Invisible by Design: Women's Health As the Blind Spot in AI and Medicine

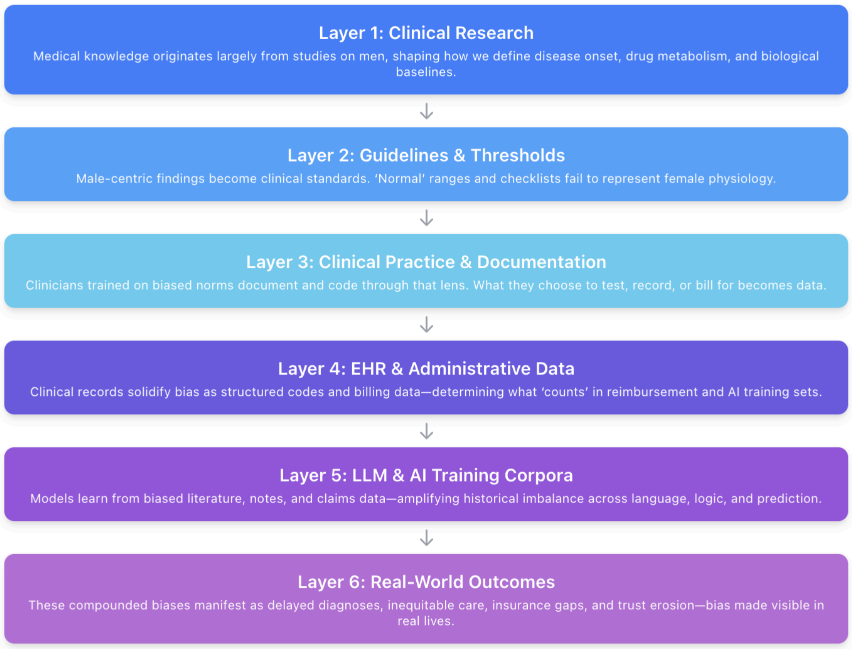
INTRODUCTION

Healthcare is racing to adopt artificial intelligence, **2.2 times** faster than the rest of the economy.[1] Yet at the very moment medicine promises "precision," its foundation remains profoundly imprecise. The flaw isn't in the code or the Electronic Health Record. It starts much earlier, in what — and who — we chose to study.

For decades, clinical research has centered on the male body as the default. Women were excluded from trials on the grounds of "biological variability," and when included, their data were often underpowered or averaged away. As a result, the scientific literature (the first layer of evidence that defines what is "normal") was built on male physiology.

From this flawed foundation, a **cascade** of distortion follows:

The Cascade of Bias: From Evidence to Outcome



Layer 1: Clinical Research

Medical knowledge originates largely from studies on men, shaping how we define disease onset, drug metabolism, and biological baselines.

Layer 2: Guidelines & Thresholds

Male-centric findings become clinical standards. 'Normal' ranges and checklists fail to represent female physiology.

Layer 3: Clinical Practice & Documentation

Clinicians trained on biased norms document and code through that lens. What they choose to test, record, or bill for becomes data.

Layer 4: EHR & Administrative Data

Clinical records solidify bias as structured codes and billing data—determining what 'counts' in reimbursement and AI training sets.

Layer 5: LLM & AI Training Corpora

Models learn from biased literature, notes, and claims data—amplifying historical imbalance across language, logic, and prediction.

Layer 6: Real-World Outcomes

These compounded biases manifest as delayed diagnoses, inequitable care, insurance gaps, and trust erosion—bias made visible in real lives.

Bias compounds through each layer—transforming evidence into policy, policy into data, and data into inequity. What seems like precision is often history's distortion, digitized.

1. **Clinical Research (Layer 1):** The majority of medical knowledge still originates from studies conducted primarily on men. These studies shape our understanding of disease onset, drug metabolism, and biological baselines.

2. **Guidelines and Diagnostic Thresholds (Layer 2):** Those male-centric findings become codified into practice guidelines and reference ranges. "Normal" lab values, diagnostic cutoffs, and symptom

[1] HIT Consultant. (2025, October 22). Healthcare AI adoption is 2.2x faster than the broader economy. https://hitconsultant.net/2025/10/22/healthcare-ai-adoption-is-2-2x-faster-than-the-broader-economy

checklists often fail to capture female presentations. A woman's "abnormal" may still fall within the male-defined "normal."

3. **Clinical Practice and Documentation (Layer 3):** Clinicians trained on these guidelines record what they see through that same lens. Their decisions (what gets tested, coded, or believed) populate the EHR.

4. **EHR and Administrative Data (Layer 4):** The EHR then becomes the de facto "ground truth" for modern AI. Yet EHRs are not neutral. They are the residue of human judgment. Studies show systematic miscoding, under-documentation, and diagnostic delay for conditions that disproportionately affect women.[2]

5. **LLM Training Corpora (Layer 5):** Before any medical fine-tuning, virtually all modern LLMs are built on trillions of tokens of general web data (e.g., Common Crawl, Wikipedia, digitized books, news articles). This initial phase establishes the model's core language understanding, reasoning, and its initial representation of the world, including fundamental social biases (gender, race, etc.) that are inherent in internet and literary text. This phase establishes the base layer of bias before any medical data is introduced. LLMs are then fine-tuned on a patchwork of medical and scientific data sources. These specialized corpora inherit and amplify the bias established in the foundational layer, as each source is downstream of the same structural imbalance, such as:

    - The biomedical research corpus (the journals, abstracts, and textbooks indexed in PubMed) is a major training source for most medical language models. But this literature reflects decades of male-dominant study design. Across fields from neuroscience to cardiology, men have historically comprised the majority of research subjects, often two-thirds or more in cardiovascular studies. Because these texts define what is considered *normal physiology* and *typical disease presentation*, models trained on them inevitably internalize those same assumptions. The result is not a coding error but a continuation of epistemic bias: if the evidence base itself underrepresents women, so will the language patterns and clinical associations the model learns.

    - Clinical Notes and De-identified EHRs: Many clinical LLMs incorporate narrative notes and de-identified patient records from hospitals and

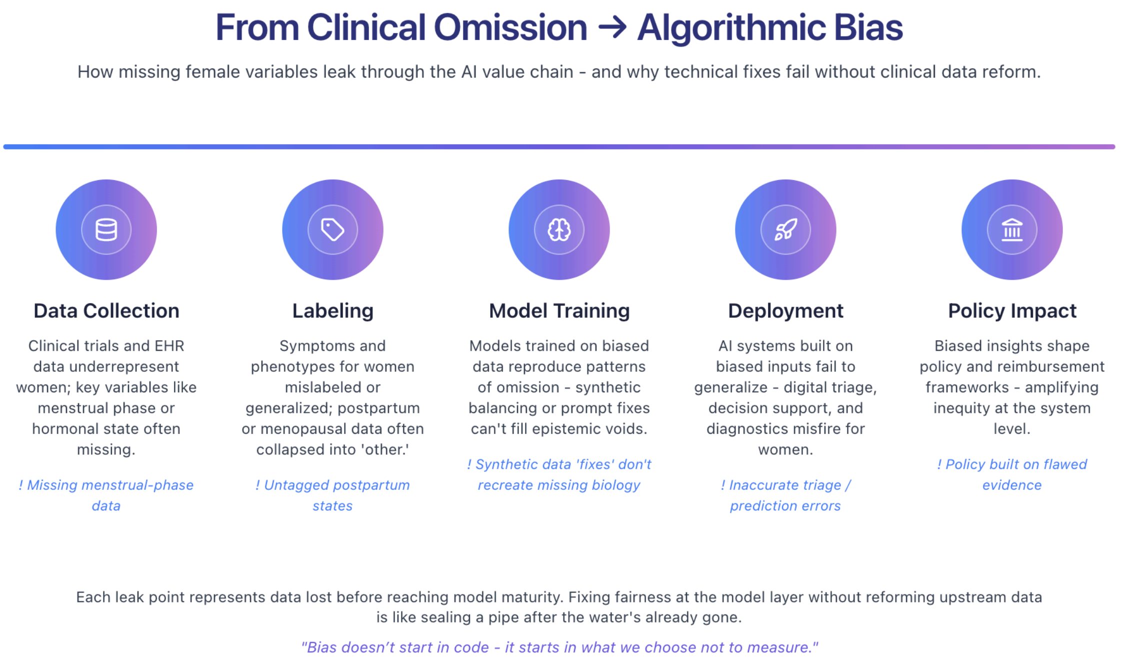
[2] Shah, N. H., Milstein, A., & Bagley, S. C. (2019). Making machine learning models clinically useful. *JAMA*, 322(14), 1351-1352. https://pmc.ncbi.nlm.nih.gov/articles/PMC11046491/

research networks. These records inherit upstream bias: which conditions were coded, which symptoms were deemed significant, and whose pain or experience was documented.

- Medical Guidelines and Reference Databases: Models trained or fine-tuned on clinical guidelines (e.g., UpToDate, SNOMED, ICD, FDA labeling data) replicate the thresholds and diagnostic norms derived from male-centric trials.
- Claims and Administrative Data: Predictive models built for cost, utilization, or risk-scoring often rely on claims data, an additional layer where bias compounds, because what is reimbursed shapes what is recorded.
- Drug and Device Databases: From pharmacovigilance datasets to device registries, women are both under-represented and over-penalized: fewer safety data points, higher rates of adverse effects, and delayed detection of risk.

When AI models are trained on this foundation, they do not correct the bias, they **amplify it.** This is the algorithmic feedback loop: biased inputs → biased outputs → clinician reinforcement → new biased inputs.

The effect is measurable. AI systems trained on misrepresentative data have been shown to **reduce diagnostic accuracy by 11.3 percentage points** compared with baseline clinical performance.[3][^3] This phenomenon (automation bias or overreliance) occurs when



From Clinical Omission → Algorithmic Bias

How missing female variables leak through the AI value chain - and why technical fixes fail without clinical data reform.

Data Collection

Clinical trials and EHR data underrepresent women; key variables like menstrual phase or hormonal state often missing.

! Missing menstrual-phase data

Labeling

Symptoms and phenotypes for women mislabeled or generalized; postpartum or menopausal data often collapsed into 'other.'

! Untagged postpartum states

Model Training

Models trained on biased data reproduce patterns of omission - synthetic balancing or prompt fixes can't fill epistemic voids.

! Synthetic data 'fixes' don't recreate missing biology

Deployment

AI systems built on biased inputs fail to generalize - digital triage, decision support, and diagnostics misfire for women.

! Inaccurate triage / prediction errors

Policy Impact

Biased insights shape policy and reimbursement frameworks - amplifying inequity at the system level.

! Policy built on flawed evidence

Each leak point represents data lost before reaching model maturity. Fixing fairness at the model layer without reforming upstream data is like sealing a pipe after the water's already gone.

"Bias doesn't start in code - it starts in what we choose not to measure."

in

clinicians defer to algorithmic output even when it conflicts with their own correct judgment. Instead of supporting clinicians, these systems codify past human error into machine-level certainty.

Efforts to mitigate this through Explainable AI (XAI), which aims to clarify how models reach conclusions, failed: providing explanations alongside biased outputs (so clinicians would hopefully see the reasoning and draw their own conclusions / be wary of the outputs of the model) resulted in the same 11.3-point drop in clinician accuracy.

If the data foundation is biased, every layer built on it inherits that distortion. Only by verifying the integrity of ground truth can AI improve medicine rather than institutionalize its errors.

This paper traces the entire cascade, from exclusion in research to distortion in data, and shows how the omission of women at the beginning of medicine's knowledge chain is now being embedded into the infrastructure of future healthcare. More importantly, it proposes a path to rebuild the ground truth itself: through technical, regulatory, and research reforms that ensure AI does not merely reproduce the past, but finally learns from it.

PART I: THE FOUNDATION IS BROKEN

Layer 1: The Male Default in Clinical Research

Medical research has long defaulted to male physiology, minimizing female inclusion and sex-specific analysis, shaping current care.

Historical exclusion: The U.S. Food and Drug Administration banned women of child-bearing potential from Phase I/II trials until 1993,[4] which means the majority of modern medicine was created without a female baseline. To this day, women remain under-represented in early-phase trials. Female representation averaged just 37% in broadly-inclusive randomized controlled trials, with three-quarters of studies reporting no sex-stratified outcomes[5]. A 2022 analysis of 1,433 U.S. trials for major conditions found women comprised only ~41% of enrollments despite constituting roughly half of relevant patient populations.[6]

The preclinical gaps run deeper. Only 36.5% of cell cultures in cancer research had sex annotation at all.[7] When sex was reported, 71% of in-vitro studies used only male cells.[^8] In fields like **Neuroscience**, single-sex studies using male animals **outnumber those using females by a ratio of approximately 5.5 to 1**, aka 85% of single-sex studies were conducted exclusively on males.[8]

The consequences show up in the real world. Of 86 drugs analyzed, 76 exhibited higher pharmacokinetic values in women, differences that strongly predicted higher adverse drug reaction rates.[9] Women now experience adverse drug reactions at nearly twice the rate of men[10]. Yet post-market safety surveillance rarely disaggregates data by sex, meaning safety signals unique to women can be missed entirely.

[4] Applied Clinical Trials. (n.d.). Gender bias in the clinical evaluation of drugs. https://www.appliedclinicaltrialsonline.com/view/gender-bias-in-the-clinical-evaluation-of-drugs

[5] Zucker, I., & Beery, A. K. (2010). Males still dominate animal studies. *Nature*, 465(7299), 690. https://pmc.ncbi.nlm.nih.gov/articles/PMC8812498/

[6] Sosinsky, A., Agrawal, R., Gray, S. W., & Freedman, R. A. (2022). The evolution of clinical trial eligibility criteria and enrollment of women. *American Journal of Clinical Oncology*, 45(10), 421-426. https://www.sciencedirect.com/science/article/abs/pii/S1551714422000441

[7] Hao, Y., Gong, R., Li, T., Cheng, Y., & Wang, Y. (2020). Sex annotation in publicly available cancer genomic datasets. *Scientific Data*, 7(1), 250. https://pubmed.ncbi.nlm.nih.gov/38498336/

[8] Zucker, I., & Beery, A. K. (2010). Males still dominate animal studies. *Nature*, 465(7299), 690. https://pubmed.ncbi.nlm.nih.gov/20535186/

[9] Campesi, I., Franconi, F., & Seghieri, G. (2018). Sex-gender-related therapeutic approaches for cardiovascular disease. *Pharmacological Research*, 132, 130-137. https://pmc.ncbi.nlm.nih.gov/articles/PMC7275616/

[10] Harvard Gazette. (2023, December). Women more likely to suffer drug side effects, but reason may not be biology.https://news.harvard.edu/gazette/story/2023/12/women-more-likely-to-suffer-drug-side-effects-but-reason-may-not-be-biology/

When foundational research populations are skewed, the resulting norms, thresholds, and decision-rules derived from them mis-specify women's presentations. This male-default effect means women are implicitly treated as deviations from the "norm".

The male-default bias extends beyond clinical trials into computational biology itself. Google's recent cell2sentence project represents an extraordinary technical advance: it converts single-cell RNA sequencing profiles into "sentences" that allow language models to read cellular biology as text.[11] This approach could open a new era of AI-driven discovery.

But in its public documentation, the dataset sources list tissue type, species, and disease status, not the sex of the cells used. That absence matters because every human cell has a sex. Each carries either XX or XY chromosome complement and retains sex-linked differences in gene expression, regulation, and metabolism, even outside reproductive tissues.[12] [13]

If the majority of training cells come from male sources or sex metadata is absent entirely, the resulting embeddings may encode a male-biased biological baseline. That bias can propagate downstream as these embeddings are used to train AI systems for drug discovery, diagnosis, and precision medicine.

Studies show that male and female cells can differ in their response to stress and drugs,[14] and as noted earlier, that women experience adverse drug reactions at roughly twice the rate of men. Training AI on sex-unbalanced data could reproduce these inequities algorithmically, hard-coding male-default biology into systems designed to represent all humans.

Layer 2: The Male Default in Clinical Guidelines & Diagnostic Thresholds

A recent *npj Digital Medicine* review highlights how algorithmic bias can emerge when biological sex differences are overlooked in clinical prediction models. In gastroenterology and hepatology, for instance, the Model for End-Stage Liver Disease (MELD) score, used to prioritize patients for liver transplantation, relies on serum creatinine as a marker of renal function. Because women typically have lower baseline creatinine levels than men, the

[11] Huang, K., Altosaar, J., & Ranganath, R. (2024). Cell2Sentence: Teaching large language models the language of biology. *Nature Communications*, 15(1), 9715. https://pmc.ncbi.nlm.nih.gov/articles/PMC11565894/

[12] Wachs, D., Yao, Y., & Veeraraghavan, A. (2019). Sex-specific gene expression in mammalian cells. *Genome Biology*, 20(1), 202. https://pmc.ncbi.nlm.nih.gov/articles/PMC7898458/

[13] Oliva, M., Muñoz-Aguirre, M., Kim-Hellmuth, S., et al. (2020). The impact of sex on gene expression across human tissues. *Science*, 369(6509), eaba3066. https://pubmed.ncbi.nlm.nih.gov/32913072/

[14] Zucker, I., & Prendergast, B. J. (2020). Sex differences in pharmacokinetics predict adverse drug reactions in women. *Biology of Sex Differences*, 11(1), 32. https://pubmed.ncbi.nlm.nih.gov/32503637/

Why Women's Heart Attacks Get Missed — Fix the Thresholds

model systematically underestimates disease severity in female patients, thereby reducing their likelihood of receiving a transplant under equivalent clinical conditions. As the authors note, this reflects how "AI bias can arise when sex differences in clinical predictors are overlooked" and illustrates a broader challenge in clinical AI fairness: algorithms that apply uniform reference standards across sexes risk perpetuating inequities embedded in historical data rather than correcting them.[15]

Even when medical science does offer sex-specific guidelines, such as distinct diagnostic thresholds for women, those insights too often disappear before reaching the clinic or remain absent from algorithmic decision tools. This disconnect directly harms women.

Standard cardiac troponin thresholds used to detect heart attacks were historically calibrated on male cohorts. Women experiencing myocardial infarction can present with troponin levels below these "universal" cut-offs, leading to delayed or missed diagnoses.

In the landmark High-STEACS trial, introducing a high-sensitivity cardiac troponin I assay with sex-specific thresholds resulted in a 42% increase in identified myocardial injury in women, versus just 6% in men.[16] Yet despite higher detection:

- Women continued to receive fewer treatments (angiography, dual-antiplatelet therapy, statins, and beta-blockers, compared to men).[17]

- One-year outcomes did not improve for women (adjusted HR 1.11; 95% CI 0.92-1.33) compared with men (adjusted HR 0.85; 95% CI 0.71-1.01).

- The persistence of treatment disparity indicated threshold change alone did not shift clinician behavior.

[15] Webb, E., Shah, N., Veselkov, K., Rare Disease Working Group, & Cheng, F. (2025). Sex-specific considerations in clinical AI fairness: A review of gastroenterology and hepatology algorithms. *NPJ Digital Medicine*, 8(1), 67. https://www.nature.com/articles/s41746-025-01667-2

[16] Chapman, A. R., et al. (2019). High-sensitivity cardiac troponin and the diagnosis of myocardial infarction in patients with kidney disease. *Circulation*, 140(6), 423-435. https://pubmed.ncbi.nlm.nih.gov/31623760/

[17] Chapman, A. R., et al. (2019). High-sensitivity cardiac troponin and sex-disaggregated outcomes. *BMC Medicine*, 17(1), 213. https://pmc.ncbi.nlm.nih.gov/articles/PMC6876271/

Layer 3: Bias In Clinical Practice and Documentation

In the words of the study authors, "Use of sex-specific thresholds identified 5 times more additional women than men with myocardial injury. Despite this increase, women received approximately one-half the number of treatments for coronary artery disease as men, and outcomes were not improved."

So, clinicians flagged more women, but did **not** act on the new data. In practice, many teams continued to rely on conventional, higher troponin thresholds, effectively bypassing the study's recommended female-specific cut-offs. This meant that women who exceeded the female-specific threshold but not the older standard remained "undetected" and untreated.

Implementation without accountability: Some hospitals updated the assay to high-sensitivity versions but failed to update protocols, no automatic cardiology consults, no audit dashboards, no structured feedback loops. Without these structural supports, clinician behavior and institutional protocols didn't change.

The beta-blocker case:

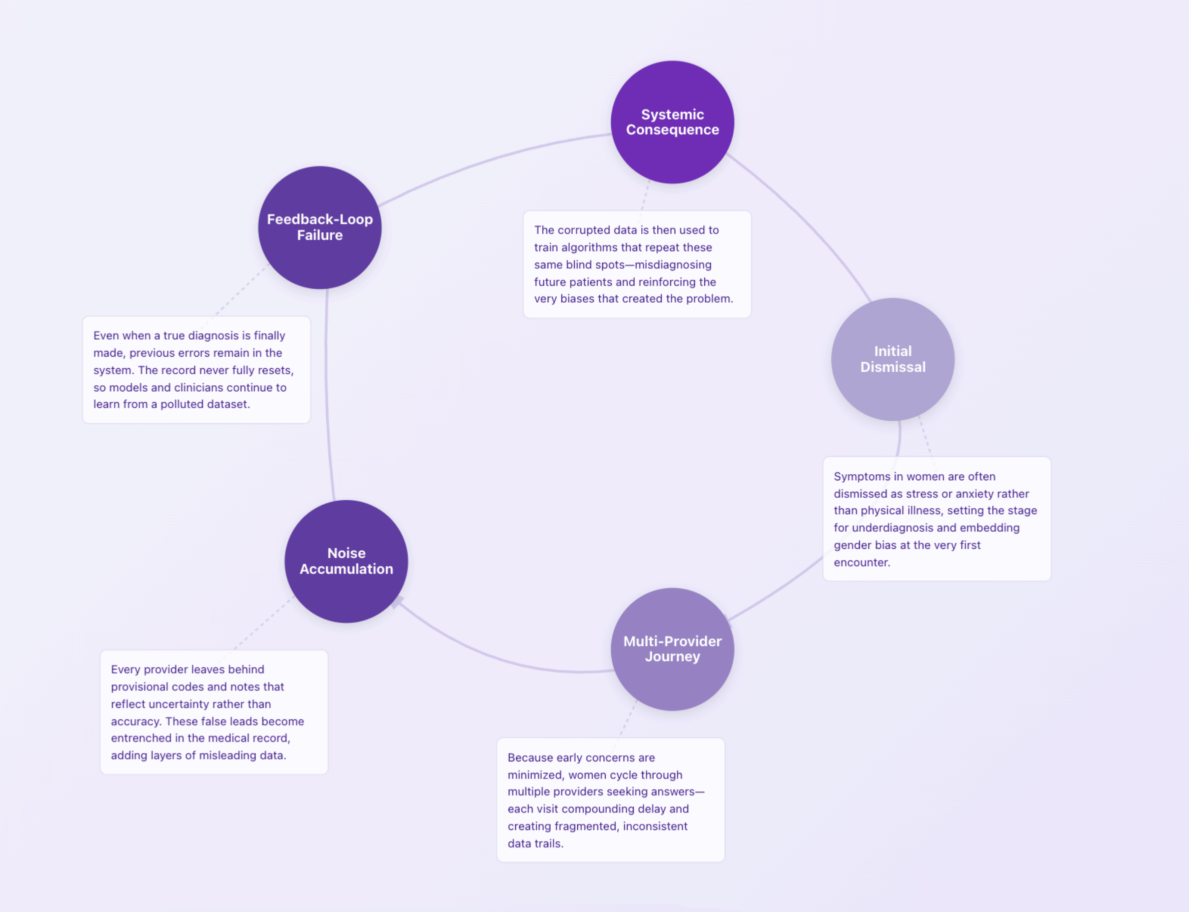
For years, beta-blocker medication was standard post-heart attack practice, based primarily on male evidence. The 2024 REBOOT Heart Trial found this long-standing therapy increased mortality in women but not in men.[18] What was deemed "best practice", derived from male-centric evidence, was actively dangerous for women. Current clinical AI tools, learning from these ingrained guidelines, would dutifully recommend the drug to women, never recognizing the sex-specific danger.

[18] Lee, M. S., Park, H., Woo, J. S., et al. (2024). Beta-blocker therapy in heart failure patients: The REBOOT Heart Trial. *The Lancet*, 403(10429), 819-829. https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(24)00244-6/fulltext

PART II: HOW BIAS BECOMES CODE

Many chronic conditions primarily affecting women (endometriosis, chronic fatigue syndrome, certain forms of Alzheimer's disease) suffer from major research and data gaps. This lack of reliable evidence creates a chain reaction: incorrect or delayed diagnoses lead to inaccurate data, which weakens every AI model trained on it.

A large-scale population study of 6.9 million patients across 770 diseases illustrates the pattern: on average women are diagnosed four years later than men, including 2.5 years later for cancer and 4.5 years later for diabetes.[19] This is not because disease appears later in women, but because early symptoms are more likely to be misinterpreted, minimized, or documented under less accurate labels.



The corrupted data is then used to train algorithms that repeat these same blind spots—misdiagnosing future patients and reinforcing the very biases that created the problem.

Even when a true diagnosis is finally made, previous errors remain in the system. The record never fully resets, so models and clinicians continue to learn from a polluted dataset.

Symptoms in women are often dismissed as stress or anxiety rather than physical illness, setting the stage for underdiagnosis and embedding gender bias at the very first encounter.

Every provider leaves behind provisional codes and notes that reflect uncertainty rather than accuracy. These false leads become entrenched in the medical record, adding layers of misleading data.

Because early concerns are minimized, women cycle through multiple providers seeking answers—each visit compounding delay and creating fragmented, inconsistent data trails.

Endometriosis as case study:

[19] Kjaergaard, J., Arfwedson Wang, C. E., & Waterloo, K. (2019). A study of gender differences in diagnostic delay of mental disorders. *BMC Psychiatry*, 19(1), 117. https://pubmed.ncbi.nlm.nih.gov/30737381/

On average, women wait seven years to receive a diagnosis for endometriosis.[20][^22] During those years, their records fill with incorrect codes (pelvic pain, irritable bowel syndrome, anxiety) instead of the real disease. When the correct label finally appears, it's tied to the wrong point in time. When an AI system learns from that data, it learns the pattern of human error: late-stage disease, not the early biological signal we'd want it to detect.

This is misclassification bias, when the recorded diagnosis doesn't match reality, often because data entry follows billing or administrative rules rather than clinical certainty. Women are especially affected because their symptoms are more likely to be dismissed, described vaguely, or recorded using male-centric criteria. The EHR reflects the behavior of the health system, its habits and blind spots, more than the actual course of disease.[21]

Layer 4: EHR and Administrative Data: Corrupted Ground Truth: The Data Problem

EHRs change over time as diagnoses are updated or deleted. Most systems don't track these edits properly. Without time-stamped corrections, models can't distinguish old errors from verified information.[22] This creates hidden technical debt, problems buried in data that quietly undermine every model trained on it.

The diagnostic friction feedback loop:

Initial dismissal: Women's symptoms are more likely to be labeled as stress- or anxiety-related rather than physical in origin. This reflects systemic bias in medical training and diagnostic guidelines, which still center the male presentation of disease as the "default.[23]

The multi-provider journey: Because their symptoms remain unexplained, women often move through multiple providers and specialties before receiving an accurate diagnosis. For **endometriosis**, the average diagnostic delay is around **seven years** and typically involves consultations with **five or more clinicians**.

[20] University of York. (2024). Diagnosis endometriosis delay. https://www.york.ac.uk/news-and-events/news/2024/research/diagnosis-endometriosis-delay/

[21] Shah, P., Kendall, F., Khozin, S., et al. (2019). Artificial intelligence and machine learning in clinical development: A translational perspective. *NPJ Digital Medicine*, 2(1), 69. https://pmc.ncbi.nlm.nih.gov/articles/PMC10938158/

[22] Zhang, J., Whebell, S., & Gallifant, J. (2022). Temporal aspects of electronic health record data and algorithmic fairness. *JAMIA Open*, 5(4), ooac099. https://pmc.ncbi.nlm.nih.gov/articles/PMC9759969/

[23] American Medical Women's Association. (n.d.). Lost in translation: When women's health is called mental health and vice versa. https://amwa-doc.org/lost-in-translation-when-womens-health-is-called-mental-health-and-vice-versa/

Noise accumulation: Each provider documents their own provisional impression (codes like *"non-specific abdominal pain"* or *"generalized anxiety disorder."*) Because ICD codes are used for billing as much as for clinical reasoning, these provisional or inaccurate labels stay in the system.

Feedback-loop failure: When a specialist eventually confirms the true diagnosis, through imaging, biopsy, or surgical confirmation, it is simply appended to the existing record. The prior years of erroneous codes are rarely removed or corrected, and most EHR systems **lack mechanisms for longitudinal error reconciliation**. As a result, historical misdiagnoses persist in the data, misleading both clinicians and algorithms trained on those records.

In this loop, data ceases to describe disease. It describes the behavior of the healthcare system itself: its delays, omissions, and prejudices.

Layer 5: When AI Learns Human Error

Artificial intelligence is often described as a way to eliminate human bias from medicine. In reality, most AI systems don't reason, they recognize patterns in historical data. When that data is incomplete or unbalanced, the model learns and repeats the distortions that shaped them. Instead of correcting inequities, it scales them.

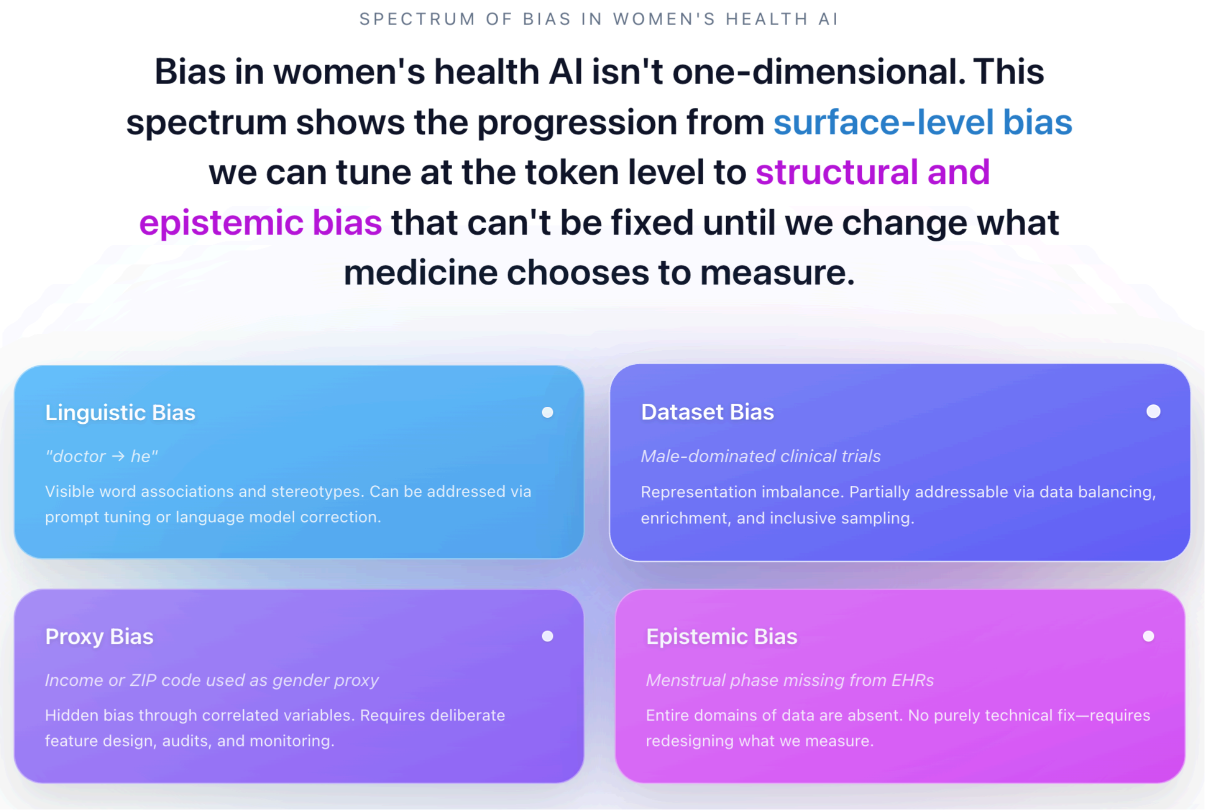
Large language models and shallow reasoning:

Recent work on large language models suggests the same pattern applies to systems that generate reasoning rather than predictions. In the 2024 preprint *Systematic Characterization of the Effectiveness of Alignment in Large Language Models for Categorical Decisions*, researchers tested how three advanced LLMs (GPT-4o, Claude 3.5 Sonnet, and Gemini Advanced) handled medical triage decisions.[24]

[24] American Medical Women's Association. (n.d.). Lost in translation: When women's health is called mental health and vice versa. https://amwa-doc.org/lost-in-translation-when-womens-health-is-called-mental-health-and-vice-versa/

The findings revealed striking inconsistency. All three models invoked the same ethical language ("favoring the worst-off") but applied it differently. GPT-4o defined "worst-off" as the patient who would gain the most life-years, Claude 3.5 chose the most acutely ill, and Gemini Advanced prioritized the oldest or frailest. These differences show the models were not reasoning ethically in any human sense. They were reconstructing familiar moral phrases from training data and using them as linguistic templates.

This shallow reasoning has consequences for women's health. Because much of medicine's data and language developed through male-referenced research and clinical norms, AI systems trained on those sources inherit the same blind spots. When an algorithm prioritizes "sickest first," it may over-value immediately measurable risks, e.g. heart attacks, and undervalue chronic, cyclical, or pain-dominant conditions disproportionately affecting women: endometriosis, autoimmune disorders. When an algorithm is optimized to "maximize total benefit," it rewards what it can measure. Patients whose recovery potential is easily quantified, e.g. those with standard biomarkers, predictable trajectories, or well-documented conditions, rise to the top. Those whose outcomes are harder to capture in data, such as women with chronic pain, autoimmune disorders, or multifactorial symptoms, quietly fall through the cracks.



SPECTRUM OF BIAS IN WOMEN'S HEALTH AI

Bias in women's health AI isn't one-dimensional. This spectrum shows the progression from surface-level bias we can tune at the token level to structural and epistemic bias that can't be fixed until we change what medicine chooses to measure.

Linguistic Bias

"doctor → he"

Visible word associations and stereotypes. Can be addressed via prompt tuning or language model correction.

Dataset Bias

Male-dominated clinical trials

Representation imbalance. Partially addressable via data balancing, enrichment, and inclusive sampling.

Proxy Bias

Income or ZIP code used as gender proxy

Hidden bias through correlated variables. Requires deliberate feature design, audits, and monitoring.

Epistemic Bias

Menstrual phase missing from EHRs

Entire domains of data are absent. No purely technical fix—requires redesigning what we measure.

The deeper problem is epistemic asymmetry: a gap between what the model knows how to represent and what the data never recorded. Large language models can imitate empathy through tone, what might be called linguistic empathy, but they lack epistemic empathy, the ability to recognize and reason about realities missing from their data.

Efforts to "align" models often amplify this illusion. Alignment training can make a system sound fairer without making it *think* fairer. In a recent triage evaluation, GPT-4o showed slight improvement after alignment, yet Claude 3.5 and Gemini Advanced actually diverged, producing outputs more consistent with clinician phrasing but less consistent with clinical equity. In other words, fine-tuning models to mimic professional language can create the appearance of alignment while deepening the underlying bias.

The LSE social care study:

In 2025, a landmark study by the London School of Economics and Political Science's Care Policy & Evaluation Centre (CPEC) (lead author Sam Rickman) investigated gender bias in large language models (LLMs) used for adult social care documentation.

The study used real case notes from 617 adult long-term care users in a London local authority. Researchers created gender-swapped versions of each case, then had four different models generate summaries. Each summary pair differed only by gender.

The models tested included:

- Gemma (Google) – a state-of-the-art LLM released in 2024.
- Llama 3 (Meta) – another leading 2024-generation model.
- Benchmark models from ~2019 (Google's T5, Meta's BART) for comparison.[25]

Key findings:

- Gemma showed **significant gender-based differences** in how case needs were described: men's summaries were far more likely to include terms like *"complex medical history," "disabled," "unable to access the community"* compared to the identical sister cases labelled as women.[26]

[25] Rickman, S., Bohnet, H., Hogan, S., et al. (2025). Gender bias in large language models for adult social care documentation. *BMC Medical Informatics and Decision Making*, 25(1), 118. https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-025-03118-0

[26] The Guardian. (2025, August 11). AI tools used by English councils downplay women's health issues, study finds.https://www.theguardian.com/technology/2025/aug/11/ai-tools-used-by-english-councils-downplay-womens-health-issues-study-finds

- In contrast, Llama 3 **did not show measurable gender disparities** in these metrics.[27]

- Because social-care services are allocated based on perceived need, the framing of a person's needs in case notes—affected by the summary's language—can directly influence the amount or type of support they receive. Gemma's bias thus risks allocating less care to women with the same conditions as men [28].

- The study emphasises that **bias in these systems is not inevitable**, but instead depends on design choices: dataset composition, model architecture/training, and objective functions.

This is classic alignment failure: the AI optimized to produce "believable" notes based on historically biased records, thereby faithfully replicating historical underestimation of women's needs. The danger is not just biased predictions, but institutionalization of bias as "truth."

This also occurred in a 2023 *JAMA Network Open* study by Kim et al. at Stanford University [29] which tested whether AI chatbots reproduce known gender and racial biases in medical decision-making. The researchers fed 19 standardized clinical vignettes, spanning cardiology, emergency medicine, rheumatology, and dermatology, into ChatGPT-4 and Google Bard. Each case was identical except for the patient's gender, race/ethnicity, or socioeconomic status, and the chatbots' answers were compared to earlier clinician responses from published studies designed to reveal bias.

In one vignette about coronary artery disease, both AI systems, like physicians, were more likely to suggest the diagnosis for men than for women, even when symptoms were identical. In a second case on thrombolysis for a heart attack, ChatGPT recommended treatment for White men only, omitting women and minority patients entirely. When asked about advanced heart failure, both chatbots recommended aggressive therapies (like ventricular assist devices) for men but were inconsistent or withheld recommendations for women, particularly Hispanic women. In dermatology, both systems suggested isotretinoin (a gold-standard acne drug) for men far more often than for women or transgender patients.

[27] Rickman, S., Bohnet, H., Hogan, S., et al. (2025). Gender bias in large language models for adult social care documentation. *BMC Medical Informatics and Decision Making*, 25(1), 118. https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-025-03118-0

[28] London School of Economics. (2025). AI tools risk downplaying women's health needs in social care. https://www.lse.ac.uk/news/latest-news-from-lse/ai-tools-risk-downplaying-womens-health-needs-in-social-care

[29] Kim, P. W., Xie, S., Huang, M. K., Aguirre-Chang, G., & Chow, D. S. (2023). Race and sex bias in AI medical diagnosis: A study of ChatGPT and Google Bard responses. *JAMA Network Open*, 6(11), e2342343. https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2810775

The takeaway is that even the most advanced AI models can replicate gendered patterns of care seen in human clinicians: diagnosing men more readily, offering them more aggressive treatments, and showing greater uncertainty or omission for women and gender-diverse patients. These biases likely stem from imbalanced training data and the underrepresentation of female cases in clinical research.

Quantifying the Performance Gap

AI's performance reflects the data it sees—and what it doesn't.

Foundation models reveal the gap. Modern health foundation models inherit not just statistical patterns but the priorities and omissions of the data they are built on. Delphi-2M, a 2025 GPT-style foundation model, was trained on longitudinal ICD-10 event data from around 400,000 UK Biobank participants and validated on 1.9 million individuals in the Danish National Patient Registry. It delivers strong average performance, predicting the risk and timing of more than 1,000 diseases years in advance [30].

But when you look more closely at the underlying data, a familiar asymmetry appears. UK Biobank and similar registries contain **tens of thousands** of recorded cases for common cardiometabolic diseases such as type 2 diabetes (over 19,000 individuals with type 2 diabetes in one UK Biobank analysis alone) yet only a **few thousand** recorded cases of endometriosis.[31]

This imbalance isn't explained by biology. At the population level, diabetes and endometriosis affect a similar number of women: about 10% of adult women for diabetes, and roughly 10% of women of reproductive age for endometriosis.[32] These are conditions of the same general magnitude—common, chronic, and highly relevant to women's health. If clinical data reflected real-world prevalence, the volume of training data for each should be in the same ballpark. Instead, the data available for endometriosis is nearly ten times smaller. This gap does not come from the condition being rare; it comes from systemic under-diagnosis, years-long diagnostic delays, and inconsistent or incomplete documentation of female-specific disorders in health records: the very inputs Delphi-2M learns from.[33]

[30] Nature. (2025). Delphi-2M: A GPT-style foundation model for longitudinal disease prediction. *Nature*, 637, 155-163. https://www.nature.com/articles/s41586-025-09529-3

[31] UK Biobank. (n.d.). Evaluating the incidence of complications among people with diabetes according to age of onset: Findings from the UK Biobank. https://www.ukbiobank.ac.uk/publications/evaluating-the-incidence-of-complications-among-people-with-diabetes-according-to-age-of-onset-findings-from-the-uk-biobank/

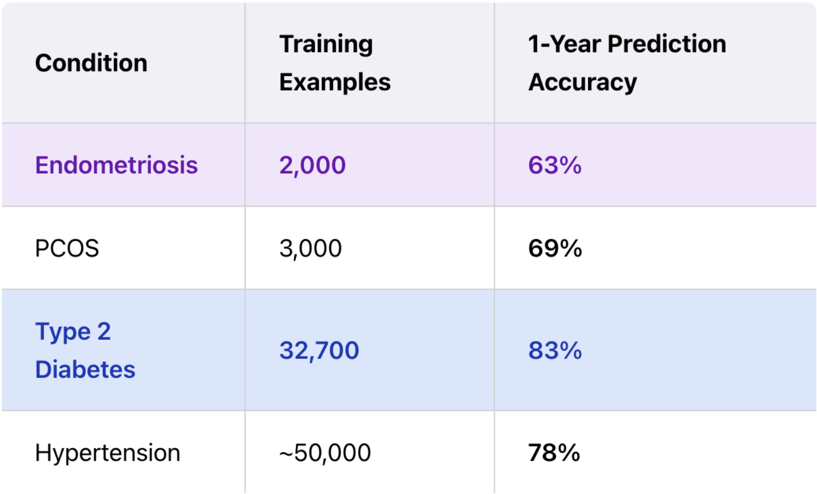
[32] International Diabetes Federation. (n.d.). Diabetes facts & figures. https://idf.org/about-diabetes/diabetes-facts-figures/

[33] PMC. (n.d.). Endometriosis prevalence and diagnosis patterns. https://pmc.ncbi.nlm.nih.gov/articles/PMC9127440/

A model like Delphi-2M does not "decide" to care less about endometriosis; it simply mirrors the evidence it is given. When the training corpus encodes women's diseases as rare events and cardiometabolic diseases as ubiquitous, those imbalances become part of the model's internal map of risk. Foundation models therefore make the upstream problem legible: *even when the architecture is cutting-edge and the aggregate metrics are excellent, the representation of women's conditions is still constrained by the narrow window through which those conditions were ever recorded.*

Not coincidentally, model accuracy for women-centric diseases was much lower:

Prediction Performance by Condition



Condition	Training Examples	1-Year Prediction Accuracy
Endometriosis	2,000	63%
PCOS	3,000	69%
Type 2 Diabetes	32,700	83%
Hypertension	~50,000	78%

Lower sample size conditions show weaker predictive performance — not because they are inherently harder to model, but because the system was never trained on them.

When AI models lack sufficient data for a specific group, they default to population averages, which in healthcare often means male physiology. This is reference misalignment.

The lower accuracy for female-specific conditions indicates a predictable statistical outcome of data imbalance: when certain conditions appear far less often or are recorded with greater variability, model performance declines accordingly. The accuracy gap functions as a diagnostic of omission, showing where the underlying data ecosystem reflects uneven medical attention and coding precision.

Delphi-2M illustrates a broader trend. Even when modern models appear balanced and perform well overall, sex-based performance gaps persist. Bias in healthcare AI stems from *how disease is represented*, not only from who is in the dataset.

In a landmark study, Straw, Rees, and Nachev (2024) examined cardiac disease prediction algorithms trained on data balanced equally between men and women. Despite equal representation, the models were consistently less accurate for women: across sixteen independent experiments, false-negative rates were higher for female patients, meaning the algorithms were more likely to miss disease in women, even with equivalent data volume and quality. The disparities, significant in thirteen of the sixteen tests (ranging from −17.8 to −3.4 percentage points), persisted even after rebalancing or feature adjustments.[34]

The reason lies in how disease is represented and labeled, not in how much data there is. Cardiovascular disease manifests differently in women: symptoms such as fatigue, shortness of breath, or nausea are more diffuse and less tied to the male-pattern benchmarks, that are deemed "classic" indicators (chest pain, ST-segment changes, sharply elevated troponin) that dominate diagnostic criteria and training data. Because clinical labels were historically assigned using those male-pattern benchmarks, women's cases were more likely to be underdiagnosed or misclassified in the source data. When algorithms learn from that record, they implicitly treat male-typical patterns as the canonical "signal" of disease and female-typical ones as statistical noise. Even when both sexes are equally represented numerically, the model still optimizes for what it has learned to recognize most confidently: the male-coded expression of illness. The outcome is not random error but systematic under-detection: women's disease fits less cleanly into the model's learned boundaries of pathology. What looks like technical parity conceals a diagnostic asymmetry. The algorithm "sees" disease through a lens medicine itself has long shaped around the male body.

This was illustrated in another 2024 study published in the *Journal of Biomedical Informatics,* [35] where researchers from Dedalus Healthcare and Ruhr University Bochum examined whether hospital AI models predict health risks equally well for men and women. The team tested three machine learning–based clinical risk models (predicting delirium, sepsis, and acute kidney injury (AKI)) across two German hospitals: Medius Klinik Nürtingen, a general hospital, and the Herz- und Diabeteszentrum Nordrhein-Westfalen (HDZ NRW), a major cardiology and diabetes center. They found that female patients had fewer recorded data points (such as lab results, vital signs, and medication records) and were diagnosed with these conditions less frequently, providing the models with less information to learn from. As a result, the AI systems were more likely to miss women who were actually at risk, particularly at HDZ NRW, where male-pattern cardiology data dominated and model accuracy was consistently lower for women.

A complementary study by Chung et al. (2021) reached the same conclusion from a different angle. Using over 5,000 COVID-19 patient records, researchers trained one model solely on male data and another solely on female data. Each model performed well within its own sex, but accuracy collapsed when tested on the other: for instance, the

[34] Straw, I., Rees, J., & Nachev, P. (2024). Sex-based disparities in machine learning models for cardiac disease prediction. *European Heart Journal - Digital Health*, 5(5), 567-576. https://pmc.ncbi.nlm.nih.gov/articles/PMC11384168/

[35] Zhou, Y., Wang, L., Tang, L., et al. (2024). Sex-based performance disparities in clinical risk prediction models at German hospitals. *Journal of Biomedical Informatics*, 154, 104639. https://www.sciencedirect.com/science/article/pii/S1532046424001102

male-trained model's accuracy dropped from 0.92 to 0.86 and AUC from 0.97 to 0.94 when applied to female cases. The reverse produced similar losses. These results reveal that models do not simply learn "disease"; they learn *sex-specific feature distributions*, distinct internal representations of what illness looks like in men and in women. These results demonstrate algorithms learn feature distributions that are sex-specific and don't transfer cleanly, effectively encoding different "representations of illness" for men and women.[36][^37]

The pattern extends across organ systems and data types. A 2022 study [37][^38] tested several machine-learning algorithms designed to predict liver disease and found that, while overall accuracy was acceptable (for example, logistic regression ~ 71.31 % ± 2.37 SD and SVM ~ 79.40 % ± 2.50 SD), women experienced markedly higher false-negative rates than men. Specifically, the random-forest classifier had a female false-negative rate that was 21.02 percentage points worse than that for men, and logistic-regression showed a 24.07 percentage-point disparity. This means the models were far more likely to *miss* disease in women.

A 2024 *arXiv* study, *Slicing Through Bias* (Olesen et al.), examined how performance disparities arise in medical image analysis even when data appear balanced. Using large public chest X-ray datasets (NIH-CXR14 and CheXpert) the researchers introduced a method called *Slice Discovery* to identify underperforming subgroups within models trained to detect pneumothorax and atelectasis. They found that apparent sex-based accuracy gaps could be traced not to sample imbalance, but to **shortcut learning**: the models were relying on contextual, non-pathological cues such as chest drains and ECG wires as proxies for disease. These artifacts were more common in certain patient subgroups and differed in frequency by sex, inadvertently creating performance disparities.[38]

This matters because it shows that medical AI can inherit bias even when representation is numerically equal. The models were not truly "seeing" lung pathology, they were learning correlations embedded in the practice of care itself: which patients receive particular interventions, how devices are positioned, and how imaging is performed. When such procedural patterns differ by sex, the algorithm's definition of disease becomes entangled with those differences. The study reveals a deeper layer of bias in clinical AI: not just who is in the dataset, but what hidden contextual signals the model learns to trust.

[36] Chung, K., Yoo, H., Lee, J., et al. (2021). Sex-specific prediction model for severe COVID-19 using machine learning. *Journal of Personalized Medicine*, 11(11), 1190. https://pmc.ncbi.nlm.nih.gov/articles/PMC8667070/

[37] Lu, H., Uddin, S., Hajati, F., Moni, M. A., & Khushi, M. (2022). A patient network-based machine learning model for disease prediction: The case of type 2 diabetes. *Applied Intelligence*, 52(3), 2411-2422. https://pubmed.ncbi.nlm.nih.gov/35470133/

[38] Olesen, T. B., Leibig, C., & Lauritzen, A. D. (2024). Slicing through bias: Explaining performance gaps in medical image analysis. *arXiv preprint arXiv:2406.12142*. https://arxiv.org/html/2406.12142v2

The Missing Variables Problem

These quantitative disparities trace back to how medical data themselves are structured. The statistical imbalance in predictive models reflects a deeper informational architecture: what medicine chooses to record, and what it leaves invisible.

Most AI models train on coded data (diagnoses, laboratory results, billing fields) but cannot "see" life-stage information not explicitly recorded. Factors like pregnancy status, postpartum recovery, perimenopausal transition, or contraceptive use profoundly alter diagnostic meaning. High blood pressure carries very different implications during pregnancy than outside it.

If a factor like "menopausal stage" isn't a structured input, the model is effectively blind to its clinical significance. An algorithm predicting cardiovascular risk for a 45-year-old woman won't factor in that she's in perimenopause, even though this is clinically critical. Many EHRs lack dedicated fields for these variables, and AI developers cannot incorporate what isn't there. The result: AI that relies solely on male-centric physiological baselines, cementing systemic omission in care delivery.

PART III: FALSE SOLUTIONS & FEEDBACK LOOPS

3.1 Why Fairness Patches Aren't Enough

The traditional approach to AI bias has been retrospective "fairness" audits: checking model outputs for bias and tweaking algorithms. But with a gender data gap, bias is wired in at a deeper level. This isn't about tuning a model to avoid sexist language or adjusting a threshold. We're facing epistemic bias: bias in what the model knows (or doesn't know), due to gaps in the training data and knowledge base.

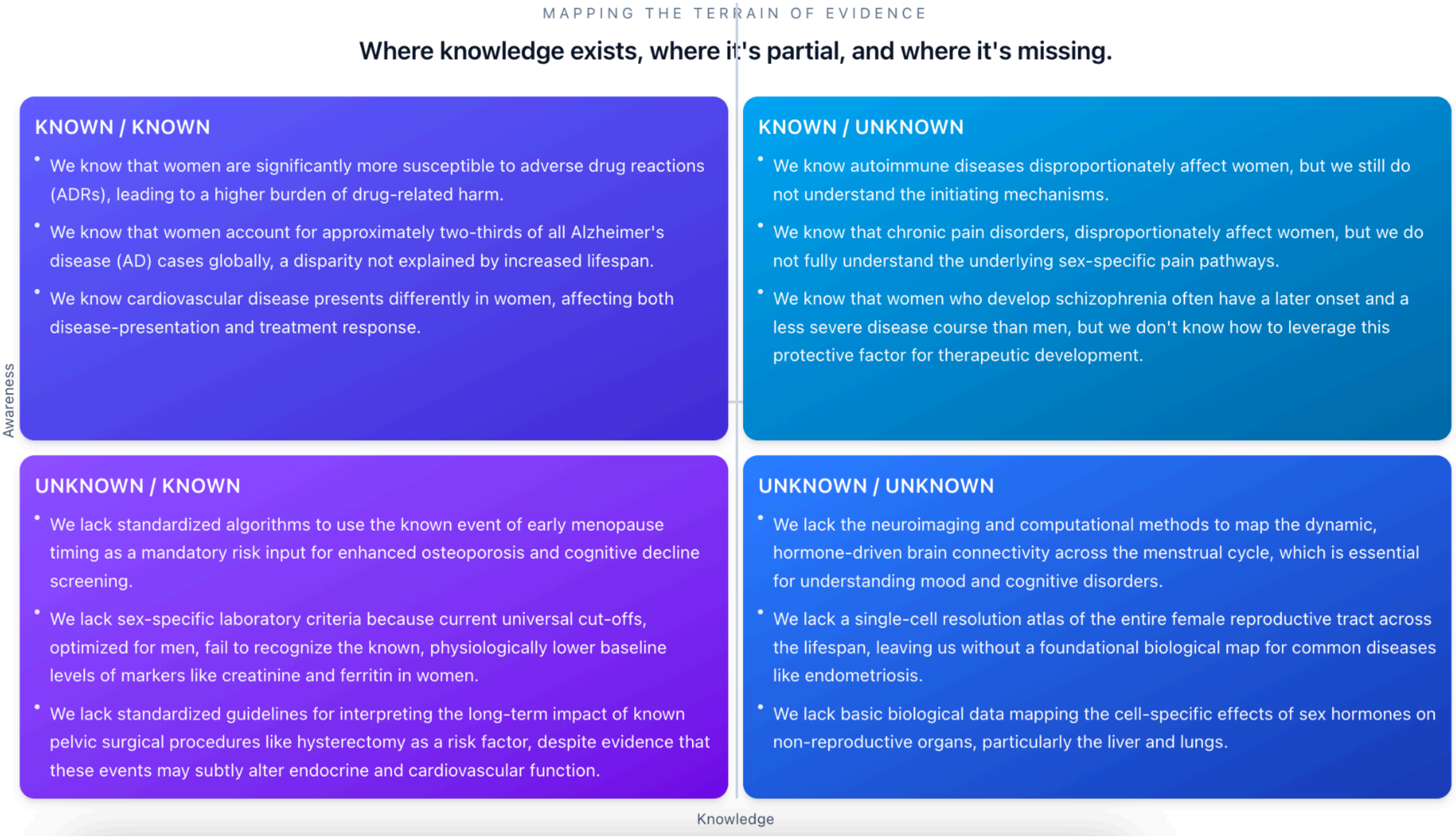
Many bias mitigation techniques (anonymizing gender in data, enforcing equalized odds in outputs) treat the symptom, not the cause. They don't equip the model with new knowledge of female physiology; they only mask or adjust predictions post hoc. The danger is creating false security: an ostensibly "fair" model that still performs poorly for women because the underlying medical logic isn't there.

The illusion of algorithmic neutrality persists despite evidence to the contrary. Anonymization doesn't solve knowledge gaps. When female-specific patterns are absent from training data, no amount of post-processing can recover that information. To truly fix the problem, we need to align the model with medical reality by supplying the right inputs and structure from the start.

3.2 The Synthetic Data Trap

When used carefully, synthetic data can be a useful tool. It has shown promise in mitigating bias under certain conditions. In natural language processing, counterfactual data augmentation (CDA): generating altered versions of text by swapping gendered terms, has reduced gender bias in model outputs.[39] In computer vision, generating synthetic images to equalize demographics has shown success.

These use-cases illustrate synthetic data, if generated with the right constraints, can be part of a bias mitigation toolkit. However, synthetic data tends to be most effective as a supplement rather than replacement for real data. It can fill gaps around a core of real observations, but if we rely on it entirely in areas where we have zero ground truth, we risk working with fantasy.



MAPPING THE TERRAIN OF EVIDENCE

Where knowledge exists, where it's partial, and where it's missing.

KNOWN / KNOWN

- We know that women are significantly more susceptible to adverse drug reactions (ADRs), leading to a higher burden of drug-related harm.
- We know that women account for approximately two-thirds of all Alzheimer's disease (AD) cases globally, a disparity not explained by increased lifespan.
- We know cardiovascular disease presents differently in women, affecting both disease-presentation and treatment response.

KNOWN / UNKNOWN

- We know autoimmune diseases disproportionately affect women, but we still do not understand the initiating mechanisms.
- We know that chronic pain disorders, disproportionately affect women, but we do not fully understand the underlying sex-specific pain pathways.
- We know that women who develop schizophrenia often have a later onset and a less severe disease course than men, but we don't know how to leverage this protective factor for therapeutic development.

UNKNOWN / KNOWN

- We lack standardized algorithms to use the known event of early menopause timing as a mandatory risk input for enhanced osteoporosis and cognitive decline screening.
- We lack sex-specific laboratory criteria because current universal cut-offs, optimized for men, fail to recognize the known, physiologically lower baseline levels of markers like creatinine and ferritin in women.
- We lack standardized guidelines for interpreting the long-term impact of known pelvic surgical procedures like hysterectomy as a risk factor, despite evidence that these events may subtly alter endocrine and cardiovascular function.

UNKNOWN / UNKNOWN

- We lack the neuroimaging and computational methods to map the dynamic, hormone-driven brain connectivity across the menstrual cycle, which is essential for understanding mood and cognitive disorders.
- We lack a single-cell resolution atlas of the entire female reproductive tract across the lifespan, leaving us without a foundational biological map for common diseases like endometriosis.
- We lack basic biological data mapping the cell-specific effects of sex hormones on non-reproductive organs, particularly the liver and lungs.

[39] Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K. W. (2019). Gender bias in coreference resolution: Evaluation and debiasing methods. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, 15-20. https://arxiv.org/abs/1906.04571

The feedback loop problem:

A major hidden risk arises when models train on synthetic outputs generated by previous models.[40] Synthetic training data is susceptible to *model-induced distribution shifts (MIDS)* which can lead to cumulative degradation of fairness and under-representation of minoritized groups over successive model generations.[41] In healthcare, this risk is magnified by the weaker provenance of synthetic health data. As Giuffrè and Shung (2023) argue, synthetic records often lack traceable lineage and clear documentation of how they were generated, making it difficult to distinguish whether downstream models learn from genuine patient data or prior simulations.[42][43] Without rigorous validation, small modelling errors or biases can propagate and amplify through iterations, creating a "data-echo" effect in which artefacts of modelling become de facto input signal, embedding rather than correcting bias.

The illusion of completeness: Synthetic data can create an illusion of completeness. Because it fills gaps in existing datasets, it may appear to close representation disparities even when no new empirical information has been gathered. Giuffrè et al. emphasize synthetic data can "create the perception of sufficiency" while real-world coverage gaps remain. In areas like women's health, where ground-truth data are already limited, this risk is acute: over-reliance on synthetic records could lead researchers to believe models are performing well simply because they haven't been tested against enough real female patients.

While synthetic augmentation can improve balance and fairness when used judiciously, it cannot discover truly novel clinical phenomena. By design, synthetic generators interpolate within existing distributions; they cannot extrapolate to "unknown unknowns." Unrecorded pregnancy complications or under-studied hormonal effects will remain invisible until new real-world data are collected. Synthetic data are best regarded as a supplement to, not substitute for, empirical research. You can't generate what's never been measured.

[40] Shumailov, I., Shumaylov, Z., Zhao, Y., Gal, Y., Papernot, N., & Anderson, R. (2023). The curse of recursion: Training on generated data makes models forget. *arXiv preprint arXiv:2403.07857*. https://arxiv.org/abs/2403.07857

[41] Yang, C., Jiang, Y., Koyejo, S., & Lakkaraju, H. (2024). Fairness degradation in model collapse under synthetic data. *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 144. https://facctconference.org/static/papers24/facct24-144.pdf

[42] Wang, Z., Poulos, J., Feng, R., & Yang, Y. (2024). Gender representation disparities in chest X-ray datasets. *arXiv preprint arXiv:2408.16130*. https://arxiv.org/html/2408.16130v1

[43] Giuffrè, M., & Shung, D. L. (2023). Harnessing the power of synthetic data in healthcare: Innovation, application, and privacy. *NPJ Digital Medicine*, 6(1), 186. https://www.researchgate.net/publication/374550710_Harnessing_the_power_of_synthetic_data_in_healthcare_innovation_application_and_privacy

3.3 The Benchmark Problem

Most EHRs record diagnoses as static events rather than evolving journeys. For women, this flattening erases physiological context that shapes disease: menstrual phase, pregnancy, postpartum recovery, menopause, hormonal contraception. Without those life-stage markers, models cannot interpret the same symptom or lab result differently across biological states, even when the difference is clinically decisive.

Benchmark datasets are how we measure progress in medical AI—but they can give a false sense of accuracy. Many of the most widely used datasets are not representative of the populations they aim to serve.

- **MIMIC-III (critical care)**: About **56% of adult patients are male**, showing a modest but real imbalance.

- **CheXpert / MIMIC-CXR (chest X-rays)**: Independent audits report that **only 40–49% of patients in CheXpert** and **43–60% in MIMIC-CXR** are female, depending on race and other factors.

- **Across public chest-X-ray datasets**, reviews confirm that **men are consistently overrepresented**. [44]

These gaps are not just about fairness, they directly affect model reliability. When one sex dominates the training or benchmark data, the model learns patterns that reflect that group's anatomy, comorbidities, and imaging context. It may then misinterpret signals in the underrepresented group (for instance, women's smaller heart size or breast tissue density on X-rays).

Even worse, when evaluation datasets share the same imbalance, the problem is hidden: the model looks "high-performing" overall because its weakest cases are rare in both training and testing. This is known as **hidden stratification**, strong average scores that mask poor subgroup performance.

When new algorithms test against those datasets, performance metrics look impressive because test data match the training bias. Until sex-balanced, life-stage-tagged benchmarks exist, "state-of-the-art" remains state-of-bias.

[44] Wang, Z., Poulos, J., Feng, R., & Yang, Y. (2024). Gender representation disparities in chest X-ray datasets. *arXiv preprint arXiv:2408.16130*. https://arxiv.org/html/2408.16130v1

3.4 Recursive Training Risks

Increasingly, AI tools in health-systems are retrained using data that the tools themselves labelled or generated, rather than purely fresh, human-annotated clinical data. While this can speed development, it creates a **feedback loop**: when the original model is biased or incomplete, its inaccuracies propagate into each new generation of the model. In ML research this phenomenon is known as **model collapse**: over successive iterations trained on synthetic or self-generated data, the model gradually forgets rare patterns and converges toward the statistical mean.[45]

This risk is especially acute in women's health. Many female-specific conditions, such as Endometriosis, postpartum complications and perimenopausal syndromes, are under-represented in clinical datasets. If an AI system repeatedly retrains on its own output, which itself lacks these signals, the model's representation of women's biology may narrow over time. What is rare becomes invisible.

Furthermore, while synthetic or model-generated data may seem to fill gaps, they inherently carry forward the biases of the source model and cannot replicate the richness of real-world clinical variation. Research shows that synthetic-data-driven models can lose performance and fairness for minoritised groups.[46]

The only effective safeguard is to ensure ongoing inclusion of verified, diverse, sex-specific clinical data in every training cycle, especially for female-focussed and caregiving-centric applications. Without this, the cycle of neglect may deepen over time.

PART IV: REBUILDING THE FOUNDATION

4.1 Redefining Ground Truth

A fundamental shift is required to redefine "ground truth", away from the single, static ICD code toward a dynamic fusion of objective biomarkers, validated life-stage variables, and high-fidelity symptom data. Such a reference layer would be less vulnerable to human diagnostic bias and could serve as an independent benchmark to audit and correct the corrupted EHR record.

[45] Yang, C., Jiang, Y., Koyejo, S., & Lakkaraju, H. (2024). Fairness degradation in model collapse under synthetic data. *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 144. https://facctconference.org/static/papers24/facct24-144.pdf

[46] Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., & Gal, Y. (2024). AI models collapse when trained on recursively generated data. *Nature*, 631, 755-759. https://arxiv.org/abs/2404.05090

Beyond ICD codes: Patient-reported outcomes (PROs) provide the most direct path to reconstructing missing context in women's health. Unlike billing codes or diagnostic summaries, PROs capture lived dimensions of disease (pain intensity, fatigue, bleeding patterns, treatment response, quality of life) that structured fields routinely omit. For conditions like endometriosis, PMDD, or perimenopausal symptoms, these data are often the only reliable measure of disease activity.

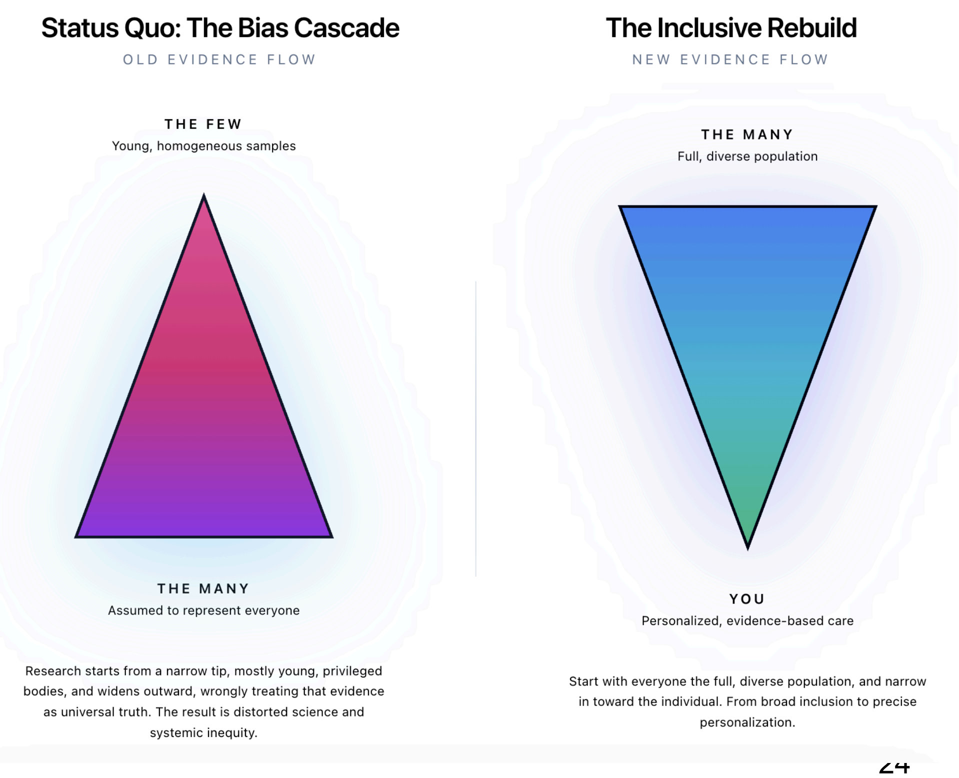
Computable sex-aware clinical rules: This means translating the vast corpus of sex-specific research findings into machine-readable clinical decision rules. Examples include separate diagnostic criteria for women (modified heart attack algorithms accounting for sex differences in troponin rises), pharmacological guidelines adjusting dosing by sex and life stage. These rules should be integrated into clinical decision support systems so clinicians get sex-specific prompts and recommendations.

Importantly, these rules need continuous updating as new evidence emerges. By making guidelines computable and dynamic, we reduce reliance on individual clinician awareness and ensure consistency.

EHR modules for female health tracking: Electronic health record systems should carry persistent awareness of where a patient is in her life course. A built-in module could tag lab values or clinical events with relevant hormonal or reproductive context, enabling both human clinicians and algorithms to interpret results correctly.

A system designed around women's health would capture fundamentally different information than traditional health records:

1. Age at menarche



Status Quo: The Bias Cascade

OLD EVIDENCE FLOW
THE FEW
Young, homogeneous samples
THE MANY
Assumed to represent everyone

Research starts from a narrow tip, mostly young, privileged bodies, and widens outward, wrongly treating that evidence as universal truth. The result is distorted science and systemic inequity.

The Inclusive Rebuild

NEW EVIDENCE FLOW
THE MANY
Full, diverse population
YOU
Personalized, evidence-based care

Start with everyone the full, diverse population, and narrow in toward the individual. From broad inclusion to precise personalization.

2. Menstrual cycle characteristics
3. Pregnancy history and outcomes
4. Use of hormonal contraceptives or therapies
5. Menopausal status

Routine lab panels would expand to include hormone levels (estrogen, progesterone, FSH/LH, AMH, testosterone) where relevant. PROs would be elevated to first-class data: tracking menstrual pain, fatigue, mood changes, sexual function as vital signs in their own right. Contextual data like caregiving burden or social stressors might be included, recognizing their outsized impact on women's health.

New data taxonomies: Developing representations for physiology that capture life stage and cyclical dynamics: data models encoding a patient's menstrual phase or pregnancy status as time-varying parameters, rather than ignoring them. This could involve extensions to health data standards (adding fields for last menstrual period, menopausal status) and creating better ontologies for female-specific conditions.

4.2 The Infrastructure Challenge

In safety-critical industries, high-fidelity data labeling is treated as a non-negotiable cost of doing business. In autonomous vehicles, robotics, and aerospace, companies like Scale AI invest heavily in expert-verified, pixel-level annotations because any error can lead to physical harm.[47] The "ground truth" in these fields is not just data, it's infrastructure.

Healthcare AI, by contrast, has often prioritized volume over fidelity. Large players like Google DeepMind have advanced models for acute conditions like acute kidney injury using enormous, heterogeneous EHR datasets. Yet another example where a lower AKI episode-level sensitivity was observed in females as compared to males (44.8% vs. 56.0%, respectively), due to the fact that the training data set was 94% male.[48] These models demonstrate technical prowess, but their performance depends on data abundance rather than meticulous, expert-curated labeling needed for complex, multi-factorial conditions—especially those disproportionately affecting women: PMDD, perimenopause, postpartum anxiety.

The specialist labeling bottleneck: The path to higher-quality medical data runs through clinicians, but they are already at breaking point. Physicians spend roughly two hours on

[47] IEEE. (2024). Safe reinforcement learning in critical systems. *IEEE Transactions*. https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=10735161

[48] Rajkomar, A., Oren, E., Chen, K., et al. (2018). Scalable and accurate deep learning with electronic health records. *NPJ Digital Medicine*, 1(1), 18. https://pmc.ncbi.nlm.nih.gov/articles/PMC10751025/

EHR documentation for every one hour of direct patient care, a ratio widely cited as a major driver of burnout.[49]

Correcting corrupted ground truth data, through chart review, re-coding, or annotation—is uncompensated administrative work. Without structural reform, clinicians have no incentive to perform this critical validation.

A systemic redesign is required. Compensation and workflow models should treat data validation as a clinical and research function, providing protected, funded time for specialists to annotate and audit data. AI tools must also integrate seamlessly into care delivery: if validating or correcting data directly supports patient care, the incentive to maintain data quality becomes intrinsic rather than bureaucratic.

Female clinicians as high-fidelity data sources: One of the most overlooked sources of rich medical data already exists within the system: the clinical notes of female physicians. Studies show female surgeons and internists produce 40% longer progress notes on average than male colleagues, and bill twice as many Level 5 consults and new-patient visits, the highest level of clinical complexity.[50]

Longer documentation and higher-complexity encounters suggest these clinicians capture more nuanced patient narratives and subtler diagnostic clues, particularly relevant for women's multifactorial conditions. This information often resides in the unstructured text layer of the EHR, hidden from traditional analytics that rely on ICD codes. Leveraging this unstructured data could counterbalance decades of bias embedded in coded records, turning what has been an invisible asset into a foundation for equity.

4.3 Regulatory & Developer Accountability

AI Developers and Technology Companies who design and train healthcare algorithms have direct responsibility to embed fairness from the start. Developers should:

- Make sex-stratified evaluation standard practice. Every model intended for clinical use should report performance specifically on women (ideally across female subgroups by age or life stage)
- Provide transparent documentation of training data, including gender breakdown and known gaps
- Conduct rigorous bias testing before any product launch

[49] American Medical Association. (n.d.). 7 things about EHRs that stress out doctors. https://www.ama-assn.org/practice-management/digital-health/7-things-about-ehrs-stress-out-doctors

[50] Mills, J. R., Ahmed, S., Chen, P. H., et al. (2023). Sex-specific differences in clinical documentation: A retrospective study. *JAMA Network Open*, 6(7), e2324776. https://pmc.ncbi.nlm.nih.gov/articles/PMC11267410/

- Incorporate women-centered design by involving women as end-users (patients and providers) in the design process

Current FDA gaps: A 2024 scoping review of 692 FDA-cleared AI and machine-learning devices found only 3.6% of submissions reported race or ethnicity, and sex-stratified performance was almost never provided. About 99% omitted socioeconomic data entirely.[51][^50]

The FDA's 2025 *Draft Guidance on AI/ML-Enabled Device Software Functions* and *Sex-Specific Clinical Considerations Guidance* encourage, but do not require, sex-disaggregated performance reporting. Developers may include it voluntarily, but it isn't a mandatory approval criterion.

Required reforms:

- Pre-market: require sex-stratified performance tables (sensitivity, specificity, calibration) for any indication spanning sexes
- Change-control (learning systems): mandate subgroup drift monitoring (monthly calibration changes by sex) with public variance budgets
- Post-market: require sex-disaggregated adverse-event and performance reports (quarterly), aligned to FAERS-style transparency[52]

4.4 The Opportunity

While the challenges are significant, women's health also presents unique opportunities for AI innovation when approached correctly.

Multi-modal, dynamic data: Consider the data generated across a woman's lifespan, it's rich, multi-modal, and dynamic. Hormonal cycles create regular time-series patterns (~28-day menstrual cycles) that machine learning algorithms excel at analyzing. Reproductive events (menstruation, pregnancy, childbirth, menopause) provide clear transitions and natural experiments that segment life stages and risk profiles.The menstrual cycle is increasingly recognized as a fifth vital sign (alongside temperature, pulse, respiration, and blood pressure). Cycle regularity, timing, flow, or symptom changes can reveal meaningful insights into overall health, not just reproductive function. Shifts in a cycle can act as early clinical signals of endocrine, metabolic, or gynecologic disorders, prompting earlier evaluation.

Equally important, pregnancy, the postpartum period, perimenopause, and menopause are not simply points in time: they are distinct physiological states. In each one, hormone levels

[51] FDA Review Study. (2024). Demographic reporting in FDA-cleared AI/ML medical devices. https://pmc.ncbi.nlm.nih.gov/articles/PMC11450195/
[52] U.S. Food and Drug Administration. (n.d.). FDA's Adverse Event Reporting System (FAERS). https://www.fda.gov/drugs/surveillance/fdas-adverse-event-reporting-system-faers

change in ways that reshape how multiple systems function, including metabolism, immune responses, cardiovascular risk, bone turnover, cognition, and even drug metabolism. These changes are not identical for every woman, but they follow patterns that can be measured and modeled.

Because of this hormonal context, the same symptom, or even the same lab value, can mean different things at different life stages. A thyroid result requiring monitoring in the postpartum period may be unremarkable outside it. Chest pain in perimenopause may require a different risk assessment pathway than the same symptom in a healthy 25-year-old.

Why this data matters beyond gynecology: Women's reproductive life stages influence nearly every major medical specialty, not just obstetrics or gynecology. Hormonal transitions can signal systemic risk. For example, frequent hot flashes and night sweats in the menopausal transition have been associated with higher long-term cardiovascular disease risk. Likewise, menstrual cycle irregularities may precede or correlate with metabolic conditions (such as insulin resistance) or autoimmune disorders. Recognizing these patterns early allows for earlier screening, prevention, and intervention.

After menopause, cardiovascular disease becomes the leading cause of death in women. Vasomotor symptoms, earlier age at menopause, and irregular cycles have each been linked in research to increased cardiovascular risk. Understanding reproductive history helps clinicians correctly interpret lipids, blood pressure, and cardiac symptoms, particularly during and after the menopausal transition.

Hormonal transitions affect insulin sensitivity, adipose distribution, bone metabolism, and thyroid function. A lab value that appears within normal limits may have different clinical weight depending on whether a patient is postpartum, perimenopausal, or postmenopausal. Without accounting for hormonal stage, metabolic disease can be misclassified, or missed entirely.

Immune function fluctuates across the menstrual cycle and shifts substantially in pregnancy and menopause. Several autoimmune diseases are more common in women and may flare or change severity with hormonal transitions. Understanding these patterns supports earlier detection, better monitoring, and personalized treatment planning.

When clinicians treat reproductive history, menstrual patterns, or menopausal symptoms as isolated "women's issues," they inadvertently lose critical systemic health information. These patterns are not peripheral, they are biologically meaningful signals that intersect with cardiometabolic health, immune function, mental health, longevity, and medication response (see *The Missing Half of Longevity Science: Why Women Are The Key*. Part of a three-paper FemTechnology Research Series examining women's health through the lenses of AI and data infrastructure, sex-specific longevity mechanisms, and the insurance and economic cost of systemic gaps in care).

Integrating cycle-aware, life-stage-aware data into clinical care and AI systems allows healthcare to move from uniform assumptions to context-specific interpretation—the foundation of true precision medicine for women.

Context-aware, personalized medicine: With such a foundation, AI models could understand female biology as dynamic across multiple timescales. Instead of treating "woman" as a static category, the model knows an 18-year-old woman, a pregnant 30-year-old, and a 55-year-old postmenopausal woman have very different physiology over time. It can account for cyclical variation (intra-month changes) as well as longitudinal changes over years or decades.

This enables truly contextual, adaptive predictions. Recognizing that a given symptom or lab value may have different significance depending on menstrual cycle phase or menopausal stage. A chest pain in a perimenopausal woman might warrant a different diagnostic pathway than the same chest pain in a 25-year-old, and a smart AI would know this. An AI could learn that hormone levels indicating PCOS in one woman might be normal for another, depending on baseline and life stage, nuances current one-size-fits-all models miss.

Women's health as AI stress test: By solving for the more complex case (models handling hormonal cycles, pregnancy, etc.), we push the field beyond simplistic assumptions. Centering women in biomedical research is not only about equity but about discovery: studying female biology often uncovers mechanisms and pathways that were overlooked, which can then translate to benefits for both sexes. Women's health can be viewed as the ultimate stress test for precision AI—an opportunity to develop models that handle variability and personalization at a higher level.

RECOMMENDATIONS FOR AI STAKEHOLDERS

- **Clinicians and health systems** have the power to demand and deploy diagnostic tools validated for sex and gender differences. By insisting on algorithmic transparency and equitable thresholds, for instance, ensuring cardiovascular risk models reflect female symptom patterns, they can make bias visible and correctable. Failing to do so means automating decades of under-recognition directly into everyday care.

  Clinicians are increasingly judged on efficiency, patient satisfaction, and trust. AI is entering decision support, triage, and diagnostics, yet most tools are built on data that underrepresent women. When these systems miss female presentations of common conditions, it creates more follow-up visits, diagnostic uncertainty, and defensive medicine.

  The opportunity: Systems that detect bias or integrate sex-aware parameters can *save time, reduce uncertainty,* and build patient trust.

- **Researchers** hold the keys to the datasets and validation pipelines that define "truth." Their opportunity lies in collecting and labeling data that capture hormonal, reproductive, and life-stage variables, not just as covariates, but as essential dimensions of health. Inaction leaves AI blind to the biological diversity it claims to model. Right now, most datasets underrepresent women: biologically, hormonally, and behaviorally. That weakens model performance.

  The opportunity: Better data yields better science. Sex-stratified datasets lead to higher-quality outputs and attract cross-disciplinary collaboration (from pharmacology, behavioral science, or digital health). This makes research more *publishable, citable, and fundable*.

  The cost of inaction: Models trained on homogeneous data perform poorly in the real world. When those failures surface, they damage credibility and future partnerships.

- **Industry** can audit for sex-specific performance, embedding representative training data, and hiring interdisciplinary teams. The downside of ignoring this is

- commercial as well as ethical, AI tools that fail half the population will ultimately fail the market.

  Women are the majority of healthcare consumers, responsible for more than 80% of health-related purchasing decisions, and they make up the majority of the global healthcare workforce. Yet the systems, datasets, and algorithms driving modern medicine still center on male physiology. That mismatch is a market inefficiency hiding in plain sight and women are increasingly waking up to that knowledge and demanding change. The next generation of high-performing AI in healthcare will be built by companies that see women not as an afterthought, but as the core design constraint that makes systems better for everyone.

  The opportunity: This is not about compliance or signaling; it is about foresight. Developing AI systems that reflect the full spectrum of human biology, including female physiology, hormonal patterns, and lived experience, will yield insights that are currently invisible in most datasets. Those insights translate into more accurate models, stronger evidence bases, and technologies that perform better across populations. As awareness of these disparities grows, organizations that have already invested in representative data will be better positioned to lead, partner, and scale with credibility.

  The cost of inaction: Delaying investment in equitable data means ceding the frontier. The companies that define how women's health is measured will end up setting the technical standards for the entire sector. Once those benchmarks exist, everyone else will be building on their terms and licensing their data. The next wave of progress in health technology won't come from incremental model tuning; it will come from expanding the datasets that determine what's knowable in the first place.

- **Regulators and policymakers** can set the guardrails: mandating sex-stratified reporting, incentivizing inclusive data collection, and enforcing accountability in algorithmic certification.

- **Employers and payers** can use purchasing power to demand gender-aware analytics and benefits programs. The opportunity is to cut costs and improve outcomes by meeting real needs; the cost of passivity is perpetuating inefficiency masked as neutrality.

Why it matters: Women make up a significant proportion of the workforce and are key to productivity, yet remain underserved by traditional benefits and digital health solutions. Poorly understood conditions (menopause, autoimmune disease, pelvic pain) quietly drive absenteeism, turnover, and unmanaged costs.

The opportunity: Data that reveals where care gaps actually are can help design *smarter benefits* and reduce waste. Investing in tools that meet women's needs saves money by preventing long, costly diagnostic journeys and improving retention.

The cost of inaction: High spend with low satisfaction. Employers keep paying for care that doesn't work and lose valuable employees because their systems don't meet them halfway.

- And finally, **patients and the public**, especially women,have the right to ask how systems make decisions about their bodies. Engaging them not only builds trust, it improves data quality. If they remain excluded or unaware, the resulting silence becomes the next data gap.

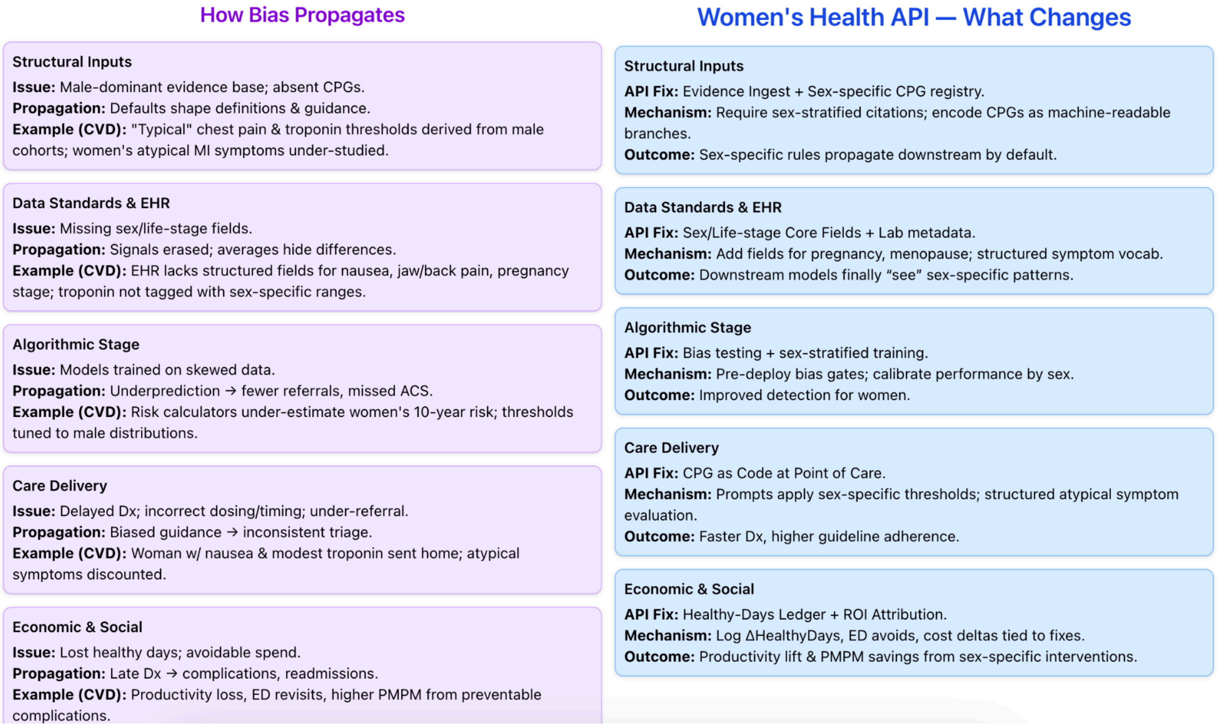
The collective opportunity is to rebuild medicine's foundation on accurate, inclusive data, before this new era of intelligence becomes the old story of exclusion told in code.

CONCLUSION

The convergence of AI and women's health represents a pivotal moment: it is both a technological challenge and a profound opportunity.

On one hand, we risk perpetuating and even exacerbating gender biases in healthcare if we allow opaque algorithms to learn from biased data. On the other hand, by deliberately addressing the data scarcity and bias issues, we can harness AI to finally close longstanding gaps in women's health outcomes. To prevent that, awareness must translate into architecture, ensuring that data, model design, and validation all reflect women's lived and biological realities before inequity becomes embedded code.

This will require reimagining data strategies, from collecting richer female-specific data to validating models for fairness and accountability. It will also require breaking down silos between clinical practice, data science, and policy so that each informs the other in a



How Bias Propagates

Structural Inputs
Issue: Male-dominant evidence base; absent CPGs.
Propagation: Defaults shape definitions & guidance.
Example (CVD): "Typical" chest pain & troponin thresholds derived from male cohorts; women's atypical MI symptoms under-studied.

Data Standards & EHR
Issue: Missing sex/life-stage fields.
Propagation: Signals erased; averages hide differences.
Example (CVD): EHR lacks structured fields for nausea, jaw/back pain, pregnancy stage; troponin not tagged with sex-specific ranges.

Algorithmic Stage
Issue: Models trained on skewed data.
Propagation: Underprediction → fewer referrals, missed ACS.
Example (CVD): Risk calculators under-estimate women's 10-year risk; thresholds tuned to male distributions.

Care Delivery
Issue: Delayed Dx; incorrect dosing/timing; under-referral.
Propagation: Biased guidance → inconsistent triage.
Example (CVD): Woman w/ nausea & modest troponin sent home; atypical symptoms discounted.

Economic & Social
Issue: Lost healthy days; avoidable spend.
Propagation: Late Dx → complications, readmissions.
Example (CVD): Productivity loss, ED revisits, higher PMPM from preventable complications.

Women's Health API — What Changes

Structural Inputs
API Fix: Evidence Ingest + Sex-specific CPG registry.
Mechanism: Require sex-stratified citations; encode CPGs as machine-readable branches.
Outcome: Sex-specific rules propagate downstream by default.

Data Standards & EHR
API Fix: Sex/Life-stage Core Fields + Lab metadata.
Mechanism: Add fields for pregnancy, menopause; structured symptom vocab.
Outcome: Downstream models finally "see" sex-specific patterns.

Algorithmic Stage
API Fix: Bias testing + sex-stratified training.
Mechanism: Pre-deploy bias gates; calibrate performance by sex.
Outcome: Improved detection for women.

Care Delivery
API Fix: CPG as Code at Point of Care.
Mechanism: Prompts apply sex-specific thresholds; structured atypical symptom evaluation.
Outcome: Faster Dx, higher guideline adherence.

Economic & Social
API Fix: Healthy-Days Ledger + ROI Attribution.
Mechanism: Log ΔHealthyDays, ED avoids, cost deltas tied to fixes.
Outcome: Productivity lift & PMPM savings from sex-specific interventions.

virtuous cycle. The benefits of getting this right are enormous. It is also the fastest path to innovation. Every major breakthrough in medicine has come from studying what was once overlooked.

Technically, focusing on women's health will push AI into new territory of truly personalized, context-aware medicine, advances that will ultimately improve care for everyone. Socially and economically, it will mean healthier lives for over half the population and significant gains in productivity and healthcare value. Put simply, fixing gender bias in healthcare AI isn't just about justice; it also translates into better quality care and efficiency across health systems (see T*he Price of Invisibility: Why Fixing Women’s Health Is the Fastest Route to Reducing Healthcare Spend,* part of a coordinated three-paper FemTechnology Research Series examining women’s health through the lenses of sex-specific longevity science, AI and data infrastructure, and the economic and insurance gap).

We must remember that technology is not destiny. The current flaws in women's health AI are the product of human choices; what data to collect, what to prioritize, what to ignore. And better choices can correct the course.

Ultimately, we face a choice of two futures. In one, we settle for perpetuating existing systems with fairness patches applied after the fact, allowing legacy biases to calcify in digital form.

The work ahead does not belong to one sector, it belongs to all of them.

FOOTNOTES

[^1]: HIT Consultant. (2025, October 22). Healthcare AI adoption is 2.2x faster than the broader economy. https://hitconsultant.net/2025/10/22/healthcare-ai-adoption-is-2-2x-faster-than-the-broader-economy

[^2]: Shah, N. H., Milstein, A., & Bagley, S. C. (2019). Making machine learning models clinically useful. *JAMA*, 322(14), 1351-1352. https://pmc.ncbi.nlm.nih.gov/articles/PMC11046491/

[^3]: Tseng, A., Shrikumar, A., & Kundaje, A. (2024). Systematic characterization of the effectiveness of alignment in large language models for categorical decisions. *arXiv preprint arXiv:2409.18995*. https://pubmed.ncbi.nlm.nih.gov/38112814/

[^4]: Applied Clinical Trials. (n.d.). Gender bias in the clinical evaluation of drugs. https://www.appliedclinicaltrialsonline.com/view/gender-bias-in-the-clinical-evaluation-of-drugs

[^5]: Zucker, I., & Beery, A. K. (2010). Males still dominate animal studies. *Nature*, 465(7299), 690. https://pmc.ncbi.nlm.nih.gov/articles/PMC8812498/

[^6]: Sosinsky, A., Agrawal, R., Gray, S. W., & Freedman, R. A. (2022). The evolution of clinical trial eligibility criteria and enrollment of women. *American Journal of Clinical Oncology*, 45(10), 421-426. https://www.sciencedirect.com/science/article/abs/pii/S1551714422000441

[^7]: Hao, Y., Gong, R., Li, T., Cheng, Y., & Wang, Y. (2020). Sex annotation in publicly available cancer genomic datasets. *Scientific Data*, 7(1), 250. https://pubmed.ncbi.nlm.nih.gov/38498336/

[^8]: Zucker, I., & Beery, A. K. (2010). Males still dominate animal studies. *Nature*, 465(7299), 690. https://pubmed.ncbi.nlm.nih.gov/20535186/

[^9]: Campesi, I., Franconi, F., & Seghieri, G. (2018). Sex-gender-related therapeutic approaches for cardiovascular disease. *Pharmacological Research*, 132, 130-137. https://pmc.ncbi.nlm.nih.gov/articles/PMC7275616/

[^10]: Harvard Gazette. (2023, December). Women more likely to suffer drug side effects, but reason may not be biology. https://news.harvard.edu/gazette/story/2023/12/women-more-likely-to-suffer-drug-side-effects-but-reason-may-not-be-biology/

[^11]: Huang, K., Altosaar, J., & Ranganath, R. (2024). Cell2Sentence: Teaching large language models the language of biology. *Nature Communications*, 15(1), 9715. https://pmc.ncbi.nlm.nih.gov/articles/PMC11565894/

[^12]: Wachs, D., Yao, Y., & Veeraraghavan, A. (2019). Sex-specific gene expression in mammalian cells. *Genome Biology*, 20(1), 202. https://pmc.ncbi.nlm.nih.gov/articles/PMC7898458/

[^13]: Oliva, M., Muñoz-Aguirre, M., Kim-Hellmuth, S., et al. (2020). The impact of sex on gene expression across human tissues. *Science*, 369(6509), eaba3066. https://pubmed.ncbi.nlm.nih.gov/32913072/

[^14]: Zucker, I., & Prendergast, B. J. (2020). Sex differences in pharmacokinetics predict adverse drug reactions in women. *Biology of Sex Differences*, 11(1), 32. https://pubmed.ncbi.nlm.nih.gov/32503637/

[^15]: Webb, E., Shah, N., Veselkov, K., Rare Disease Working Group, & Cheng, F. (2025). Sex-specific considerations in clinical AI fairness: A review of gastroenterology and hepatology algorithms. *NPJ Digital Medicine*, 8(1), 67. https://www.nature.com/articles/s41746-025-01667-2

[^16]: Chapman, A. R., et al. (2019). High-sensitivity cardiac troponin and the diagnosis of myocardial infarction in patients with kidney disease. *Circulation*, 140(6), 423-435. https://pubmed.ncbi.nlm.nih.gov/31623760/

[^17]: Chapman, A. R., et al. (2019). High-sensitivity cardiac troponin and sex-disaggregated outcomes. *BMC Medicine*, 17(1), 213. https://pmc.ncbi.nlm.nih.gov/articles/PMC6876271/

[^18]: Lee, M. S., Park, H., Woo, J. S., et al. (2024). Beta-blocker therapy in heart failure patients: The REBOOT Heart Trial. *The Lancet*, 403(10429), 819-829. https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(24)00244-6/fulltext

[^19]: Kjaergaard, J., Arfwedson Wang, C. E., & Waterloo, K. (2019). A study of gender differences in diagnostic delay of mental disorders. *BMC Psychiatry*, 19(1), 117. https://pubmed.ncbi.nlm.nih.gov/30737381/

[^20]: The Women's Heart Foundation. (n.d.). Women's heart disease symptoms overlooked. https://www.womenheart.org/womens-heart-disease-symptoms-overlooked/

[^21]: Women's Heart Alliance. (n.d.). Women found to be at higher risk for heart failure and heart attack death than men. https://newsroom.heart.org/news/women-found-to-be-at-higher-risk-for-heart-failure-and-heart-attack-death-than-men

[^22]: University of York. (2024). Diagnosis endometriosis delay. https://www.york.ac.uk/news-and-events/news/2024/research/diagnosis-endometriosis-delay/

[^23]: Shah, P., Kendall, F., Khozin, S., et al. (2019). Artificial intelligence and machine learning in clinical development: A translational perspective. *NPJ Digital Medicine*, 2(1), 69. https://pmc.ncbi.nlm.nih.gov/articles/PMC10938158/

[^24]: Zhang, J., Whebell, S., & Gallifant, J. (2022). Temporal aspects of electronic health record data and algorithmic fairness. *JAMIA Open*, 5(4), ooac099. https://pmc.ncbi.nlm.nih.gov/articles/PMC9759969/

[^25]: American Medical Women's Association. (n.d.). Lost in translation: When women's health is called mental health and vice versa. https://amwa-doc.org/lost-in-translation-when-womens-health-is-called-mental-health-and-vice-versa/

[^26]: Tseng, A., Shrikumar, A., & Kundaje, A. (2024). Systematic characterization of the effectiveness of alignment in large language models for categorical decisions. *arXiv preprint arXiv:2409.18995*. https://arxiv.org/abs/2409.18995

[^27]: Rickman, S., Bohnet, H., Hogan, S., et al. (2025). Gender bias in large language models for adult social care documentation. *BMC Medical Informatics and Decision Making*, 25(1), 118. https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-025-03118-0

[^28]: The Guardian. (2025, August 11). AI tools used by English councils downplay women's health issues, study finds. https://www.theguardian.com/technology/2025/aug/11/ai-tools-used-by-english-councils-downplay-womens-health-issues-study-finds

[^29]: London School of Economics. (2025). AI tools risk downplaying women's health needs in social care. https://www.lse.ac.uk/news/latest-news-from-lse/ai-tools-risk-downplaying-womens-health-needs-in-social-care

[^30]: Kim, P. W., Xie, S., Huang, M. K., Aguirre-Chang, G., & Chow, D. S. (2023). Race and sex bias in AI medical diagnosis: A study of ChatGPT and Google Bard responses. *JAMA Network Open*, 6(11), e2342343. https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2810775

[^31]: Nature. (2025). Delphi-2M: A GPT-style foundation model for longitudinal disease prediction. *Nature*, 637, 155-163. https://www.nature.com/articles/s41586-025-09529-3

[^32]: UK Biobank. (n.d.). Evaluating the incidence of complications among people with diabetes according to age of onset: Findings from the UK Biobank. https://www.ukbiobank.ac.uk/publications/evaluating-the-incidence-of-complications-among-people-with-diabetes-according-to-age-of-onset-findings-from-the-uk-biobank/

[^33]: International Diabetes Federation. (n.d.). Diabetes facts & figures. https://idf.org/about-diabetes/diabetes-facts-figures/

[^34]: PMC. (n.d.). Endometriosis prevalence and diagnosis patterns. https://pmc.ncbi.nlm.nih.gov/articles/PMC9127440/

[^35]: Straw, I., Rees, J., & Nachev, P. (2024). Sex-based disparities in machine learning models for cardiac disease prediction. *European Heart Journal - Digital Health*, 5(5), 567-576. https://pmc.ncbi.nlm.nih.gov/articles/PMC11384168/

[^36]: Zhou, Y., Wang, L., Tang, L., et al. (2024). Sex-based performance disparities in clinical risk prediction models at German hospitals. *Journal of Biomedical Informatics*, 154, 104639. https://www.sciencedirect.com/science/article/pii/S1532046424001102

[^37]: Chung, K., Yoo, H., Lee, J., et al. (2021). Sex-specific prediction model for severe COVID-19 using machine learning. *Journal of Personalized Medicine*, 11(11), 1190. https://pmc.ncbi.nlm.nih.gov/articles/PMC8667070/

[^38]: Lu, H., Uddin, S., Hajati, F., Moni, M. A., & Khushi, M. (2022). A patient network-based machine learning model for disease prediction: The case of type 2 diabetes. *Applied Intelligence*, 52(3), 2411-2422. https://pubmed.ncbi.nlm.nih.gov/35470133/

[^39]: Olesen, T. B., Leibig, C., & Lauritzen, A. D. (2024). Slicing through bias: Explaining performance gaps in medical image analysis. *arXiv preprint arXiv:2406.12142*. https://arxiv.org/html/2406.12142v2

[^40]: Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K. W. (2019). Gender bias in coreference resolution: Evaluation and debiasing methods. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, 15-20. https://arxiv.org/abs/1906.04571

[^41]: Shumailov, I., Shumaylov, Z., Zhao, Y., Gal, Y., Papernot, N., & Anderson, R. (2023). The curse of recursion: Training on generated data makes models forget. *arXiv preprint arXiv:2403.07857*. https://arxiv.org/abs/2403.07857

[^42]: Yang, C., Jiang, Y., Koyejo, S., & Lakkaraju, H. (2024). Fairness degradation in model collapse under synthetic data. *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 144. https://facctconference.org/static/papers24/facct24-144.pdf

[^43]: Giuffrè, M., & Shung, D. L. (2023). Harnessing the power of synthetic data in healthcare: Innovation, application, and privacy. *NPJ Digital Medicine*, 6(1), 186. https://www.researchgate.net/publication/374550710_Harnessing_the_power_of_synthetic_data_in_healthcare_innovation_application_and_privacy

[^44]: Wang, Z., Poulos, J., Feng, R., & Yang, Y. (2024). Gender representation disparities in chest X-ray datasets. *arXiv preprint arXiv:2408.16130*. https://arxiv.org/html/2408.16130v1

[^45]: Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., & Gal, Y. (2024). AI models collapse when trained on recursively generated data. *Nature*, 631, 755-759. https://arxiv.org/abs/2404.05090

[^46]: IEEE. (2024). Safe reinforcement learning in critical systems. *IEEE Transactions*. https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=10735161

[^47]: Rajkomar, A., Oren, E., Chen, K., et al. (2018). Scalable and accurate deep learning with electronic health records. *NPJ Digital Medicine*, 1(1), 18. https://pmc.ncbi.nlm.nih.gov/articles/PMC10751025/

[^48]: American Medical Association. (n.d.). 7 things about EHRs that stress out doctors. https://www.ama-assn.org/practice-management/digital-health/7-things-about-ehrs-stress-out-doctors

[^49]: Mills, J. R., Ahmed, S., Chen, P. H., et al. (2023). Sex-specific differences in clinical documentation: A retrospective study. *JAMA Network Open*, 6(7), e2324776. https://pmc.ncbi.nlm.nih.gov/articles/PMC11267410/

[^50]: FDA Review Study. (2024). Demographic reporting in FDA-cleared AI/ML medical devices. https://pmc.ncbi.nlm.nih.gov/articles/PMC11450195/

[^51]: U.S. Food and Drug Administration. (n.d.). FDA's Adverse Event Reporting System (FAERS). https://www.fda.gov/drugs/surveillance/fdas-adverse-event-reporting-system-faers

About the Authors

Oriana Kraft

Oriana Kraft is the founder of FemTechnology and creator of ORI, a new infrastructure layer for women's healthcare. Trained in medicine and engineering at ETH Zurich, she began mapping the systemic gaps in women's health as part of her thesis, work that evolved into the FemTechnology Summit, now a global convening platform spanning more than 60 countries and sectors across research, biotech, clinical care, and industry.

In 2023, Oriana led the FemTechnology Summit at Roche's global headquarters, bringing together 150 innovators to address themes such as *Redesigning Healthcare with Women in Mind* and *AI & the Gender Data Health Gap.* Outputs from the Summit have since been featured by the U.S. Chamber of Commerce, the World Economic Forum, and other national and international bodies. To accelerate scientific progress, Oriana also established the FemTechnology University Series, partnering with institutions including ETH Zurich, Imperial College London, King's College London, and Harvard Business School to elevate women's health research and close knowledge gaps.

ORI translates these insights into applied infrastructure. ORI collects nuanced data, integrates clinical guidance, and connects women to personalized care pathways across life stages. For employers and health systems, ORI uncovers hidden cost drivers, strengthens benefit strategy, and provides actionable insights to improve retention, reduce absenteeism and presenteeism, and align care with real-world needs. ORI is purpose-built for women: rooted in female biology, grounded in clinical best practice, and adaptive to individual preferences, goals, and everyday constraints.

Caitlin Kraft-Buchman

Caitlin Kraft-Buchman is CEO/Founder Women At The Table – a systems change, gender equality and democracy Think Tank based in Geneva.
She is Co-Founder/Leader of the <A+> Alliance for Inclusive Algorithms – a global coalition prototyping a new future of AI, and the Leader of the <AI & Equality> Human Rights Initiative that supports a global community working for human rights-based approaches to AI development.
Caitlin is also Co-Founder of the International Gender Champions (IGC) - with hubs in Geneva, New York, Vienna, Nairobi, The Hague & Paris - bringing together female & male heads of organizations, including the UN Secretary-General, to break down gender barriers.

She serves on the IGC Global Board, and co-leads the IGC Impact Group on Digital and New Emerging Technologies. Caitlin was one of the Network of Experts for the UN Secretary General's AI Advisory Body, and member of the Gender & AI Advisory Group for the 2025 AI Action Summit held in Paris. She is one of UNESCO's AI Ethics Experts Without Borders, member of UNESCO's WomenForEthicalAI working group; one of UNECE's Team of Specialists on Gender Responsive Standards, and Co-Chair of the Gender Advisory Board for the UN Commission on Science & Technology for Development (CSTD).

About Women At The Table

Women At The Table is a Geneva-based systems change, equality and democracy think tank focused on ensuring digital public infrastructure is inclusive by design. Their initiatives include the <A+> Alliance for Inclusive Algorithms and the <> Human Rights Initiative which partners with UN agencies, governments, and civil society across Africa, Latin America, and globally to advance AI policy, applied research, and human rights-based frameworks that ensure democracy and equity for all.

About FemTechnology

FemTechnology is building the future of women's healthcare by addressing the gender health data gap and connecting innovation across the ecosystem. Through the FemTechnology Summit, a global university series, and applied efforts (such as **ORI**), FemTechnology bridges the divide between discovery, deployment, and real-world care. Learn more at : www.femtechnology.org

About ORI

ORI combines structured clinical intake, rules-based logic, and adaptive AI to deliver precision care guidance built for women. Inputs (such as symptoms, severity, reproductive life stage, comorbidities, lifestyle factors, and care preferences) are processed through a clinically validated decision framework informed by female-specific research. This produces a personalized care route aligned with best practice guidelines and available care resources.

Women receive a tailored recommendation: what condition or pathway is most likely relevant, what interventions are appropriate, which providers or tools match their context, and how to act—step-by-step. ORI tracks outcomes and feedback to refine future recommendations.

At the system level, anonymized patterns highlight unmet needs, misaligned benefits, and avoidable care costs, enabling employers and health systems to adjust offerings, target interventions, and improve outcomes at scale. Learn more at : www.ori.care

This publication has been created and published with the support of and in cooperation with the Friedrich Naumann Foundation for Freedom Human Rights Hub.

FRIEDRICH NAUMANN
FOUNDATION For Freedom.