

Toward demographic robustness in digital patient twins: Addressing the gender data gap

Dana Mahr¹ , Meike Hebich² and Nora Weinberger¹ 

Abstract

In this opinion piece, we argue that sex- and gender-based equity must become a foundational criterion in the design and implementation of digital patient twins. Digital patient twins offer a promising avenue for precision medicine by simulating individual health states and treatment responses. However, their clinical utility and fairness depend on whether diverse patient populations are adequately represented and accounted for in the data and devices on which these models are built. Drawing on evidence from cardiology, endocrinology, mental health, and medical device research, this article shows how current digital patient twin initiatives often overrepresent male, white, and socioeconomically privileged populations, while women, gender-diverse individuals, and people of color remain underrepresented. These imbalances can lead to systematic misdiagnoses, misinterpretation of physiological variation, and measurement inaccuracies. Documented examples include under-recognition of heart failure with preserved ejection fraction in women, omission of menstrual cycle-related changes in glycemic control, underdiagnosis of depression in women by speech-based AI models, and oxygen saturation overestimation in patients with darker skin tones. We argue that these disparities are rooted in structural biases in clinical research and are perpetuated when sex- and gender-specific variables, intersectional factors, and subgroup validation are absent from model design. Addressing these limitations requires balanced data representation, integration of sex- and gender-informed knowledge, participatory design with diverse patient groups, subgroup performance testing, transparent reporting, and mitigation of device-related bias. We contend that these are not optional refinements but prerequisites for realizing the promise of personalized care without reproducing or deepening existing health inequities.

Keywords

Digital patient twins, gender data gap, health equity, precision medicine, bias in AI, sex and gender differences

Received: 28 March 2025; accepted: 24 September 2025

Introduction

Digital patient twins (DPTs) are virtual models that integrate multi-modal data—from imaging and genomics to continuous sensor streams—to simulate an individual's physiological state and disease trajectory. They support precision medicine by allowing clinicians and patients to test “what-if” scenarios and tailor treatments to the person rather than the population average. However, the clinical value and ethical legitimacy of DPTs depend on the quality of data and assumptions upon which they are built. Machine learning algorithms, often integral to DPT architectures, can replicate and intensify sex, gender, racial, and socioeconomic biases when trained on unbalanced datasets.¹ Historically, medicine has used male bodies as the default model, and merely including women without analyzing sex-specific differences does not correct this bias.² Clinical

AI systems risk perpetuating such structural legacies when their design overlooks social and cultural dynamics embedded in the data.¹ When training data lack demographic breadth, algorithmic decisions may lead to misclassification, inappropriate treatment decisions, and reduced clinical safety.

¹Institute for Technology Assessment and Systems Analysis, Karlsruhe Institute of Technology, Karlsruhe, Germany

²KIT-Center Humans and Technology, Karlsruhe Institute of Technology, Karlsruhe, Germany

Corresponding author:

Dana Mahr, Institute for Technology Assessment and Systems Analysis, Karlsruhe Institute of Technology, Karlstraße 11, 76133 Karlsruhe, Germany.

Email: dana.mahr@kit.edu



Several studies of DPT initiatives point to a persistent gender data gap. Many models rely on datasets that over-represent male, white, and affluent populations, while women, trans and non-binary persons, and other marginalized groups remain underrepresented. For example, a philosophical and bioethical examination of trans-specific digital twins highlights that underlying medical datasets predominantly reflect the “male, cis-gendered body,” leaving female and queer individuals largely invisible.³ Likewise, a scoping review of digital twins warns that biased health datasets can be heavily skewed (by gender or race) and generate suboptimal treatment recommendations.⁴ These concerns echo findings from broader studies on AI in medicine, which show that algorithms trained predominantly on male populations tend to misidentify heart attacks in women and that 81% of genome-wide association data derive from people of European ancestry, limiting their predictive accuracy for more diverse populations.⁵ These patterns point not only to demographic imbalance, but also to a deeper misalignment between the andro- and Eurocentric normative claims of precision medicine and the design of its computational foundations.

The root of the problem often lies in upstream data sources. Clinical trials and biomedical datasets have historically excluded women and gender-diverse individuals. A systematic review reports that persistent underrepresentation of women remains in multiple research areas and a lack of sex-stratified analyses, despite higher adverse drug reactions in female patients.⁶ Early-phase trials continue to exclude women due to perceived reproductive risks and the lingering assumption that the male physiology represents the human standard.⁷ A meta-analysis of machine-learning studies in rheumatoid arthritis found that most papers failed to acknowledge sex bias; only three assessed model performance by sex, and none implemented any corrections.⁸ Because DPTs often draw on these sources, their clinical performance may be selectively reliable, optimized for some, but not for all.

Addressing these limitations requires intersectional and inclusive frameworks. A lexicon for “digital health diversity” stresses that categories such as gender, ethnicity, age, and sexual identity must be considered jointly to tackle digital inequity. It also calls for greater attention to social determinants of health, like access to technology and digital literacy.⁹ Empirical work demonstrates that sex-specific models yield more accurate predictions: for instance, classifiers for women outperformed generic models in predicting antiarrhythmic drug response. Katsoulakis et al. emphasize that fair and representative datasets, transparent governance, and protections against socioeconomic inequalities are essential to prevent DPTs from widening health disparities.⁴ Similarly, research on digital twins for female pelvic floor conditions confirms their utility if privacy and access are adequately addressed.¹⁰ These examples illustrate how inclusion, when operationalized, can enhance the precision and safety of DPTs.

In sum, many current DPT systems are limited by narrow population assumptions and inherited data structures. Realizing the promise of truly personalized medicine requires the systematic integration of sex- and gender-specific data, socioeconomic context, intersectional perspectives, and co-creative development processes. This entails active involvement of diverse patient groups in data acquisition, model training, and validation to ensure fairness and inclusivity in future DPT systems.

As an opinion piece, this article draws on evidence from cardiology, endocrinology, mental health, and medical device research to show how demographic biases manifest in DPT-relevant domains. We argue that equity must be embedded from the outset, across design, validation, and governance, if DPTs are to fulfill their promise for all patients.

Evidence across clinical domains

Sex differences in cardiovascular disease

Heart failure is not a single disease but comprises multiple phenotypes. Heart failure with preserved ejection fraction (HFpEF) is more prevalent in women than in men and is associated with distinct risk factors such as obesity, hypertension, pregnancy-related complications, and hormonal changes.¹¹ Women represent approximately 55% of HFpEF patients, but only 29% of those with heart failure with reduced ejection fraction.¹¹ Historically, clinical trials and registries have under-enrolled women, leading to predictive models and therapeutic guidelines that are primarily based on male-dominant data. If DPTs for heart failure rely on these datasets, they may underrecognize HFpEF and misclassify women’s symptoms.

Hormonal cycles and glycemic control

In women with type 1 diabetes, insulin dosing is typically adjusted based on glucose levels, dietary intake, and physical activity. However, hormonal fluctuations over the menstrual cycle also affect how the body responds to insulin and how stable glucose levels remain, factors that are often not considered in both clinical care and model development. In a prospective study using continuous glucose monitoring. At the same time, their bodies responded less effectively to insulin, a condition known as insulin sensitivity, even though total daily insulin doses remained stable.¹² Further evidence from a simulation-based analysis confirmed this pattern. Researchers use a technique known as euglycemic clamp, a gold-standard method to measure how much insulin is needed to keep blood glucose stable. The results showed that during the luteal phase, the glucose infusion rate, a proxy for insulin sensitivity, was significantly lower. This meant that participants spent more time with a glucose level over 180 mg/dl, while the recommended target after meals is

typically below 140 mg/dl after 2 h (postprandial). When algorithms were informed by cycle-related hormonal variations in insulin sensitivity, the model's performance improved, and glucose stability across the cycle increased.¹³ A systematic review supports these findings, noting that a subset of women with type 1 diabetes consistently experiences higher blood sugar levels and reduced insulin sensitivity in the luteal phase compared to the first half of the cycle (the follicular phase).¹⁴

These results imply that DPTs for diabetes that do not include menstrual cycle information may misinterpret hormone-driven blood sugar fluctuations as poor adherence or lifestyle-related failure. This can lead to unnecessary insulin increases and a higher risk of treatment-related complications, such as a reduction in blood glucose concentration (<60 mg/dl; hypoglycemia). Incorporating cycle-phase data and sex-specific physiology into model inputs, along with sex-stratified validation, can ensure that recommendations are personalized and help prevent iatrogenic hypoglycemia.^{12,13}

Mental health AI and gendered patterns of detection

Recent work in computational psychiatry has highlighted that AI systems used to assess mental health often rely on data such as voice recordings, written text, facial expressions, and/or physiological signals. These inputs, however, are shaped by gender, language use, cultural norms, and social context.¹⁵ A recent review of bias in mental health AI warns that emotional expression varies significantly across gender, race, and ethnicity. Tools based on natural language processing (NLP), which aim to detect mood or distress from spoken or written language, often perform differently across demographic groups.¹⁵ For instance, automated speech transcription methods, which are often used as input features, produce more word-detection errors for women and minorities than for non-Hispanic white men.¹⁵ A study found that speech-based recognition systems flagged equal numbers of men and women as being at risk of depression, even though the women in the sample had higher depression rates. This suggests that the system underdiagnosed distress in women, likely due to misaligned language features or model assumptions. Recognizing these limitations, researchers in mental health AI have argued that a one-size-fits-all approach is unrealistic. Instead, models may need to be tailored to specific population groups and trained on datasets that include relevant demographic variables. This allows the system to learn population-specific patterns between behavior and mood,¹⁶ and to detect risk more equitably across groups.

These insights are directly relevant for DPTs designed to simulate mental health trajectories. If such models do not account for differences in gendered communication styles, symptom presentation, and social context, they may fail to recognize clinically significant patterns. Ensuring that training datasets are balanced by sex and gender, including

behavioral and physiological markers specific to different groups and calibrating model outputs accordingly, can help prevent under- or over-diagnosis.¹⁵ Without such safeguards, divergence from model expectations may be wrongly attributed to patient non-compliance or lead to incorrect diagnostic labeling, rather than being understood as a limitation of the system's demographic sensitivity.

Device and measurement biases across skin tones

Digital twins depend not only on clinical and observational data but also on accurate continuous inputs from sensor-based devices. These measurements are often treated objectively, yet the devices themselves can introduce demographic bias. One widely discussed example is the pulse oximeter, a device used to estimate blood oxygen saturation (SpO₂). Studies have shown that readings from pulse oximeters tend to be systematically higher in individuals with darker skin compared to direct arterial blood gas measurements. A review of 22 studies found that this overestimation was consistent and that the variability was greater in Black patients than in white patients.¹⁷ As a result, some Black patients with dangerously low oxygen levels, below 88% arterial saturation, still displayed normal or near-normal SpO₂ values (between 92% and 96%) on the device.¹⁷

If DPTs integrate sensor data without accounting for such measurement bias, they may misjudge a patient's oxygenation status. This could lead to inappropriate clinical recommendations, such as insufficient oxygen therapy or missed escalation of care. Moreover, since pulse oximeters are often used in continuous monitoring and as triggers for clinical interventions, these biases can become embedded in DPT logic without ever being questioned. Ensuring that DPTs function equitably across populations, therefore, requires more than diverse training data. It also means scrutinizing the hardware inputs themselves. Device calibration, validation across skin tones, and the use of alternative measurement technologies should be considered part of demographic robustness, not as afterthoughts, but as foundational design requirements.

Equity by design: Strategies for demographic robust DPTs

These examples demonstrate that demographic variation is not a marginal concern in DPT development. When models are trained on narrow datasets, or when measurement tools systematically misrepresent specific groups, the resulting systems may fail entire patient populations. Addressing these limitations requires more than technical refinement. Therefore, we argue for a systematic shift in how data are collected, how models are developed and validated, and how fairness is defined and operationalized across the DPT lifecycle. To move from diagnostic critique to

actionable reform, we outline in this section, six strategies for designing DPTs that are not only technically sophisticated but also socially responsive, ethically responsible, and clinically equitable:

1. *Diversify datasets and ensure balanced representation.* Efforts to diversify training data should begin with proactive recruitment of underrepresented populations in clinical studies and registries. Multi-institutional collaborations can help capture variation across socio-economic, geographic, and racial contexts. Transparent reporting of data provenance and demographic composition should be encouraged, where legally and ethically feasible. Open science initiatives can support reproducibility and enable external validation. Several large-scale research consortia (e.g., All of Us) have already begun incorporating demographic diversity criteria in data governance.
2. *Integrate sex- and gender-specific knowledge.* Where clinically relevant and supported by high-quality data, physiological variables such as menstrual cycle phase, pregnancy history, hormonal levels, and sex-specific disease phenotypes should be systematically incorporated in model development. For example, cardiovascular applications should reflect the prevalence and distinct pathophysiology of HFpEF in women, and the endocrine model should incorporate menstrual tracking data when informing insulin dosing.
3. *Perform subgroup-stratified evaluation and external validation.* Model performance should be evaluated separately for key demographic subgroups, with attention to intersectional variation where feasible. When disparities are identified, approaches such as reweighting, resampling, adversarial debiasing, or subgroup-specific modeling may be applied. External validation across diverse institutions can help assess generalizability and robustness. While not all institutions may be able to implement full intersectional testing immediately, stratified performance monitoring is increasingly recognized as best practice in clinical AI development.
4. *Foster interdisciplinary and inclusive teams.* Equity-oriented system design requires collaboration between clinicians, data scientists, ethicists, social scientists, and patient advocates, including representatives from historically marginalized communities.¹⁸ Participatory design processes help surface overlooked biases and ensure that performance metrics align with patient priorities. In gender-sensitive DPT development, co-creation is not an accessory; it is a precondition for relevance. Numerous real-world lab approaches and

participatory design frameworks are already available and can be adapted to the DPT context.

5. *Ensure transparency and accountability.* Developers should clearly document data sources, modeling choices, validation procedures, and (subgroup-specific) demographic performance metrics. Transparent reporting enables peer review, regulatory scrutiny, clinical decision-making, and informed trust. Post-deployment monitoring and structured feedback mechanisms are essential to detect emerging biases and update models accordingly.¹⁸ Existing initiatives, such as model cards and datasheets for datasets, provide practical templates for implementing transparency.
6. *Address socio-economic and infrastructural factors.* Equity in digital health is also shaped by access, affordability, and digital literacy. Design process should consider these factors early on. This includes developing low-barrier interfaces, ensuring multilingual access, and accounting for sensor limitations (see skin-tone-related bias in pulse oximetry above). Policy-level interventions, such as public infrastructure for digital inclusion or clinical safety boards for AI, can further support implementation. Not all solutions lie within individual projects, but even local design choices can reduce exclusion risks.

Beyond DPT design: Toward structural change

DPTs hold immense promise for advancing personalized medicine. Yet this promise is contingent on representativeness and quality of the data, assumptions, and measurement tools they rely on. Evidence from cardiology, endocrinology, mental health, and medical devices shows that when women and racialized populations are underrepresented, the consequences are not hypothetical: they include misdiagnoses, misinterpretation of physiological variation, and clinically significant measurement errors.^{2,11,17}

But bias is not an isolated design flaw. It reflects broader systematic patterns of exclusion across biomedical research, clinical trials, and regulatory frameworks, patterns that become operational through model architectures and decision rules. Design is not the origin of these structures, but it is where they are enacted. And because it is a site of enactment, it is also a site of intervention. Nevertheless, more equitable design alone will not suffice. Technical corrections cannot compensate for upstream exclusions in data generation, research prioritization, or health system access. Efforts to build fair and inclusive DPTs must therefore be accompanied, if not preceded, by structural reforms that challenge the conditions under which exclusion is produced and legitimized. Only when design and system change are pursued in tandem can personalization become truly inclusive.

If precision medicine aspires to live up to its name, it must be precise for *everyone*. A digital twin that works flawlessly for the “average” patient but fails those at the margins is not a breakthrough. It is a high-tech repetition of systemic inequity. The choice is not whether DPTs will reflect social values, but which ones. The systems we build will carry forward the structures we fail to question.

Opportunities and future directions

Encouragingly, the DPT paradigm itself can be part of the solution. When these models are deliberately designed for inclusion, they can highlight and even help correct biases in healthcare. For example, incorporating female-specific physiological variations into a diabetes DPT has been shown to improve model accuracy and patient glucose control.¹³ Likewise, digital twin projects focusing on women’s health have demonstrated the technology’s potential: a recent female pelvic-floor digital twin confirmed that such tailored models are both feasible and clinically valuable.¹⁰ These cases exemplify how paying attention to sex and gender can directly enhance the performance and fairness of digital twin models.

More broadly, new initiatives and research efforts are laying the groundwork for more equitable DPTs. Large-scale programs like the NIH’s All of Us are enriching datasets with diverse participants, which will, in turn, improve the robustness of future DPT models. Researchers are also exploring participatory, co-creative development processes to ensure digital twins serve the needs of marginalized communities.² Each strategy outlined in this paper represents an opportunity for innovation: by embracing diverse data, interdisciplinary collaboration, and structural change, DPT developers can build systems that actively reduce disparities. With growing awareness and commitment, the next generation of DPTs could become powerful tools for closing the gender data gap and fostering health equity, ultimately fulfilling the promise of precision medicine for all.

Acknowledgements

The authors would like to thank Dr Theresa Ahrens and the Digital Health Engineering team at Fraunhofer IESE for their valuable input and collaboration. We also acknowledge the Research Group “Life, Innovation, Health, and Technology (LIGHT)” at the Institute for Technology Assessment and Systems Analysis (ITAS), Karlsruhe Institute of Technology (KIT), for their constructive feedback and support.

ORCID iDs

Dana Mahr  <https://orcid.org/0009-0005-2214-7761>

Nora Weinberger  <https://orcid.org/0000-0002-1148-7470>

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was funded by the European Union under *Horizon Europe*: IANUS—INspiring and ANchoring TrUst in Science, Grant Agreement No. 101058158 (Project duration: 1 June 2022–31 May 2025). Further information: CORDIS record 101058158.

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

1. Joshi A. Big data and AI for gender equality in health: bias is a big challenge. *Front Big Data* 2024; 7: 1436019.
2. Weinberger N, Hery D, Mahr D, et al. Beyond the gender data gap: co-creating equitable digital patient twins. *Front Digit Health* 2025; 7: 1584415.
3. Guerrero Quiñones JL and Puzio A. Digital twins for trans people in healthcare: queer, phenomenological and bioethical considerations. *J Med Ethics* Epub ahead of print 4 March 2025. DOI: 10.1136/jme-2024-110403.
4. Katsoulakis E, Wang Q, Wu H, et al. Digital twins for health: a scoping review. *NPJ Digit Med* 2024; 7: 77.
5. Norori N, Hu Q, Aellen FM, et al. Addressing bias in big data and AI for health care: a call for open science. *Patterns (N Y)* 2021; 2: 100347.
6. Goldstein KM, Kung LCY, Dailey SA, et al. Strategies for enhancing the representation of women in clinical trials: an evidence map. *Syst Rev* 2024; 13. doi:10.1186/s13643-023-02408-w
7. Waltz M, Lyerly AD and Fisher JA. Exclusion of women from phase I trials: perspectives from investigators and research oversight officials. *Ethics Hum Res* 2023; 45: 19–30.
8. Talwar A, Turner S, Maw C, et al. Sex bias consideration in healthcare machine-learning research: a systematic review in rheumatoid arthritis. *BMJ Open* 2025; 15: e086117.
9. Sharma Y, Saha A and Goldsack JC. Defining the dimensions of diversity to promote inclusion in the digital era of health care: a lexicon. *JMIR Public Health Surveill* 2024; 10: e51980.
10. Egorov V. Digital twin of the female pelvic floor. *Open J Obstet Gynecol* 2024; 14: 1687–1694.
11. Kaur G and Lau E. Sex differences in heart failure with preserved ejection fraction: from traditional risk factors to sex-specific risk factors. *Womens Health (Lond)* 2022; 18: 17455057221140209.
12. Brown SA, Jiang B, McElwee-Malloy M, et al. Fluctuations of hyperglycemia and insulin sensitivity are linked to menstrual cycle phases in women with T1D. *J Diabetes Sci Technol* 2015; 9: 1192–1199.
13. Díaz CJL, Fabris C, Breton MD, et al. Insulin replacement across the menstrual cycle in women with type 1 diabetes: an in silico assessment of the need for ad hoc technology. *Diabetes Technol Ther* 2022; 24: 832–841.

14. Gamarra E and Trimboli P. Menstrual cycle, glucose control and insulin sensitivity in type 1 diabetes: a systematic review. *J Pers Med* 2023; 13: 74.
15. Timmons AC, Duong JB, Simo Fiallo N, et al. A call to action on assessing and mitigating bias in artificial intelligence applications for mental health. *Perspect Psychol Sci* 2023; 18: 1062–1096.
16. Adler DA, Stamatis CA, Meyerhoff J, et al. Measuring algorithmic bias to analyze the reliability of AI tools that predict depression risk using smartphone-sensed behavioral data. *NPJ Ment Health Res* 2024; 3: 17.
17. Al-Halawani R, Charlton PH, Qassem M, et al. A review of the effect of skin pigmentation on pulse oximeter accuracy. *Physiol Meas* 2023; 44: 05TR01.
18. Nazer LH, Zatarah R, Waldrup S, et al. Bias in artificial intelligence algorithms and recommendations for mitigation. *PLOS Digit Health* 2023; 2: e0000278.