From Awareness to Action: Understanding and Overcoming the Research-Practice Gap in Algorithmic Fairness for Public Health

Sara Altamirano
Informatics Institute
University of Amsterdam
Science Park 900, 1098 XH Amsterdam
s.e.altamirano@uva.nl

Tijs Portegies
Informatics Institute
University of Amsterdam
Science Park 900, 1098 XH Amsterdam
t.c.portegies@uva.nl

Sennay Ghebreab
Informatics Institute
University of Amsterdam
Science Park 900, 1098 XH Amsterdam
s.ghebreab@uva.nl

Abstract

Algorithmic fairness is essential for responsible ML-driven public health research, yet its practical implementation remains limited. To investigate this awareness–action gap, we conducted a sequential mixed-methods study comprising expert interviews, an online survey, and systematic mapping. The expert interviews informed the design of the survey, which in turn revealed fragmented definitions of fairness, limited training and guidance, reliance on external sources, and rare use of formal assessment, mitigation, or monitoring. These findings were subsequently mapped onto three established research–practice gap lenses: the Knowledge–Practice Gap, the Knowledge-to-Action Cycle, and the Knowing–Doing Gap, each offering complementary perspectives. Building on this synthesis, we introduce the Fairness-to-Action framework, which integrates methodological, organizational, and systemic dimensions to identify where translation of algorithmic fairness knowledge stalls. Our analysis shows that fairness remains weakly institutionalized, translation mechanisms are externally driven, and system-level priorities continue to emphasize accuracy over fairness. These insights suggest critical leverage points for advancing safe, fair, and ethical ML-driven public health research practice.

1 Introduction

Algorithmic decision-making (ADM) is reshaping public health through applications in disease surveillance, outcome prediction, resource allocation, and identification of vulnerable groups [1–6]. By integrating large-scale data and predictive modeling, machine learning (ML) enables new forms of population-level monitoring and intervention. Yet these opportunities carry significant risks: models trained on incomplete or skewed data can reproduce socioeconomic, racial, or geographic disparities [7–9]. In public health, such risks can be amplified because determinants of health are multi-layered, trade-offs across groups are often unavoidable, and decisions often span multiple institutions [10–12]. Technical validity alone is therefore insufficient to ensure safe or equitable outcomes [13, 14], making fairness a central pillar of responsible ML.

Article presented at the Second IASEAI Conference (IASEAI'26).

Accordingly, algorithmic fairness has emerged as a central ethical and societal concern. Approaches range from quantitative criteria such as demographic parity [15] and equalized odds [16] to procedural measures emphasizing transparency and stakeholder participation [17–19]. Growing awareness reflects a shift from predictive performance toward equity and accountability as conditions for responsible ML. Yet awareness alone rarely translates into systematic mitigation or transparent reporting, and practices remain fragmented across domains [20–22]. Even in advanced health systems, fairness research often remains disconnected from public health implementation [13, 10]. This persistent gap limits the real-world impact of fairness and calls for empirical insight into how researchers themselves interpret and apply it.

To address this gap, we examine how researchers in ML-driven public health *perceive* and *operationalize* fairness across the ML lifecycle: conceptualization, awareness, evaluation, design, and application (adapted from [23–25]). These five dimensions structure our research sub-questions (RSQ1 to RSQ5). To interpret the observed disconnect between fairness research and practice, we draw on established theories from implementation science: the Knowledge–Practice Gap [26], the Knowledge-to-Action (KTA) Cycle [27], and the Knowing–Doing Gap [28]. Using a sequential mixed-methods design, we first conducted expert interviews to inform the design of an online survey, then analyzed survey responses to map researcher perceptions and practices against these frameworks. This synthesis yields the Fairness-to-Action (F2A) framework, which identifies where fairness knowledge stalls before reaching practice and outlines conditions for more effective translation in ML-driven public health research. We ground the study in the Netherlands, a salient case as it combines mature public health and ML capacity, advanced data infrastructures, and an active artificial intelligence (AI) governance landscape with documented inequities among ethnic minorities and socioeconomically disadvantaged groups [29–31]. These features make it an informative case with lessons for high-resource settings.

Statement of Contributions. First, we empirically examine how researchers in ML-driven public health perceive and operationalize algorithmic fairness across the ML lifecycle through an interview-informed survey. Second, we integrate evidence with implementation science by situating the observed awareness–action gap within three established theoretical lenses, identifying pathways for improving fairness adoption (see Section 4). Third, we extend this theoretical synthesis to the fairness domain and introduce the F2A framework (see Section 5), which can help explain how fairness knowledge becomes sustained practice and where the process often stalls. Together, these contributions synthesize empirical and theoretical insights to advance safe, fair, and ethical ML-driven public health research practice.

2 Related Work

Reviews of ML-driven public health consistently identify sources of bias and equity risks, ranging from data absenteeism and representational gaps to mis-specified objectives and model misuse [4, 10, 32, 22, 9]. Empirical studies further demonstrate subgroup disparities and performance variation: racial bias in commercial risk scores [7], underdiagnosis in underserved populations [33], poor external generalizability [34], gender imbalance in imaging datasets [35], and inequities embedded in race corrections [36]. In response, mitigation strategies emphasize fairness metrics, explainability, and lifecycle-based awareness [37, 38, 25, 14, 13], complemented by structured reporting and evaluation standards [39–44]. Conceptual work stresses transparency, accountability, and inclusion as pillars for algorithmic fairness [45], yet metrics and documentation alone are insufficient: over-reliance may reinforce inequities [46, 12, 8], particularly where population-level trade-offs, structural determinants, and fragmented governance complicate evaluation and mitigation [47]. Collectively, this literature underscores the limits of technical fixes and highlights the need for integrated governance, participatory design, and context-sensitive evaluation. Despite extensive technical and policy work, little is known about how researchers themselves engage with algorithmic fairness. Increasingly, fairness in ML-driven public health is framed as a socio-technical concern. Ethical and policy analyses emphasize justice, transparency, and inclusion [48, 45, 32, 10], while human-computer interaction and social science studies introduce tools such as model cards, datasheets, and checklists [49, 50, 21]. However, adoption remains inconsistent. Empirical studies reveal heterogeneity: industry practitioners report limited uptake of fairness practices [20], AI developers cite the lack of fair data, guidelines, and expertise as barriers [51], and systematic reviews show fragmented definitions and context-dependent judgments [52, 53]. Broader critiques link fairness to structural inequities and governance gaps [46, 12, 54]. In the Netherlands, these global trends are

reflected in a context where advanced data infrastructures coexist with persistent inequities among ethnic minorities and disadvantaged groups [29–31]. Despite policy initiatives such as the Algorithm Register [55] and revised data classification standards [56], explicit fairness evaluations in ML-driven public health remain rare [57]. This study extends the literature by centering researchers as the unit of analysis and examining the research–practice gap through an interview-informed survey.

3 Theoretical Framework

Translating algorithmic fairness into practice requires attention to methodological, organizational, and systemic factors. Implementation science offers complementary perspectives for understanding how knowledge moves, or fails to move, into practice [58–61]; F2A integrates these lenses to interpret where fairness translation falters. This approach is particularly relevant to public health, where translation spans multiple institutions, competing priorities, and unavoidable trade-offs across populations, underscoring the need to assess both implementation processes and outcomes.

To structure our methodological approach, we draw on three established theories that together specify the *what*, *how*, and *why* of fairness translation. The Knowledge–Practice Gap [26] delineates what characterizes the current state of the gap: effective knowledge rarely becomes routine practice without targeted translation activities and supportive conditions. The KTA Cycle [27] clarifies how fairness knowledge is transferred through an action cycle that adapts evidence to context, addresses barriers, implements strategies, and evaluates sustainment. The Knowing–Doing Gap [28] explains why the gap persists, emphasizing organizational dynamics that substitute discourse for action, privilege narrow metrics, or sustain risk-averse norms. These theories have been extended through models such as the Consolidated Framework for Implementation Research (CFIR), the Active Implementation Frameworks, and the Centers for Disease Control and Prevention (CDC) KTA guide [62, 58, 63]. Adaptations emphasize team-based learning, continuous feedback, and iterative improvement [64, 65].

In short, these perspectives clarify *what* defines the gap, *how* fairness knowledge is transferred or stalled, and *why* it persists. This threefold structure provides a systems-level lens: methodological factors shape how knowledge is produced, organizational factors govern its use, and systemic factors determine its sustainment. Together, these dimensions offer a foundation for analyzing fairness translation across the ML lifecycle. Section 4 outlines how we applied these lenses.

4 Methods

4.1 Empirical Study

We used a two-phase sequential exploratory mixed-methods design [66] to examine how algorithmic fairness is perceived and operationalized in ML-driven public health. Phase one comprised semi-structured expert interviews (28 questions across four sections on ADM use, fairness awareness and practice, and study-level practices adapted from [50, 21, 67]). A thematic analysis grounded constructs, aligned practitioner vocabulary, and shaped specific prompts for the survey. Instruments were piloted (interview guide with two colleagues; survey with five) with edits limited to wording and logic. Survey items were derived from coded interview themes to ensure content validity and traceability to qualitative constructs, supported by expert triangulation during iterative revision. Item to RSQ mapping is shown in Figure 1. Reporting follows CROSS [68] and CHERRIES [69]. The full survey questionnaire is available in the extended version on arXiv; additional documentation is available on request.

Interviewees were recruited purposively and via snowballing. Of 33 invitations, four experts participated, meeting two criteria: willingness to reflect on fairness and demonstrated expertise. Two interviews were conducted via recorded video call and two via written questionnaires; all were in English and lasted 30 to 45 minutes. Participants represented diverse academic, policy, and technical backgrounds, spanning varying levels of seniority and institutional roles. While thematic saturation was not sought, the interviews were used for instrument development and conceptual coverage in an exploratory design. The survey targeted researchers from Dutch institutions working at the ML–public health interface, identified through authorship records, professional networks, and public listings; interviewees were excluded. We distributed 105 personalized invitations and supplemented outreach via social media, academic events, and word of mouth, so the total reach exceeded 105. The study used a cross-sectional online survey administered in Qualtrics (English, February–August

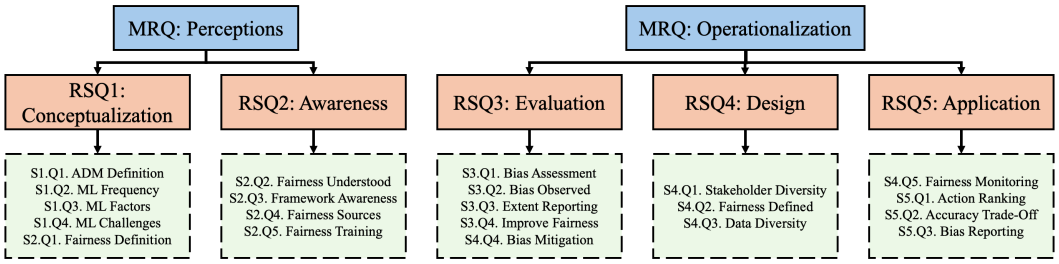

Figure 1: Mapping of survey items to Main Research Question (MRQ) and Research Sub-Questions (RSQs). Solid borders represent research questions, dashed borders indicate grouped items.

2025) via a single anonymous link. We obtained 73 valid responses, defined as submissions with verifiable institutional email addresses and $\geq$90% completion. Given the lack of a complete sampling frame for this niche researcher population, we treat the data as exploratory and use them to identify recurring patterns rather than estimate prevalence. A filter question excluded respondents without ML experience, and the estimated completion time was 15 minutes. Data-quality controls included required consent, completeness checks, and a post hoc email check to limit duplicates; participants could revise responses before submission. Respondents could enter a raffle for five €10 gift cards; contact details were used solely for validation and prize notification, stored separately from survey data, and deleted after use.

Quantitative analysis included descriptive statistics, frequency distributions, and cross-tabulations by S0.Q1 (Primary Field), S0.Q3 (Years Experience), and S0.Q5 (Institution Type). Multiple-answer items were summarized as counts and percentages; the ranking item S5.Q1 by mean rank and rank-1 frequency; Likert items by category counts and percentages, with occasional grouping for interpretability as reported in Section 5. The survey included multiple choice, Likert-scale, ranking, and open-text items. Inferential testing was not performed because subgroup counts fell below thresholds for valid χ^2 or regression analyses; descriptive reporting was emphasized instead. Data cleaning removed direct identifiers and high-risk pseudo-identifiers, deleted linkage keys, validated institutional email domains, and removed emails flagged by the post hoc check. The resulting dataset is anonymized with no retained keys. Item non-response was lower than 1%; no imputation was performed. All analyses were performed in Python (v3.13).

Open-text responses were analyzed using a keyword-assisted inductive approach. A structured codebook defined six themes with descriptions and keyword lists. Two coders independently assigned each response to a single theme based on keyword matches and human judgment; when no keywords matched, the main conceptual idea guided coding. Disagreements were resolved through discussion until consensus. Coders were blinded to respondent metadata. Intercoder reliability statistics were not calculated; iterative consensus coding was used instead. Category frequencies are reported; codebooks with themes, descriptions, and keywords are available in the extended version.

Separate consent was obtained for interviews and surveys. The study involved no sensitive data or vulnerable populations, followed University of Amsterdam ethical guidance and GDPR, and allowed withdrawal at any time. Further details are available in the extended version.

4.2 Theoretical Embedding

To examine fairness translation mechanisms, survey items were embedded within the three frameworks presented in Section 3, with F2A used as an analytical scaffold rather than proposing a new theoretical model. Items were mapped along three axes: *what* fairness knowledge exists or is missing [26], *how* fairness knowledge is transferred and where translation stalls [27], and *why* organizational mechanisms sustain the research-practice gap [28]. Framework constructs were adapted into fairness-specific diagnostic questions to guide mapping. Two researchers independently coded items using a structured spreadsheet; discrepancies were resolved through discussion, and coding decisions were retained for traceability. During interpretation, convergence across frameworks and framework-specific divergences were identified, and subgroup patterns by field, institution type, and ML experience were assessed.

The outputs of this process were organized into three framework-specific tables that present survey findings in terms of the adapted diagnostic questions. Integration was conducted at the interpretive stage by comparing convergence and divergence across the three frameworks. This approach allowed us to preserve the integrity of each framework while highlighting their complementary insights into the fairness translation gap. Full item-to-framework mappings, including coder notes, are available upon request.

5 Results

5.1 Empirical Study

Sample size considerations. Our final sample consisted of 73 respondents. Results are presented using descriptive statistics, cross-tabulations where relevant, and illustrative excerpts from open-ended responses. Percentages are descriptive indicators of patterns in this sample, not population estimates. Detailed distributions for all survey items from RSQ1 to RSQ5 appear in Tables 2 to 6.

Respondent Characteristics: Who Are the Researchers? We analyzed 73 valid responses ($\geq$90% completion, verified contacts). As summarized in Table 1, respondents were from technical and applied fields, with academics the largest role group and universities the most common affiliation. Early-career researchers were concentrated in universities, while those with $\geq$4 years of experience were more often in hospitals and government institutes. Overall, the majority of respondents reported more than four years of experience, with smaller numbers affiliated to the private sector and non-governmental organizations (NGOs).

Table 1: Characteristics of survey respondents.

Primary field	n (%)	**Current role**	n (%)	**Years experience**	n (%)	**Institution type**	n (%)
DS and ML	23 (32)	Academic/Researcher	39 (53)	$<$1 year	2 (3)	University	30 (41)
PH Research	21 (29)	PH Professional	14 (19)	1–3 years	21 (29)	Hospital	20 (27)
Epidemiology/Biostatistics	15 (21)	DS/ML Engineer	12 (16)	4–7 years	27 (37)	Government	16 (22)
PH Policy/Management	8 (11)	Policymaker	7 (10)	8–10 years	11 (15)	NGO	6 (8)
Other[1]	6 (8)	Other[2]	1 (1)	$>$10 years	12 (16)	Private sector	1 (1)

[1] Includes: Medicine; Neuroscience; Sexual and Reproductive Health and Rights/digital rights; Atmospheric Sciences; (Biological) Psychology; Mathematics/Computational Science.
[2] Civil society organization (advisor).
Abbreviations: DS, data science; ML, machine learning; PH, public health; Govt., government; NGO, non-governmental organization. Percentages are rounded and may not total exactly 100%.

RSQ1. Conceptualization: How do ML-driven public health researchers understand and engage with algorithmic decision-making (ADM) and fairness?

Respondents described ADM mainly as decision support or data-based modeling, emphasizing algorithms that assist expert judgment or derive predictions from data, for example, "*Building AI-driven tools to assist decision-making.*" Algorithmic fairness was most often defined as equal treatment or equal performance across groups, typically expressed as "*Ensuring models do not underperform for minority groups.*" In (S1.Q2 to S1.Q4), ML use was frequent among data scientists but less common in public health; data quality and availability were the most cited enablers, while data interoperability and limited ML expertise were leading barriers. In general, ADM and fairness were interpreted through disciplinary lenses: technical for data scientists; ethical and practical for public health researchers, reflecting uneven familiarity and institutional constraints related to roles, data access, and governance.

RSQ2. Awareness: How familiar are researchers with algorithmic fairness in ML-driven public health research?

Awareness of algorithmic fairness was generally low. In (S2.Q2) Fairness Understood, 79% reported little to moderate familiarity, and only one in five indicated strong or full understanding. In (S2.Q3) Framework Awareness, 73% were aware of existing frameworks but without concrete details, while the remainder were unsure or believed none exist; awareness within organizations appeared mainly

Table 2: RSQ1: Conceptualization, response distribution from survey items S1.Q1–S1.Q4 and S2.Q1.

(S1.Q1) ADM Definition		**(S2.Q1) Fairness Definition**		**(S1.Q2) ML Frequency**		**(S1.Q3) ML Factors**		**(S1.Q4) ML Challenges**	
Answers	*n (%)*	*Answers*	*n (%)*	*Answers*	*n (%)*	*Answers*	*n (%)*	*Answers*	*n (%)*
Decision Supp.	22 (30)	Equal Treatment	21 (29)	Rarely	1 (1)	Data Q&A	62 (85)	Data access	50 (68)
Data Analysis	14 (19)	Protected Attributes	20 (27)	Occasionally	14 (19)	Model costs	35 (48)	Computing power	29 (40)
Model Focus	13 (18)	Outcome Fairness	14 (19)	Sometimes	13 (18)	Interoperability	31 (42)	Legal/Regulatory	30 (41)
Pattern Recog.	10 (14)	Standards/Processes	10 (14)	Frequently	21 (29)	Transparency	24 (33)	Interoperability	52 (71)
Rules Based	5 (7)	Context & Account.	0 (0)	Extensively	24 (33)	Ethics	13 (18)	ML expertise	39 (53)
Governance	4 (5)	Equity & Disparities	5 (7)	–	–	Regulatory reqs	24 (33)	Transparency	20 (27)
Unclassified	5 (7)	Unclassified	3 (4)	–	–	Other[1]	8 (11)	Other[2]	2 (3)
Total: 73 (100%)		*Total:* 73 (100%)		*Total:* 73 (100%)		*Base:* N=73 (mult. answer)		*Base:* N=73 (mult. answer)	

[1] "Other" (S1.Q3): trade-off explainability vs accuracy, implementation limits, complex or high-dimensional data, limited support. [2] "Other" (S1.Q4): evaluating personalized causal effects or combined challenges. Percentages are rounded and may not total exactly 100%. Abbreviations: Supp., Support; Recog., Recognition; Account., Accountability; Q&A, quality & availability; reqs., requirements; mult., multiple.

in hospitals. In (S2.Q4) Fairness Sources, most relied on academic or professional channels (e.g., peer-reviewed articles, networks, and conferences) whereas 37% did not stay updated and none cited internal training. In (S2.Q5) Fairness Training, about 70% reported no or minimal learning opportunities, with structured programs rare outside hospitals. In general, fairness knowledge was uneven and concentrated in academic settings, reflecting unequal institutional capacity and limited access to formal training.

Table 3: RSQ2: Awareness, response distribution from survey items S2.Q2–S2.Q5.

(S2.Q2) Fairness Understood		**(S2.Q3) Framework Awareness**		**(S2.Q4) Fairness Sources**		**(S2.Q5) Fairness Training**	
Answers	*n (%)*	*Answers*	*n (%)*	*Answers*	*n (%)*	*Answers*	*n (%)*
Not at all	2 (3)	Yes, within group/org	7 (10)	Academic networks	26 (36)	Regular training	2 (3)
Minimally	31 (43)	Only in field	17 (23)	Peer-reviewed articles	34 (47)	Periodic workshops	9 (12)
Moderately	25 (34)	Aware: no details	29 (40)	Conferences/workshops	18 (25)	External training	13 (18)
Strongly	11 (15)	None exist	6 (8)	Social media/online	20 (27)	Plan to start	6 (8)
Fully	4 (5)	Not sure	11 (15)	Internal training	0 (0)	No programs	37 (51)
–	–	Other[1]	3 (4)	Do not stay updated	27 (37)	Other[3]	6 (8)
–	–	–	–	Other[2]	2 (3)	–	–
Total: 73 (100%)		*Total:* 73 (100%)		*Base:* N=73 (multiple answer)		*Total:* 73 (100%)	

[1] "Other" (S2.Q3): limited or no awareness of organizational fairness guidelines; one cited public resource. [2] "Other" (S2.Q4): fairness discussed informally within teams or during peer review. [3] "Other" (S2.Q5): uncertainty or lack of knowledge about existing training or standards. Percentages are rounded and may not total exactly 100%. Abbreviations: org., organization.

RSQ3. Evaluation: How do researchers assess and report algorithmic bias in ML-driven public health research?

Evaluation of algorithmic bias remained largely informal and descriptive. In (S3.Q1) Bias Assessment, most relied on subgroup sensitivity or demographic analyses and informal discussion, while formal approaches such as fairness metrics or external review were rare. In (S3.Q2) Bias Observed, reported issues mainly concerned prediction accuracy gaps and data quality or sample diversity, with one fifth reporting none. In (S3.Q3) Extent Reporting, 75% described minimal or moderate reporting of bias. In (S3.Q4) Improve Fairness, actions centered on open science and transparency, such as code sharing, algorithm availability, and stakeholder involvement, while formal accountability mechanisms were uncommon. In (S4.Q4) Bias Mitigation, 72% reported none or minimal mitigation. In general, bias evaluation practices were ad hoc and descriptive rather than methodological.

RSQ4. Design: How is algorithmic fairness addressed during the design phase of public health ML projects?

In (S4.Q1) Stakeholder Diversity, 64% reported moderate or stronger inclusion of diverse stakeholders, while one third reported minimal or none. Hospitals showed the highest levels of engagement, and universities the lowest. In (S4.Q2) Fairness Defined, 58% indicated fairness was minimally or not defined, with fewer than 10% reporting strong integration. Data scientists were more likely to mention explicit fairness definitions, while government respondents reported the weakest attention.

Table 4: RSQ3: Evaluation, response distribution from survey items S3.Q1–S3.Q4 and S4.Q4.

(S3.Q1) Bias Assessment		(S3.Q2) Bias Observed		(S3.Q3) Extent Reporting		(S3.Q4) Improve Fairness		(S4.Q4) Bias Mitigation	
Answers	*n (%)*	*Answers*	*n (%)*	*Answers*	*n (%)*	*Answers*	*n (%)*	*Answers*	*n (%)*
Sensitivity analysis	49 (67)	Accuracy gaps	32 (44)	None	3 (4)	Code sharing	45 (62)	Not at all	33 (45)
Demographic vars.	39 (53)	Data quality issues	29 (40)	Minimal	38 (52)	Algorithm availability	41 (56)	Minimal	20 (27)
Informal discussion	38 (52)	Sample diversity lim.	23 (32)	Moderate	17 (23)	Stakeholders involved	32 (44)	Moderate	6 (8)
Fairness metrics	10 (14)	Underrepresentation	21 (29)	Strong	9 (12)	Audit reports	7 (10)	Strong	13 (18)
External review	8 (11)	No bias observed	16 (22)	Full	1 (1)	Other[2]	5 (7)	Full	1 (1)
No measures	9 (12)	Other[1]	3 (4)	–	–	–	–	–	–
Base: N=73 (multiple answer)		*Base:* N=73 (multiple answer)		*Total:* 73 (100%)		*Base:* N=73 (multiple answer)		*Total:* 73 (100%)	

[1] "Other" (S3.Q2): not directly assessing fairness or limited relevance to population analyses. [2] "Other" (S3.Q4): references to internal meetings, peer-reviewed publications, or new research projects. Percentages are rounded and may not total exactly 100%. Abbreviations: vars., variables; lim., limits.

In (S4.Q3) Data Diversity, 63% reported moderate or stronger consideration of subgroup diversity, led by hospitals and public health researchers. In general, fairness considerations were more often implicit through stakeholder and data diversity than explicitly defined during design.

Table 5: RSQ4: Design, response distribution from survey items S4.Q1–S4.Q3.

(S4.Q1) Stakeholder Diversity		(S4.Q2) Fairness Defined		(S4.Q3) Data Diversity	
Answers	*n (%)*	*Answers*	*n (%)*	*Answers*	*n (%)*
Not at all	3 (4)	Not at all	10 (14)	Not at all	7 (10)
Minimal	23 (32)	Minimal	32 (44)	Minimal	20 (27)
Moderate	33 (45)	Moderate	25 (34)	Moderate	29 (40)
Strong	12 (16)	Strong	5 (7)	Strong	16 (22)
Fully	2 (3)	Fully	0 (0)	Fully	1 (1)
Total: 73 (100%)		*Total:* 73 (100%)		*Total:* 73 (100%)	

Percentages are rounded and may not total exactly 100%.

RSQ5. Application: How do researchers apply and justify trade-offs or interventions to operationalize fairness in practice?

In (S4.Q5) Fairness Monitoring, 81% reported no or minimal monitoring, and only 11% indicated moderate or stronger activity. In (S5.Q1) Action Ranking, top priorities were ensuring recommendations were not worse for any income group (mean rank 2.12) and equal across groups (2.74), followed by equal recommendations for similar health needs (2.94) and prioritizing urgent needs (3.21); updating recommendations ranked lower (4.08). In (S5.Q2) Accuracy Trade-off, 84% accepted some accuracy reduction for fairness, while 10% rejected trade-offs and 6% accepted a significant loss. In (S5.Q3) Bias Reporting, 72% described or proposed improvements to bias findings, while 26% implemented corrective actions. In general, respondents were open to modest accuracy–fairness trade-offs but rarely maintained systematic monitoring or corrective mechanisms.

Table 6: RSQ5: Application, response distribution from survey items S4.Q5, S5.Q1-S5.Q3.

(S4.Q5) Fairness Monitoring		(S5.Q1) Action Ranking		(S5.Q2) Accuracy Trade-off		(S5.Q3) Bias Reporting	
Answers	*n (%)*	*Top priorities (mean rank)*	*Score*	*Answers*	*n (%)*	*Answers*	*n (%)*
Not at all	41 (56)	Equal across groups	2.74	No reduction	7 (10)	No action	0 (0)
Minimal	18 (25)	No group worse off	2.12	Small	45 (62)	Acknowledge bias only	0 (0)
Moderate	6 (8)	Similar needs	2.94	Moderate	16 (22)	Suggest research	28 (38)
Strong	7 (10)	Prioritize urgent needs	3.21	Significant	4 (6)	Propose improvements	25 (34)
Fully	1 (1)	Update over time	4.08	Severe	0 (0)	Suggest corrections	11 (15)
–	–	Other (specify)	6.00	–	–	Immediate correction	8 (11)
–	–	–	–	–	–	–	–
Total: 73 (100%)		*Base: mean ranks per item (N = 64–68)*		*Total:* 72 (100%)		*Total:* 72 (100%)	

Percentages are rounded and may not total exactly 100%. Mean ranks calculated as weighted averages of rank positions (1–6; lower = higher priority).

5.2 Theoretical Embedding

Survey findings were systematically mapped onto three theories on the research–practice gap: the Knowledge–Practice Gap (the What), the KTA Cycle (the How), and the Knowing–Doing Gap (the Why). Figure 2 presents the resulting F2A framework, composed of three panels mirroring these perspectives. The left panel lists diagnostic questions about the gap's current state: what fairness knowledge exists, who holds it, where it is obtained, how it is accessed, and with what effect. The center panel shows KTA as two linked components—a knowledge-creation funnel (inquiry, synthesis, tools) and an action cycle (identify needs, select and adapt knowledge, assess barriers, implement, monitor, sustain). The right panel translates mechanisms from the Knowing–Doing Gap literature into diagnostic questions about why translation stalls, such as whether talk substitutes for action or accuracy metrics limit fairness. Horizontal connectors indicate flows from knowledge to action and feedback from sustainment to these explanatory mechanisms. Arrows are interpretive, depicting plausible feedbacks rather than causal claims. The framework offers an organizing scaffold consistent with observed patterns in ML-driven public health research and intended for use alongside implementation and governance guidance. Condensed mappings appear in the extended version; full matrices are available on request.

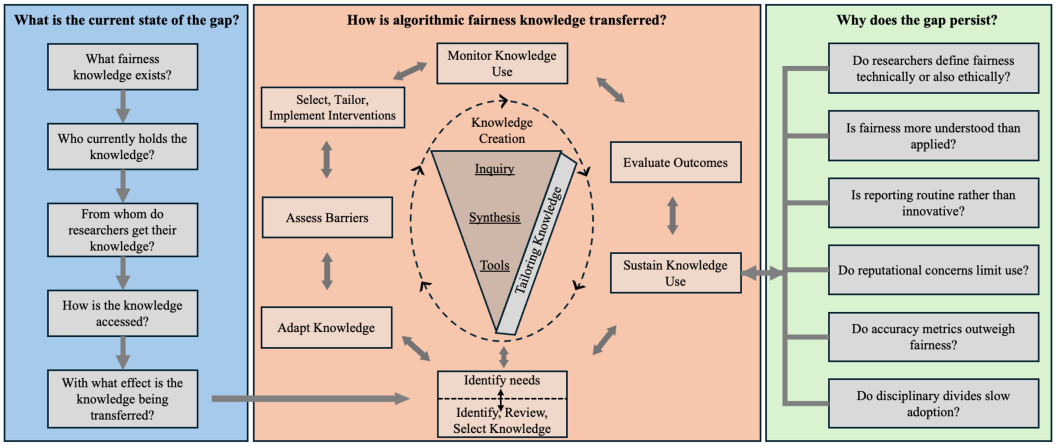

Figure 2: Fairness-to-Action (F2A) framework: analytical mapping of fairness translation across methodological (What), organizational (How), and systemic (Why) levels.

What is the current state of the gap? Regarding what fairness knowledge exists, respondents most often described fairness as equal treatment or protection of sensitive attributes, while smaller groups emphasized outcome-based definitions. Concerning who holds this knowledge, most reported limited understanding, with strong knowledge rare and particularly scarce in epidemiology and biostatistics, though higher levels appeared among data scientists and public health researchers. As to from whom researchers obtain it, dissemination relies mainly on external sources, while internal channels are weak or absent. With respect to how it is accessed, structured training opportunities are minimal, especially in universities and NGOs compared with hospitals. Finally, with what effect, reporting of algorithmic bias is typically minimal to moderate, with few comprehensive cases; reports often acknowledge bias or propose improvements rather than implement corrective actions. Acceptance of trade-offs between accuracy and fairness is widespread, though some respondents reject any accuracy reduction.

In summary, fairness knowledge exists but remains weakly institutionalized and not routine. It resides in individuals rather than organizations, depends on ad hoc external sources instead of structured training, and lacks feedback from reporting to remediation. Limitations are most pronounced where understanding is low, frameworks absent, and training missing.

How is algorithmic fairness knowledge transferred? In the knowledge-creation funnel, inquiry depends on external sources such as academic publications, conferences, and professional networks, while internal training channels remain weak. Synthesis is limited: most respondents reported minimal to moderate understanding, and framework awareness was common but often superficial, suggesting

limited systematic integration of fairness knowledge. As for tools and products, frameworks or guidelines were rarely applied, and practices such as fairness audits or subgroup evaluation were inconsistent and limited in scope, with few examples of strong corrective action.

In the action cycle, problem recognition is partial: bias is occasionally observed, yet fairness reporting is minimal to moderate, indicating that issues are noticed but seldom documented. Selection and adaptation of knowledge are uneven: framework awareness varies by field, bias assessment is inconsistent, and attention to data diversity remains moderate, leaving subgroups under-considered. Barriers include technical and regulatory obstacles, limited ML expertise, and lack of training, with these gaps most evident in university and NGO contexts. Interventions are irregular—checking models for bias or evaluating performance across subgroups is common, while other actions such as involving stakeholders, publishing transparency reports, or integrating fairness into training are less frequent. Monitoring mechanisms are minimal, follow-up actions uncommon, and evaluation often framed as accuracy–fairness trade-offs without subsequent corrective steps.

Taken together, fairness knowledge transfer remains fragmented: inquiry relies on external channels, synthesis and tool use are superficial, and the action cycle stalls at implementation and monitoring.

Why does the gap persist? Regarding whether knowing what to do is sufficient, fairness was primarily defined in technical terms, with project-level definitions most often minimal or moderate, indicating limited integration of ethical or equity framings. On whether talk substitutes for action, understanding outpaced implementation: bias assessment and improvement practices were inconsistently applied across groups. Regarding whether memory substitutes for thinking, reporting was generally minimal or moderate, and uptake of new improvement practices remained uneven beyond routine activity. Concerning whether fear prevents acting on knowledge, bias was recognized, yet responses were cautious: although many accepted small or moderate accuracy reductions, follow-through into corrective action was limited. On whether measurement obstructs good judgment, bias mitigation was mostly none or minimal, while accuracy continued to dominate decision making. Finally, on whether internal competition fragments collective effort, differences across primary fields and institution types in ML use, training, and reporting indicated fragmentation rather than shared practice.

Overall, these results help explain why the research-practice gap persists: definitions remain narrow, understanding exceeds implementation, reporting is routine rather than innovative, accuracy concerns limit corrective action, fairness metrics lag behind accuracy measures, and disciplinary and institutional divides constrain collective progress.

6 Discussion

This study examined how researchers in ML-driven public health in the Netherlands perceive and operationalize algorithmic fairness. Extending earlier literature-based assessments of fairness reporting [57], it provides an empirical view behind the scenes of how fairness is understood, where it is applied, and why translation often stalls within research practice.

Interpreted through the F2A framework, the findings reveal a multi-level challenge, with F2A serving as an evidence-informed diagnostic scaffold rather than a prescriptive or predictive model. At the methodological level (the What), fairness knowledge exists yet remains inconsistently defined, weakly institutionalized, and seldom formalized into design or reporting frameworks. At the organizational level (the How), knowledge flows primarily through external networks rather than internal structures, formal training is scarce, tools are applied unevenly and largely limited to subgroup checks or bias audits, and monitoring is minimal. At the systemic level (the Why), incentives continue to favor accuracy and throughput, reporting often substitutes for corrective action, and disciplinary as well as institutional divides fragment shared practice. Together, these conditions illustrate mechanisms that likely extend beyond the national setting, where fairness awareness often exceeds implementation.

The F2A framework helps explain where fairness translation falters: from defining and disseminating fairness knowledge to selecting and adapting it, from adaptation and barrier assessment to implementation, and from implementation and reporting to monitoring and sustained use. Each fragile transition reflects limited codification of fairness into design and reporting, weak organizational support, and incentives that undervalue iterative improvement. Rather than listing deficits, F2A highlights leverage points for strengthening translation: formalize definitions as design requirements,

build internal knowledge pipelines and training, link reporting to corrective action, and align incentives with continuous monitoring and learning. Advancing fairness in practice therefore requires coordinated interventions across methodological, organizational, and systemic levels so that fairness becomes routine, evaluation continuous, and accountability shared. Algorithmic fairness in public health thus emerges not only as a methodological issue but also as an institutional and governance challenge. Sustained progress depends on infrastructures and accountability systems that anchor fairness as standard research practice rather than as an ad hoc or externally driven task.

Recommendations. We prioritize researcher- and group-level actions as most immediately feasible and high impact, while institutional, funding, and governance interventions represent longer-term, system-level levers. The three highest-impact near-term actions are to formalize fairness definitions as explicit design requirements, strengthen internal knowledge pipelines and training, and link bias reporting to concrete corrective action. At the individual and group levels, researchers can normalize fairness assessment alongside accuracy and robustness, cultivate shared accountability, reward fairness artifacts, and embed monitoring into routine workflows. Journals, conferences, and funding bodies can reinforce these practices by requiring fairness reporting, including subgroup evaluation, bias audits, and impact statements, and by supporting shared benchmarking datasets. Institutions and governments can enable sustained uptake through training, governance mechanisms, oversight, and standards that connect reporting to remediation. Finally, affected communities should be involved so that fairness definitions and metrics reflect lived experience rather than abstract proxies. Coordinated action across these levels targets the fragile transitions identified in F2A, helping to close the research-practice gap and ensure that ML-driven public health operates safely, transparently, and equitably, through feasible and coordinated action.

7 Conclusion

This study extends earlier assessments of algorithmic fairness in ML-driven public health by directly engaging researchers to examine how fairness is perceived and operationalized in practice. At a glance, the findings reveal three interlinked layers of translation failure: methodological (fairness knowledge is individually held and weakly codified); organizational (knowledge transfer relies on external rather than internal structures); and systemic (accuracy incentives and fragmented governance impede sustained fairness practice). Framed through the F2A framework, these patterns highlight specific leverage points for strengthening translation: formalizing definitions as design requirements, developing internal knowledge pipelines and training, linking reporting to actionable remediation, and aligning incentives with ongoing monitoring and improvement. The contribution lies not in enumerating gaps but in providing a structured map of where interventions can anchor fairness as routine practice.

We acknowledge some limitations. The modest sample size limits statistical power for subgroup analyses, and self-reporting may introduce social desirability or recall bias. The national focus constrains generalizability, though the Netherlands offers a representative context for high-resource public health systems. In addition, simplifying complex definitions into single themes may obscure conceptual nuance. Finally, reported practices were not directly observed, so implementation could not be validated empirically.

Future work should address these limitations by validating fairness implementation through observed practice, tracing it across the ML lifecycle via case studies, and evaluating domain-specific metrics and mitigation methods. It should also test institutional levers (e.g., ethics review, funding criteria, training mandates) and examine how publication norms and governance shape reporting, develop lightweight tools for routine evaluation, and compare across countries to assess how structural conditions influence translation. These efforts can move the field from awareness to safer, sustained, accountable practice.

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

[1] Kurt Benke and Geza Benke. Artificial Intelligence and Big Data in Public Health. *International Journal of Environmental Research and Public Health*, 15(12):2796, December 2018.

[2] Jason Denzil Morgenstern, Emmalin Buajitti, Meghan O'Neill, Thomas Piggott, Vivek Goel, Daniel Fridman, Kathy Kornas, and Laura C Rosella. Predicting population health with machine learning: a scoping review. *BMJ Open*, 10(10):e037860, October 2020.

[3] Timothy L. Wiemken and Robert R. Kelley. Machine Learning in Epidemiology and Health Outcomes Research. *Annual Review of Public Health*, 41(1):21–36, April 2020.

[4] Vishwali Mhasawade, Yuan Zhao, and Rumi Chunara. Machine learning and algorithmic fairness in public and population health. *Nature Machine Intelligence*, 3(8):659–666, July 2021.

[5] David B. Olawade, Ojima J. Wada, Aanuoluwapo Clement David-Olawade, Edward Kunonga, Olawale Abaire, and Jonathan Ling. Using artificial intelligence to improve public health: a narrative review. *Frontiers in Public Health*, 11:1196397, October 2023.

[6] David Jungwirth and Daniela Haluza. Artificial Intelligence and Public Health: An Exploratory Study. *International Journal of Environmental Research and Public Health*, 20(5):4541, March 2023.

[7] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, October 2019.

[8] Richard Ribón Fletcher, Audace Nakeshimana, and Olusubomi Olubeko. Addressing Fairness, Bias, and Appropriate Use of Artificial Intelligence and Machine Learning in Global Health. *Frontiers in Artificial Intelligence*, 3:561802, April 2021.

[9] Lidia Flores, Seungjun Kim, and Sean D Young. Addressing bias in artificial intelligence for public health surveillance. *Journal of Medical Ethics*, 50(3):190–194, March 2024.

[10] Paul Wesson, Yulin Hswen, Gilmer Valdes, Kristefer Stojanovski, and Margaret A. Handley. 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, April 2022.

[11] Thomas C. Tsai, Sercan Arik, Benjamin H. Jacobson, Jinsung Yoon, Nate Yoder, Dario Sava, Margaret Mitchell, Garth Graham, and Tomas Pfister. Algorithmic fairness in pandemic forecasting: lessons from COVID-19. *npj Digital Medicine*, 5(1):59, May 2022.

[12] Marshall H. Chin, Nasim Afsar-Manesh, Arlene S. Bierman, Christine Chang, Caleb J. Colón-Rodríguez, Prashila Dullabh, Deborah Guadalupe Duran, Malika Fair, Tina Hernandez-Boussard, Maia Hightower, Anjali Jain, William B. Jordan, Stephen Konya, Roslyn Holliday Moore, Tamra Tyree Moore, Richard Rodriguez, Gauher Shaheen, Lynne Page Snyder, Mithuna Srinivasan, Craig A. Umscheid, and Lucila Ohno-Machado. 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, December 2023.

[13] Nicole M. Thomasian, Carsten Eickhoff, and Eli Y. Adashi. Advancing health equity with artificial intelligence. *Journal of Public Health Policy*, 42(4):602–611, December 2021.

[14] Lama H. Nazer, Razan Zatarah, Shai Waldrip, Janny Xue Chen Ke, Mira Moukheiber, Ashish K. Khanna, Rachel S. Hicklen, Lama Moukheiber, Dana Moukheiber, Haobo Ma, and Piyush Mathur. Bias in artificial intelligence algorithms and recommendations for mitigation. *PLOS Digital Health*, 2(6):e0000278, June 2023.

[15] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and Removing Disparate Impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 259–268, Sydney NSW Australia, August 2015. ACM.

[16] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of Opportunity in Supervised Learning. *arXiv:1610.02413 [cs]*, October 2016. arXiv: 1610.02413.

[17] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pages 214–226, Cambridge Massachusetts, January 2012. ACM.

[18] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press, 2023.

[19] Simon Caton and Christian Haas. Fairness in Machine Learning: A Survey. *ACM Computing Surveys*, 56(7):1–38, July 2024.

[20] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé, Miro Dudik, and Hanna Wallach. Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need? In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–16, Glasgow Scotland Uk, May 2019. ACM.

[21] Michael A. Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. 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*, pages 1–14, Honolulu HI USA, April 2020. ACM.

[22] Feng Chen, Liqin Wang, Julie Hong, Jiaqi Jiang, and Li Zhou. 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, April 2024.

[23] Dr Winston W Royce. Managing the Development of Large Software Systems. In *IEEE WESCON*, pages 328–338, Los Angeles, 1970.

[24] Michael Molenda. In Search of the Elusive ADDIE Model: Performance Improvement. *Performance Improvement*, 54(2):40–42, February 2015.

[25] Harini Suresh and John Guttag. A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–9, – NY USA, October 2021. ACM.

[26] Jeremy M Grimshaw, Martin P Eccles, John N Lavis, Sophie J Hill, and Janet E Squires. Knowledge translation of research findings. *Implementation Science*, 7(1):50, December 2012.

[27] Ian D. Graham, Jo Logan, Margaret B. Harrison, Sharon E. Straus, Jacqueline Tetroe, Wenda Caswell, and Nicole Robinson. Lost in knowledge translation: Time for a map? *Journal of Continuing Education in the Health Professions*, 26(1):13–24, 2006.

[28] Jeffrey Pfeffer and Robert I. Sutton. *The knowing-doing gap: how smart companies turn knowledge into action*. Harvard Business School Press, Boston, Mass, 2000.

[29] Umar Z. Ikram, Anton E. Kunst, Majda Lamkaddem, and Karien Stronks. The disease burden across different ethnic groups in Amsterdam, the Netherlands, 2011–2030. *European Journal of Public Health*, 24(4):600–605, August 2014.

[30] M. Kroneman, W. Boerma, M. van den Berg, P. Groenewegen, J. de Jong, and E. van Ginneken. *The Netherlands: Health System Review*, volume 18 of *Health Systems in Transition*. European Observatory on Health Systems and Policies, Copenhagen, 2016.

[31] Onaedo Ilozumba, Tirsah S Koster, Elena V Syurina, and Ikenna Ebuenyi. Ethnic minority experiences of mental health services in the Netherlands: an exploratory study. *BMC Research Notes*, 15(1):266, July 2022.

[32] Carl Thomas Berdahl, Lawrence Baker, Sean Mann, Osonde Osoba, and Federico Girosi. Strategies to Improve the Impact of Artificial Intelligence on Health Equity: Scoping Review. *JMIR AI*, 2:e42936, February 2023.

[33] Laleh Seyyed-Kalantari, Haoran Zhang, Matthew B. A. McDermott, Irene Y. Chen, and Marzyeh Ghassemi. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature Medicine*, 27(12):2176–2182, December 2021.

[34] John R. Zech, Marcus A. Badgeley, Manway Liu, Anthony B. Costa, Joseph J. Titano, and Eric Karl Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLOS Medicine*, 15(11):e1002683, November 2018.

[35] Agostina J. Larrazabal, Nicolás Nieto, Victoria Peterson, Diego H. Milone, and Enzo Ferrante. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences*, 117(23):12592–12594, June 2020.

[36] Darshali A. Vyas, Leo G. Eisenstein, and David S. Jones. Hidden in Plain Sight: Reconsidering the Use of Race Correction in Clinical Algorithms. *The New England Journal of Medicine*, 383(9):874–882, August 2020.

[37] Irene Y. Chen, Emma Pierson, Sherri Rose, Shalmali Joshi, Kadija Ferryman, and Marzyeh Ghassemi. Ethical Machine Learning in Healthcare. *Annual Review of Biomedical Data Science*, 4(1):123–144, July 2021.

[38] Thomas Grote and Geoff Keeling. On Algorithmic Fairness in Medical Practice. *Cambridge Quarterly of Healthcare Ethics*, 31(1):83–94, January 2022.

[39] Samantha Cruz Rivera, Xiaoxuan Liu, An-Wen Chan, Alastair K Denniston, Melanie J Calvert, Hutan Ashrafian, Andrew L Beam, Gary S Collins, Ara Darzi, Jonathan J Deeks, M Khair ElZarrad, Cyrus Espinoza, Andre Esteva, Livia Faes, Lavinia Ferrante Di Ruffano, John Fletcher, Robert Golub, Hugh Harvey, Charlotte Haug, Christopher Holmes, Adrian Jonas, Pearse A Keane, Christopher J Kelly, Aaron Y Lee, Cecilia S Lee, Elaine Manna, James Matcham, Melissa McCradden, David Moher, Joao Monteiro, Cynthia Mulrow, Luke Oakden-Rayner, Dina Paltoo, Maria Beatrice Panico, Gary Price, Samuel Rowley, Richard Savage, Rupa Sarkar, Sebastian J Vollmer, and Christopher Yau. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *The Lancet Digital Health*, 2(10):e549–e560, October 2020.

[40] Xiaoxuan Liu, Samantha Cruz Rivera, David Moher, Melanie J Calvert, Alastair K Denniston, Hutan Ashrafian, Andrew L Beam, An-Wen Chan, Gary S Collins, Ara DarziJonathan J Deeks, M Khair ElZarrad, Cyrus Espinoza, Andre Esteva, Livia Faes, Lavinia Ferrante Di Ruffano, John Fletcher, Robert Golub, Hugh Harvey, Charlotte Haug, Christopher Holmes, Adrian Jonas, Pearse A Keane, Christopher J Kelly, Aaron Y Lee, Cecilia S Lee, Elaine Manna, James Matcham, Melissa McCradden, Joao Monteiro, Cynthia Mulrow, Luke Oakden-Rayner, Dina Paltoo, Maria Beatrice Panico, Gary Price, Samuel Rowley, Richard Savage, Rupa Sarkar, Sebastian J Vollmer, and Christopher Yau. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *The Lancet Digital Health*, 2(10):e537–e548, October 2020.

[41] Tina Hernandez-Boussard, Selen Bozkurt, John P A Ioannidis, and Nigam H Shah. MINIMAR (MINimum Information for Medical AI Reporting): Developing reporting standards for artificial intelligence in health care. *Journal of the American Medical Informatics Association*, 27(12):2011–2015, December 2020.

[42] Gary S Collins, Karel G M Moons, Paula Dhiman, Richard D Riley, Andrew L Beam, Ben Van Calster, Marzyeh Ghassemi, Xiaoxuan Liu, Johannes B Reitsma, Maarten Van Smeden, Anne-Laure Boulesteix, Jennifer Catherine Camaradou, Leo Anthony Celi, Spiros Denaxas, Alastair K Denniston, Ben Glocker, Robert M Golub, Hugh Harvey, Georg Heinze, Michael M Hoffman, André Pascal Kengne, Emily Lam, Naomi Lee, Elizabeth W Loder, Lena Maier-Hein, Bilal A Mateen, Melissa D McCradden, Lauren Oakden-Rayner, Johan Ordish, Richard Parnell, Sherri Rose, Karandeep Singh, Laure Wynants, and Patricia Logullo. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*, page e078378, April 2024.

[43] Karel G M Moons, Johanna A A Damen, Tabea Kaul, Lotty Hooft, Constanza Andaur Navarro, Paula Dhiman, Andrew L Beam, Ben Van Calster, Leo Anthony Celi, Spiros Denaxas, Alastair K Denniston, Marzyeh Ghassemi, Georg Heinze, André Pascal Kengne, Lena Maier-Hein, Xiaoxuan Liu, Patricia Logullo, Melissa D McCradden, Nan Liu, Lauren Oakden-Rayner, Karandeep Singh, Daniel S Ting, Laure Wynants, Bada Yang, Johannes B Reitsma, Richard D Riley, Gary S Collins, and Maarten Van Smeden. PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. *BMJ*, page e082505, March 2025.

[44] Judy Wawira Gichoya, Liam G McCoy, Leo Anthony Celi, and Marzyeh Ghassemi. Equity in essence: a call for operationalising fairness in machine learning for healthcare. *BMJ Health & Care Informatics*, 28(1):e100289, April 2021.

[45] Laura Sikstrom, Marta M Maslej, Katrina Hui, Zoe Findlay, Daniel Z Buchman, and Sean L Hill. Conceptualising fairness: three pillars for medical algorithms and health equity. *BMJ Health & Care Informatics*, 29(1):e100459, January 2022.

[46] Melissa D McCradden, Shalmali Joshi, Mjaye Mazwi, and James A Anderson. Ethical limitations of algorithmic fairness solutions in health care machine learning. *The Lancet Digital Health*, 2(5):e221–e223, May 2020.

[47] Dimitra Panteli, Keyrellous Adib, Stefan Buttigieg, Francisco Goiana-da Silva, Katharina Ladewig, Natasha Azzopardi-Muscat, Josep Figueras, David Novillo-Ortiz, and Martin McKee. Artificial intelligence in public health: promises, challenges, and an agenda for policy makers and public health institutions. *The Lancet Public Health*, 10(5):e428–e432, May 2025.

[48] Banu Aysolmaz, Rudolf Müller, and Darian Meacham. The public perceptions of algorithmic decision-making systems: Results from a large-scale survey. *Telematics and Informatics*, 79:101954, April 2023.

[49] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 220–229, Atlanta GA USA, January 2019. ACM.

[50] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, December 2021.

[51] Carina Nina Vorisek, Caroline Stellmach, Paula Josephine Mayer, Sophie Anne Ines Klopfenstein, Dominik Martin Bures, Anke Diehl, Maike Henningsen, Kerstin Ritter, and Sylvia Thun. Artificial Intelligence Bias in Health Care: Web-Based Survey. *Journal of Medical Internet Research*, 25:e41089, June 2023.

[52] Christopher Starke, Janine Baleis, Birte Keller, and Frank Marcinkowski. Fairness perceptions of algorithmic decision-making: A systematic review of the empirical literature. *Big Data & Society*, 9(2):20539517221115189, July 2022.

[53] Max Kasun, Katie Ryan, Jodi Paik, Kyle Lane-McKinley, Laura Bodin Dunn, Laura Weiss Roberts, and Jane Paik Kim. Academic machine learning researchers' ethical perspectives on algorithm development for health care: a qualitative study. *Journal of the American Medical Informatics Association*, 31(3):563–573, February 2024.

[54] Jee Young Kim, Alifia Hasan, Katherine C. Kellogg, William Ratliff, Sara G. Murray, Harini Suresh, Alexandra Valladares, Keo Shaw, Danny Tobey, David E. Vidal, Mark A. Lifson, Manesh Patel, Inioluwa Deborah Raji, Michael Gao, William Knechtle, Linda Tang, Suresh Balu, and Mark P. Sendak. Development and preliminary testing of Health Equity Across the AI Lifecycle (HEAAL): A framework for healthcare delivery organizations to mitigate the risk of AI solutions worsening health inequities. *PLOS Digital Health*, 3(5):e0000390, May 2024.

[55] Dutch Government. Algorithm register (algoritmeregister), 2023. Accessed 9 June 2025.

[56] Statistics Netherlands (CBS). CBS introduces new population classification by origin, 2022. Accessed 9 June 2025.

[57] Sara Altamirano, Arjan Vreeken, and Sennay Ghebreab. Machine Learning and Public Health: Identifying and Mitigating Algorithmic Bias through a Systematic Review. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES 2025)*, New York, NY, USA, 2025. Association for Computing Machinery (ACM).

[58] Fixsen, D. L., Naoom, S. F., Blase, K. A., Friedman, R. M. & Wallace, F. Implementation Research: A Synthesis of the Literature. Technical report, University of South Florida, Louis de la Parte Florida Mental Health Institute, National Implementation Research Network, Tampa, FL, 2005.

[59] Trisha Greenhalgh, Glenn Robert, Fraser Macfarlane, Paul Bate, and Olivia Kyriakidou. Diffusion of Innovations in Service Organizations: Systematic Review and Recommendations. *The Milbank Quarterly*, 82(4):581–629, December 2004.

[60] Enola Proctor, Hiie Silmere, Ramesh Raghavan, Peter Hovmand, Greg Aarons, Alicia Bunger, Richard Griffey, and Melissa Hensley. Outcomes for Implementation Research: Conceptual Distinctions, Measurement Challenges, and Research Agenda. *Administration and Policy in Mental Health and Mental Health Services Research*, 38(2):65–76, March 2011.

[61] Sanne Peters, Krithika Sukumar, Sophie Blanchard, Akilesh Ramasamy, Jennifer Malinowski, Pamela Ginex, Emily Senerth, Marleen Corremans, Zachary Munn, Tamara Kredo, Lucia Prieto Remon, Etienne Ngeh, Lisa Kalman, Samia Alhabib, Yasser Sami Amer, and Anna Gagliardi. Trends in guideline implementation: an updated scoping review. *Implementation Science*, 17:50, December 2022.

[62] Laura J Damschroder, David C Aron, Rosalind E Keith, Susan R Kirsh, Jeffery A Alexander, and Julie C Lowery. Fostering implementation of health services research findings into practice: a consolidated framework for advancing implementation science. *Implementation Science*, 4(1):50, December 2009.

[63] Centers for Disease Control and Prevention. Applying the Knowledge to Action (K2A) Framework: Questions to Guide Planning. *Centers for Disease Control and Prevention, US Dept of Health and Human Services*, 2014.

[64] Kilo, Charles M. A framework for collaborative improvement: lessons from the Institute for Healthcare Improvement's Breakthrough Series. *Quality Management in Healthcare*, 6(4):1–14, 1998.

[65] Noah Ivers, Sharlini Yogasingam, Meagan Lacroix, Kevin A. Brown, Jesmin Antony, Charlene Soobiah, Michelle Simeoni, Thomas A. Willis, Jacob Crawshaw, Vivi Antonopoulou, Carly Meyer, Nathan M. Solbak, Brenna J. Murray, Emily-Ann Butler, Simone Lepage, Martina Giltenane, Mary D. Carter, Guillaume Fontaine, Michael Sykes, Michael Halasy, Abdalla Bazazo, Samantha Seaton, Tony Canavan, Sarah Alderson, Catherine Reis, Stefanie Linklater, Aislinn Lalor, Ashley Fletcher, Emma Gearon, Hazel Jenkins, Jason A. Wallis, Liesl Grobler, Lisa Beccaria, Sheila Cyril, Tomas Rozbroj, Jia Xi Han, Alice Xt Xu, Kelly Wu, Geneviève Rouleau, Maryam Shah, Kristin Konnyu, Heather Colquhoun, Justin Presseau, Denise O'Connor, Fabiana Lorencatto, and Jeremy M. Grimshaw. Audit and feedback: effects on professional practice. *Cochrane Database of Systematic Reviews*, 2025(3):CD000259, March 2025. Art. No.: CD000259.

[66] John W. Creswell and Cheryl N. Poth. *Qualitative inquiry & research design: choosing among five approaches*. SAGE, Los Angeles, fourth edition edition, 2018.

[67] Karim Lekadir, Richard Osuala, Catherine Gallin, Noussair Lazrak, Kaisar Kushibar, Gianna Tsakou, Susanna Aussó, Leonor Cerdá Alberich, Kostas Marias, Manolis Tsiknakis, Sara Colantonio, Nickolas Papanikolaou, Zohaib Salahuddin, Henry C. Woodruff, Philippe Lambin, and Luis Martí-Bonmatí. FUTURE-AI: Guiding Principles and Consensus Recommendations for Trustworthy Artificial Intelligence in Medical Imaging, July 2024. arXiv:2109.09658 [cs].

[68] Akash Sharma, Nguyen Tran Minh Duc, Tai Luu Lam Thang, Nguyen Hai Nam, Sze Jia Ng, Kirellos Said Abbas, Nguyen Tien Huy, Ana Marušić, Christine L. Paul, Janette Kwok, Juntra Karbwang, Chiara De Waure, Frances J. Drummond, Yoshiyuki Kizawa, Erik Taal, Joeri Vermeulen, Gillian H. M. Lee, Adam Gyedu, Kien Gia To, Martin L. Verra, Évelyne M. Jacqz-Aigrain, Wouter K. G. Leclercq, Simo T. Salminen, Cathy Donald Sherbourne, Barbara Mintzes, Sergi Lozano, Ulrich S. Tran, Mitsuaki Matsui, and Mohammad Karamouzian. A Consensus-Based Checklist for Reporting of Survey Studies (CROSS). *Journal of General Internal Medicine*, 36(10):3179–3187, October 2021.

[69] Gunther Eysenbach. Improving the Quality of Web Surveys: The Checklist for Reporting Results of Internet E-Surveys (CHERRIES). *Journal of Medical Internet Research*, 6(3):e34, September 2004.