Risks:** Could your model worsen existing health inequities or exclude vulnerable groups? **Transparency:** Are risks and societal impacts clearly described?
Conceptualization	**Fairness Framing:** Have you defined fairness in relation to your population and goals? **At-Risk Subgroups:** Have you identified groups that may be disproportionately affected? **Bias Mechanisms:** Have you considered how the model might produce disparities? **Subgroup Relevance:** Are subgroup definitions grounded in public health disparities?
Application	**Sampling Fairness:** Have you assessed whether missing data or exclusions affect certain groups? **Sensitive Attributes:** Do key traits like sex, age, or origin influence predictions? **Subgroup Testing:** Have you tested for unequal performance across key groups? **Mitigation Actions:** Have you applied and evaluated bias reduction strategies?
Reporting	**Bias Reporting:** Are bias sources and their subgroup effects clearly described? **Disparity Disclosure:** Have you reported group disparities and their implications? **Mitigation Reporting:** Have you reported bias mitigation actions and their outcomes? **Consent Clarity:** Is consent clearly reported, including data use and fairness risks?

Table 3: Guiding questions for AB-aware PH+ML research, labeled for integration across study design, modeling, and reporting.

reflecting established biomedical norms. These insights informed the development of the ACAR framework to support fairness integration in PH+ML workflows.

5.1 Observed Insights

Final RABAT scores revealed substantial variability in how studies identified, discussed, and reported the risk of AB. Figure 2 helps us visualize question-level score distributions, highlighting two distinct reporting patterns. First, items aligned with conventional epidemiological practices—such as *Sampling and Missing Data* (Q6) and *Informed Consent* (Q10)—showed higher median scores and greater dispersion across the scale, indicating both engagement and heterogeneity in reporting practices. In particular, Q6 displayed the highest central tendency (median = 2.0) and interquartile range, reflecting wide uptake of standard reporting norms. These patterns are echoed in Table 2, where only Q6 was classified as *Low Risk* and Q10 as *Some Concerns*. In contrast, fairness-specific items—including *ML Fairness* (Q3), *Sensitive Attributes* (Q7), and *Harm Transparency* (Q9)—were clustered near zero, with minimal variance, suggesting widespread omission rather than inconsistency. Correspondingly, these three items received mean scores below 0.15 and were classified as *High Risk*. Bias-related items such as *Data Bias* (Q1), *Model Bias* (Q2), and *Subgroups at Risk* (Q5) fell in between, with low to moderate engagement and scores concentrated around minimal levels. The shape and spread of these boxplots underscore a key observation: while epidemiologic-reporting rigor was often maintained, fairness considerations remained underdeveloped, sporadic, or absent altogether. Even when sensitive features were collected or subgroup labels applied, their implications for fairness or subgroup harm were seldom analyzed. No study applied fairness metrics or conducted disaggregated performance evaluations, underscoring the persistent absence of structural AB risk assessment in routine Dutch PH+ML reporting.

5.2 Actionable Recommendations for PH+ML Researchers

Derived from RABAT findings, these recommendations support consistent AB-aware practices in PH+ML research.

1. **State and Justify AB Risks.** Identify sources of data and model bias (e.g., overrepresentation, missingness) and explain their potential effects.
2. **Define Fairness Early.** Provide a study-relevant definition of AF (e.g., group or individual fairness) and specify its relevance.
3. **Identify Subgroups Proactively.** Note any populations likely to face higher AB risk, even if not analyzed separately.
4. **Test by Subgroup.** Disaggregate performance by subgroup when feasible, and report disparities clearly.
5. **Audit for Harms.** Assess or acknowledge potential downstream harms and unintended effects.
6. **Report Bias Transparently.** Include a dedicated section on AB sources, subgroup risks, and AF limitations.
7. **Use the ACAR Guide.** Refer to ACAR during planning and reporting to ensure AF is considered throughout.

5.3 Limitations

This review has some limitations. Although Dutch-language and institutional sources were manually searched, all included studies were from Google Scholar and English-language peer-reviewed sources, which may have limited the scope of the review, excluding Dutch-language, gray, or discipline-specific literature. The use of a single reviewer for initial screening may have introduced selection bias, though a structured two-stage protocol was followed. While piloting and inter-reviewer checks supported scoring consistency, RABAT's reliability, generalizability beyond this context, and usability by non-expert reviewers have not yet been assessed. Only two reviewers per paper conducted scoring, which may limit reproducibility, though reviewer alignment was essential due to the complexity and time demands of RABAT. As the review aimed to identify reporting patterns rather than test hypotheses, no statistical analyses were conducted. The ACAR framework also remains theoretical, as it has not yet been applied in real-world PH+ML settings. Finally, findings reflect the Dutch context and may, for example, not generalize to low-resource or alternative governance settings. Importantly, this review did not aim to exhaustively catalog all Dutch PH+ML studies. Nonetheless, the purposive sample of 35 studies was sufficient to reveal consistent reporting gaps, despite potential selection bias from manual screening. Through an SLR and structured RABAT assessment, we identified representative patterns to inform the ACAR framework—proposed here as a conceptual guide, not yet validated in real-world settings, and positioned as part of a broader mixed-methods strategy to embed fairness-aware practice in PH+ML research.

5.4 Closing the Gap

Our review confirms and extends prior findings on gaps in fairness-aware practices in PH+ML research. Earlier work has highlighted the limited use of AF metrics, subgroup evaluation, and harm transparency in these studies (Mhasawade, Zhao, and Chunara 2021; Delgado et al. 2022; Chen et al. 2024; Raza et al. 2024). These reviews also identify a deeper conceptual gap: AF is often reduced to technical performance, with limited attention to structural determinants or downstream harms (Wesson et al. 2022; Chin et al. 2023; Char, Abràmoff, and Feudtner 2020). Although Rajkomar et al. (2018), Vollmer et al. (2020), and others such as Fletcher, Nakeshimana, and Olubeko (2021), Char, Abràmoff, and Feudtner (2020), Chin et al. (2023), Morgenstern et al. (2020), and Raza (2023) offer principles, checklists, or guiding criteria to address AF, most remain either high-level, clinical in scope, or insufficiently tailored to the structural and interdisciplinary challenges of PH+ML.

We respond to this gap by introducing ACAR, a PH-oriented, stage-based framework grounded in our empirical findings. Unlike prior approaches, ACAR is directly derived from RABAT-assessed reporting gaps and provides targeted guidance for fairness-aware ML design, evaluation, and reporting in PH contexts. As detailed in Table 3, the framework translates observed AB reporting gaps into actionable prompts, facilitating AF reflection throughout the PH+ML lifecycle. These guiding questions are intentionally designed for interdisciplinary teams, reflecting our observation that PH+ML research commonly involves collaborators with diverse expertise and backgrounds.

5.5 Implications for Practice and Policy

The lack of structured AB assessments in PH+ML research calls for systemic changes in education, research design, and governance. ACAR offers a practical set of questions to embed fairness reflection across PH+ML workflows. While not a technical tool, its staged structure helps diverse research teams identify AB risks, consider fairness, and improve reporting, enhancing transparency, replicability, and comparability. The questions were intentionally designed with a low barrier to adoption, allowing integration into existing projects regardless of ML maturity. ACAR complements existing protocols without requiring significant changes. However, institutionalizing these practices may face obstacles such as limited awareness, weak incentives, and unclear responsibility for fairness within teams. Targeted training, evaluation mandates, and clearer roles may help overcome these barriers. Educational programs can include fairness case studies and interdisciplinary training to build capacity for AB engagement through ACAR. Policy measures such as the Dutch Algorithm Register and the CBS classification reform can support fairness reflection in PH+ML by enhancing algorithm transparency and improving equity in population classifications. Alignment with the EU AI Act and national AI strategies can further support AB reporting, though current efforts focus on sectors like law enforcement and healthcare rather than PH. Journals and funders can also promote shifts in norms by requiring AB discussion in publications and proposals. Ultimately, AF in PH+ML must become routine, not optional, in responsible research.

5.6 Future Work

We recommend (1) empirically validating ACAR in active PH+ML projects; (2) exploring researchers' decision-making around AB (e.g., interviews or surveys) to understand whether and why fairness considerations are addressed but not documented; (3) applying RABAT and ACAR in non-Dutch contexts to test their applicability and identify any necessary context-specific adaptations; (4) developing and testing training tools (e.g., workshops or online modules) to support ACAR use; (5) using collected metadata to explore additional patterns in AB reporting, including differences by venue or discipline; and (6) extending future reviews to include local, non-indexed, or non-English sources to better capture underrepresented work.

6 Conclusion

This study provides the first systematic assessment of AB reporting in Dutch PH+ML research. Applying the RABAT framework to 35 studies, we reveal that, despite strong technical rigor in data handling, fairness considerations—especially ML-specific framing, sensitive-attribute analysis, and transparent harm disclosure—are largely absent. Our ACAR framework offers a clear, actionable pathway to embed fairness at every stage of the ML lifecycle. We urge PH+ML researchers to adopt ACAR-informed reporting practices and leverage national transparency mechanisms to ensure that future PH+ML applications advance health equity rather than exacerbate existing disparities.

Acknowledgments

This research was conducted at the Civic AI Lab (SIAS group), Informatics Institute, University of Amsterdam, in collaboration with the Municipality of Amsterdam and the Dutch Ministry of the Interior and Kingdom Relations.

References

Altman, E. I.; and Saunders, A. 1998. Credit Risk Measurement: Developments over the Last 20 Years. *Journal of Banking & Finance*, 21(11-12): 1721–1742.

Angwin, J.; Larson, J.; Mattu, S.; and Kirchner, L. 2016. Machine Bias. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. ProPublica, May 23.

Arnold, K. E.; and Pistilli, M. D. 2012. Course Signals at Purdue: Using Learning Analytics to Increase Student Success. In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, 267–270. Vancouver, BC, Canada: ACM.

Aïvodji, U.; Arai, H.; Fortineau, O.; Gambs, S.; Hara, S.; and Tapp, A. 2019. Fairwashing: The Risk of Rationalization. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, 161–170.

Barocas, S.; Hardt, M.; and Narayanan, A. 2023. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press. ISBN 9780262048613.

Barocas, S.; and Selbst, A. D. 2016. Big Data's Disparate Impact. *SSRN Electronic Journal*.

Bellamy, R. K. E.; Dey, K.; Hind, M.; Hoffman, S. C.; Houde, S.; Kannan, K.; Lohia, P.; Martino, J.; Mehta, S.; Mojsilović, A.; Nagar, S.; Ramamurthy, K. N.; Richards, J.; Saha, D.; Sattigeri, P.; Singh, M.; Varshney, K. R.; and Zhang, Y. 2019. AI Fairness 360: An Extensible Toolkit for Detecting and Mitigating Algorithmic Bias. *IBM Journal of Research and Development*, 63(4/5): 4:1–4:15.

Benke, K.; and Benke, G. 2018. Artificial Intelligence and Big Data in Public Health. *International Journal of Environmental Research and Public Health*, 15(12): 2796.

Binns, R. 2018. Fairness in Machine Learning: Lessons from Political Philosophy. In Friedler, S. A.; and Wilson, C., eds., *Proceedings of the 2018 Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, 149–159. PMLR.

Bird, S.; Dudik, M.; Edgar, R.; Horn, B.; Lutz, R.; Milan, V.; Sameki, M.; Wallach, H.; and Walker, K. 2020. Fairlearn: A Toolkit for Assessing and Improving Fairness in AI. Technical Report MSR-TR-2020-32, Microsoft Research.

Bossuyt, P. M.; Reitsma, J. B.; Bruns, D. E.; Gatsonis, C. A.; Glasziou, P. P.; Irwig, L.; Lijmer, J. G.; Moher, D.; Rennie, D.; de Vet, H. C.; Kressel, H. Y.; Rifai, N.; Golub, R. M.; Altman, D. G.; Hooft, L.; Korevaar, D. A.; and Cohen, J. F. 2015. STARD 2015: An Updated List of Essential Items for Reporting Diagnostic Accuracy Studies. *Radiology*, 277(3): 826–832.

Brown, T. 2009. *Change by Design: How Design Thinking Creates New Alternatives for Business and Society*. New York: HarperBusiness. ISBN 9780061766084.

Buolamwini, J.; and Gebru, T. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, 77–91. PMLR.

Burrell, J. 2016. How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1): 1–12.

Carrera-Rivera, A.; Ochoa, W.; Larrinaga, F.; and Lasa, G. 2022. How-to conduct a systematic literature review: A quick guide for computer science research. *MethodsX*, 9: 101895.

Caton, S.; and Haas, C. 2024. Fairness in Machine Learning: A Survey. *ACM Computing Surveys*, 56(7): 1–38.

Centraal Bureau voor de Statistiek (CBS). 2022. Migration Background No Longer Used as Standard Variable. Accessed April 21, 2025.

Chan, A.-W.; Tetzlaff, J. M.; Altman, D. G.; Laupacis, A.; Gøtzsche, P. C.; Krleža-Jerić, K.; Hróbjartsson, A.; Mann, H.; Dickersin, K.; Berlin, J. A.; Doré, C. J.; Parulekar, W. R.; Summerskill, W. S. M.; Groves, T.; Schulz, K. F.; Sox, H. C.; Rockhold, F. W.; Rennie, D.; and Moher, D. 2013. SPIRIT 2013 Statement: Defining Standard Protocol Items for Clinical Trials. *Annals of Internal Medicine*, 158(3): 200–207.

Char, D. S.; Abràmoff, M. D.; and Feudtner, C. 2020. Identifying Ethical Considerations for Machine Learning Healthcare Applications. *The American Journal of Bioethics*, 20(11): 7–17.

Chen, F.; Wang, L.; Hong, J.; Jiang, J.; and Zhou, L. 2024. Unmasking Bias in Artificial Intelligence: A Systematic Review of Bias Detection and Mitigation Strategies in Electronic Health Record-Based Models. *Journal of the American Medical Informatics Association*, 31(5): 1172–1183.

Chin, M. H.; Afsar-Manesh, N.; Bierman, A. S.; Chang, C.; Colón-Rodríguez, C. J.; Dullabh, P.; Duran, D. G.; Fair, M.; Hernandez-Boussard, T.; Hightower, M.; Jain, A.; Jordan, W. B.; Konya, S.; Moore, R. H.; Moore, T. T.; Rodriguez, R.; Shaheen, G.; Snyder, L. P.; Srinivasan, M.; Umscheid, C. A.; and Ohno-Machado, L. 2023. Guiding Principles to Address the Impact of Algorithm Bias on Racial and Ethnic Disparities in Health and Health Care. *JAMA Network Open*, 6(12): e2345050.

Chouldechova, A. 2017. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data*, 5(2): 153–163.

Collins, G. S.; Moons, K. G. M.; Dhiman, P.; Riley, R. D.; Beam, A. L.; Van Calster, B.; Ghassemi, M.; Liu, X.; Reitsma, J. B.; van Smeden, M.; et al. 2024. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*, 385: e078378.

Collins, G. S.; Reitsma, J. B.; Altman, D. G.; and Moons, K. G. M. 2015. Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD): The TRIPOD Statement. *Circulation*, 131(2): 211–219.

Corbett-Davies, S.; Gaebler, J. D.; Nilforoshan, H.; Shroff, R.; and Goel, S. 2023. The Measure and Mismeasure of Fairness. *Journal of Machine Learning Research*, 24(312): 1–117.

Delgado, J.; de Manuel, A.; Parra, I.; Moyano, C.; Rueda, J.; Guersenzvaig, A.; Ausín, T.; Cruz, M.; Casacuberta, D.; and

Puyol, A. 2022. Bias in Algorithms of AI Systems Developed for COVID-19: A Scoping Review. *Journal of Bioethical Inquiry*, 19(3): 407–419.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness Through Awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 214–226. New York, NY, USA: Association for Computing Machinery.

Essink-Bot, M.-L.; Lamkaddem, M.; Jellema, P.; Nielsen, S. S.; and Stronks, K. 2013. Interpreting Ethnic Inequalities in Healthcare Consumption: A Conceptual Framework for Research. *The European Journal of Public Health*, 23(6): 922–926.

Eubanks, V. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. New York: St. Martin's Press. ISBN 9781250074317.

European Parliament and Council of the European Union. 2024. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending certain Union legislative acts (AI Act). Official Journal of the European Union, L 2024/1689, 12 July 2024.

Feldman, M.; Friedler, S. A.; Moeller, J.; Scheidegger, C.; and Venkatasubramanian, S. 2015. Certifying and Removing Disparate Impact. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, 259–268. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-3664-2.

Ferrara, E. 2024. GenAI Against Humanity: Nefarious Applications of Generative Artificial Intelligence and Large Language Models. *Journal of Computational Social Science*, 7(1): 549–569.

Fletcher, R. R.; Nakeshimana, A.; and Olubeko, O. 2021. Addressing Fairness, Bias, and Appropriate Use of Artificial Intelligence and Machine Learning in Global Health. *Frontiers in Artificial Intelligence*, 3: 561802.

Flores, L.; Kim, S.; and Young, S. D. 2024. Addressing Bias in Artificial Intelligence for Public Health Surveillance. *Journal of Medical Ethics*, 50(3): 190–194.

Gagnier, J. J.; Kienle, G.; Altman, D. G.; Moher, D.; Sox, H.; and Riley, D. 2013. The CARE Guidelines: Consensus-Based Clinical Case Reporting Guideline Development. *Global Advances in Health and Medicine*, 2(5): 38–43.

Galanty, M.; Luitse, D.; Noteboom, S. H.; Croon, P.; Vlaar, A. P.; Poell, T.; Sanchez, C. I.; Blanke, T.; and Išgum, I. 2024. Assessing the Documentation of Publicly Available Medical Image and Signal Datasets and Their Impact on Bias Using the BEAMRAD Tool. *Scientific Reports*, 14(1): 31846.

Galhotra, S.; Brun, Y.; and Meliou, A. 2017. Fairness Testing: Testing Software for Discrimination. In *Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering*, 498–510. ACM.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Daumé III, H.; and Crawford, K. 2021. Datasheets for Datasets. *Communications of the ACM*, 64(12): 86–92.

Gianfrancesco, M. A.; Tamang, S.; Yazdany, J.; and Schmajuk, G. 2018. Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data. *JAMA Internal Medicine*, 178(11): 1544–1547.

Ginsberg, J.; Mohebbi, M. H.; Patel, R. S.; Brammer, L.; Smolinski, M. S.; and Brilliant, L. 2009. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232): 1012–1014.

Government of the Netherlands. 2024. Algoritmeregister: Overview of algorithmic systems used by public institutions. Available online, accessed 2025-03-22.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems*, volume 29.

Higgins, J. P. T.; Altman, D. G.; Gøtzsche, P. C.; Jüni, P.; Moher, D.; Oxman, A. D.; Savović, J.; Schulz, K. F.; Weeks, L.; and Sterne, J. A. C. 2011. The Cochrane Collaboration's Tool for Assessing Risk of Bias in Randomised Trials. *BMJ*, 343: d5928.

Hillebrand, M.; Khan, I.; Peleja, F.; and Oliver, N. 2020. MobiSenseUs: Inferring Aggregate Objective and Subjective Well-being from Mobile Data. In *ECAI 2020: 24th European Conference on Artificial Intelligence*, 1818–1825. IOS Press.

Holstege, L.; van de Geer, S.; Leenes, R.; and van Sluijs, J. 2025. Auditing a Dutch Public Sector Risk Profiling Algorithm Using an Unsupervised Bias Detection Tool. Preprint. Available at https://arxiv.org/pdf/2502.01713, arXiv:2502.01713v2.

Holstein, K.; Wortman Vaughan, J.; Daumé III, H.; Dudík, M.; and Wallach, H. 2019. Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need? In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–16. ACM.

Husereau, D.; Drummond, M.; Petrou, S.; Carswell, C.; Moher, D.; Greenberg, D.; Augustovski, F.; Briggs, A. H.; Mauskopf, J.; and Loder, E. 2013. Consolidated Health Economic Evaluation Reporting Standards (CHEERS) Statement. *International Journal of Technology Assessment in Health Care*, 29(2): 117–122.

Ibrahim, H.; Liu, X.; Denniston, A. K.; Fraser, H.; Keane, P. A.; Faes, L.; Geerts, B.; Chambers, D.; Corral, J.; Lee, A. M.; Wagner, M.; and et al. 2021. STARD-AI: A Reporting Guideline for Studies Using Artificial Intelligence in Diagnostic Test Accuracy Studies. *BMJ Open*, 11(6): e041411.

IEEE. 2025. IEEE Standard for Algorithmic Bias Considerations. IEEE Std 7003-2024. Pages 1–59.

Ikram, U. Z.; Kunst, A. E.; Lamkaddem, M.; and Stronks, K. 2014. The disease burden across different ethnic groups in Amsterdam, the Netherlands, 2011–2030. *The European Journal of Public Health*, 24(4): 600–605.

Ilozumba, O.; Koster, T. S.; Syurina, E. V.; and Ebuenyi, I. 2022. Ethnic minority experiences of mental health services in the Netherlands: An exploratory study. *BMC Research Notes*, 15(1): 266.

Jiang, F.; Jiang, Y.; Zhi, H.; Dong, Y.; Li, H.; Ma, S.; Wang, Y.; Dong, Q.; Shen, H.; and Wang, Y. 2017. Artificial intelligence in healthcare: past, present and future. *Stroke and Vascular Neurology*, 2(4): 230–243.

Kilkenny, C.; Browne, W. J.; Cuthill, I. C.; Emerson, M.; and Altman, D. G. 2010. Improving bioscience research reporting: the ARRIVE guidelines for reporting animal research. *PLoS Biology*, 8(6): e1000412.

Kroneman, M.; Boerma, W.; van den Berg, M.; Groenewegen, P.; de Jong, J.; and van Ginneken, E. 2016. Netherlands: health system review. *Health Systems in Transition*, 18(2): 1–240.

Kusner, M. J.; Loftus, J.; Russell, C.; and Silva, R. 2017. Counterfactual Fairness. In *Advances in Neural Information Processing Systems*, volume 30.

Lekadir, K.; Frangi, A. F.; Porras, A. R.; Glocker, B.; Cintas, C.; Langlotz, C. P.; Weicken, E.; Asselbergs, F. W.; Prior, F.; Collins, G. S.; Kaissis, G.; Tsakou, G.; Buvat, I.; Kalpathy-Cramer, J.; Mongan, J.; Schnabel, J. A.; Kushibar, K.; Riklund, K.; Marias, K.; Amugongo, L. M.; Fromont, L. A.; Maier-Hein, L.; Cerdá-Alberich, L.; Martí-Bonmatí, L.; Cardoso, M. J.; Bobowicz, M.; Shabani, M.; Tsiknakis, M.; Zuluaga, M. A.; Fritzsche, M.-C.; Camacho, M.; Linguraru, M. G.; Wenzel, M.; De Bruijne, M.; Tolsgaard, M. G.; Goisauf, M.; Cano Abadía, M.; Papanikolaou, N.; Lazrak, N.; Pujol, O.; Osuala, R.; Napel, S.; Colantonio, S.; Joshi, S.; Klein, S.; Aussó, S.; Rogers, W. A.; Salahuddin, Z.; Starmans, M. P. A.; and Consortium, F.-A. 2025. FUTURE-AI: International consensus guideline for trustworthy and deployable artificial intelligence in healthcare. *BMJ (Clinical research ed.)*, 388: e081554.

Liu, X.; Cruz Rivera, S.; Moher, D.; Calvert, M. J.; and Denniston, A. K. 2020. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nature Medicine*, 26: 1364–1374.

Madaio, M. A.; Stark, L.; Wortman Vaughan, J.; and Wallach, H. 2020. Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, 1–14. New York, NY, USA: Association for Computing Machinery.

Margetts, H.; and Naumann, A. 2017. Government as a Platform: What Can Estonia Show the World? Research paper, University of Oxford.

Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; and Galstyan, A. 2021. A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys*, 54(6): 1–35.

Mhasawade, V.; Zhao, Y.; and Chunara, R. 2021. Machine Learning and Algorithmic Fairness in Public and Population Health. *Nature Machine Intelligence*, 3(8): 659–666.

Ministry of Infrastructure and Water Management. 2024. AI Impact Assessment: The Tool for a Responsible AI Project. Technical report, Government of the Netherlands.

Ministry of the Interior and Kingdom Relations. 2024. Government-wide Vision on Generative AI of the Netherlands. Technical report, Government of the Netherlands.

Mitchell, M.; Wu, S.; Zaldivar, A.; Barnes, P.; Vasserman, L.; Hutchinson, B.; Spitzer, E.; Raji, I. D.; and Gebru, T. 2019. Model Cards for Model Reporting. In *Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency (FAT* 2019)*, 220–229. New York, NY, USA: Association for Computing Machinery.

Mittelstadt, B. D.; Allo, P.; Taddeo, M.; Wachter, S.; and Floridi, L. 2016. The Ethics of Algorithms: Mapping the Debate. *Big Data & Society*, 3(2): 2053951716679679.

Mongan, J.; Moy, L.; and Kahn, C. E. J. 2020. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A Guide for Authors and Reviewers. *Radiology: Artificial Intelligence*, 2(2): e200029.

Moons, K. G. M.; Damen, J. A. A.; Kaul, T.; Riley, R. D.; Reitsma, J. B.; van Smeden, M.; Wolff, R. F.; et al. 2025. PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. *BMJ*, 388: e082505.

Moons, K. G. M.; Wolff, R. F.; Riley, R. D.; Whiting, P. F.; Westwood, M.; Collins, G. S.; Reitsma, J. B.; Kleijnen, J.; and Mallett, S. 2019. PROBAST: A Tool to Assess Risk of Bias and Applicability of Prediction Model Studies: Explanation and Elaboration. *Annals of Internal Medicine*, 170(1): W1–W33.

Morgenstern, J. D.; Buajitti, E.; O'Neill, M.; Piggott, T.; Goel, V.; Fridman, D.; Kornas, K.; and Rosella, L. C. 2020. Predicting Population Health with Machine Learning: A Scoping Review. *BMJ Open*, 10(10): e037860.

Obermeyer, Z.; Powers, B.; Vogeli, C.; and Mullainathan, S. 2019. Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations. *Science*, 366(6464): 447–453.

O'Brien, B. C.; Harris, I. B.; Beckman, T. J.; Reed, D. A.; and Cook, D. A. 2014. Standards for Reporting Qualitative Research: A Synthesis of Recommendations. *Academic Medicine*, 89(9): 1245–1251.

Page, M. J.; McKenzie, J. E.; Bossuyt, P. M.; Boutron, I.; Hoffmann, T. C.; Mulrow, C. D.; Shamseer, L.; Tetzlaff, J. M.; Akl, E. A.; Brennan, S. E.; Chou, R.; Glanville, J.; Grimshaw, J. M.; Hróbjartsson, A.; Lalu, M. M.; Li, T.; Loder, E. W.; Mayo-Wilson, E.; McDonald, S.; McGuinness, L. A.; Stewart, L. A.; Thomas, J.; Tricco, A. C.; Welch, V. A.; Whiting, P.; and Moher, D. 2021. The PRISMA 2020 Statement: An Updated Guideline for Reporting Systematic Reviews. *BMJ*, 372: n71.

Patel, J.; Shah, S.; Thakkar, P.; and Kotecha, K. 2015. Predicting Stock Market Index Using Fusion of Machine Learning Techniques. *Expert Systems with Applications*, 42(4): 2162–2172.

Rajkomar, A.; Hardt, M.; Howell, M. D.; Corrado, G.; and Chin, M. H. 2018. Ensuring Fairness in Machine Learning to Advance Health Equity. *Annals of Internal Medicine*, 169(12): 866–872.

Raza, S. 2023. Connecting Fairness in Machine Learning with Public Health Equity. In *Proceedings of the 2023 IEEE 11th International Conference on Healthcare Informatics (ICHI)*, 704–708. IEEE.

Raza, S.; Shaban-Nejad, A.; Dolatabadi, E.; and Mamiya, H. 2024. Exploring Bias and Prediction Metrics to Characterise the Fairness of Machine Learning for Equity-Centered Public Health Decision-Making: A Narrative Review. *IEEE Access*, 12: 180815–180829.

Rostamzadeh, N.; Mincu, D.; Roy, S.; Smart, A.; Wilcox, L.; Pushkarna, M.; Schrouff, J.; Amironesei, R.; Moorosi, N.; and Heller, K. 2022. Healthsheet: Development of a Transparency Artifact for Health Datasets. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 1943–1961. Association for Computing Machinery.

Saleiro, P.; Kuester, B.; Hinkson, L.; London, J.; Stevens, A.; Anisfeld, A.; Rodolfa, K. T.; and Ghani, R. 2018. Aequitas: A Bias and Fairness Audit Toolkit. arXiv preprint arXiv:1811.05577.

Selbst, A. D.; boyd, d.; Friedler, S. A.; Venkatasubramanian, S.; and Vertesi, J. 2019. Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency (FAT*)*, 59–68. Association for Computing Machinery.

Sikstrom, L.; Maslej, M. M.; Hui, K.; Findlay, Z.; Buchman, D. Z.; and Hill, S. L. 2022. Conceptualising Fairness: Three Pillars for Medical Algorithms and Health Equity. *BMJ Health & Care Informatics*, 29(1): e100459.

Suresh, H.; and Guttag, J. V. 2019. A Framework for Understanding Unintended Consequences of Machine Learning. arXiv preprint arXiv:1901.10002.

Teunissen, E.; Van Bavel, E.; Van den Driessen Mareeuw, F.; MacFarlane, A.; Van Weel-Baumgarten, E.; Van den Muijsenbergh, M.; and Van Weel, C. 2015. Mental Health Problems of Undocumented Migrants in the Netherlands: A Qualitative Exploration of Recognition, Recording, and Treatment by General Practitioners. *Scandinavian Journal of Primary Health Care*, 33(2): 82–90.

Thomasian, N. M.; Eickhoff, C.; and Adashi, E. Y. 2021. Advancing Health Equity with Artificial Intelligence. *Journal of Public Health Policy*, 42(4): 602–611.

Tong, A.; Sainsbury, P.; and Craig, J. 2007. Consolidated Criteria for Reporting Qualitative Research (COREQ): A 32-Item Checklist for Interviews and Focus Groups. *International Journal for Quality in Health Care*, 19(6): 349–357.

Tramèr, F.; Atlidakis, V.; Geambasu, R.; Hsu, D.; Hubaux, J.-P.; Humbert, M.; Juels, A.; and Lin, H. 2017. FairTest: Discovering Unwarranted Associations in Data-Driven Applications. In *2017 IEEE European Symposium on Security and Privacy (EuroS&P)*, 401–416. IEEE.

Tsai, T. C.; Arik, S.; Jacobson, B. H.; Yoon, J.; Yoder, N.; Sava, D.; Mitchell, M.; Graham, G.; and Pfister, T. 2022. Algorithmic Fairness in Pandemic Forecasting: Lessons from COVID-19. *NPJ Digital Medicine*, 5(1): 59.

Verma, S.; and Rubin, J. 2018. Fairness Definitions Explained. In *Proceedings of the International Workshop on Software Fairness*, 1–7. ACM.

Vollmer, S.; Mateen, B. A.; Bohner, G.; Király, F. J.; Ghani, R.; Jonsson, P.; Cumbers, S.; Jonas, A.; McAllister, K. S.; Myles, P.; and et al. 2020. Machine Learning and Artificial Intelligence Research for Patient Benefit: 20 Critical Questions on Transparency, Replicability, Ethics, and Effectiveness. *BMJ*, 368: l6927.

von Elm, E.; Altman, D. G.; Egger, M.; Pocock, S. J.; Gøtzsche, P. C.; and Vandenbroucke, J. P. 2007. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: Guidelines for Reporting Observational Studies. *The Lancet*, 370(9596): 1453–1457.

Wesson, P.; Hswen, Y.; Valdes, G.; Stojanovski, K.; and Handley, M. A. 2022. Risks and Opportunities to Ensure Equity in the Application of Big Data Research in Public Health. *Annual Review of Public Health*, 43(1): 59–78.

Wexler, J.; Pushkarna, M.; Bolukbasi, T.; Wattenberg, M.; Viégas, F.; and Wilson, J. 2020. The What-If Tool: Interactive Probing of Machine Learning Models. *IEEE Transactions on Visualization and Computer Graphics*, 26(1): 56–65.

Wiemken, T. L.; and Kelley, R. R. 2020. Machine Learning in Epidemiology and Health Outcomes Research. *Annual Review of Public Health*, 41: 21–36.

Wiśniewski, J.; and Biecek, P. 2022. fairmodels: A Flexible Tool for Bias Detection, Visualization, and Mitigation in Binary Classification Models. *The R Journal*, 14(1): 227–243.

Wolff, R. F.; Moons, K. G. M.; Riley, R. D.; Whiting, P. F.; Westwood, M.; Collins, G. S.; Reitsma, J. B.; Kleijnen, J.; and Mallett, S. 2019. PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. *Annals of Internal Medicine*, 170(1): 51–58.

World Health Organization. 2021. Ethics and Governance of Artificial Intelligence for Health. Accessed 2025-05-22.

Xu, J.; Xiao, Y.; Wang, W. H.; Ning, Y.; Shenkman, E. A.; Bian, J.; and Wang, F. 2022. Algorithmic Fairness in Computational Medicine. *eBioMedicine*, 84: 104250.