The Patient as a Prediction: Four Epistemological Challenges of Digital Twins in Health

Martina Baltuzzi[1,2,3]

[1]University of Bologna, Italy
[2]University of Torino, Italy
[3]KU Leuven, Belgium

Abstract

Health Digital Twins (HDTs) are continuously updated, patient-specific computational models designed to simulate physiological states and anticipate future clinical trajectories. They represent one of the most ambitious expressions of the turn towards personalised medicine. This paper examines them from an epistemological perspective, asking what kind of knowledge HDTs produce and on what terms that knowledge can be trusted. To do so, the paper adopts the framework developed by Matsumi and Solove in The Prediction Society, which identifies four problems raised by algorithmic predictions: the fossilisation problem, the unfalsifiability problem, the preemptive intervention problem, and the self-fulfilling prophecy problem. The paper argues that, in the context of HDTs, these four problems form a compounding cycle, driven by the continuous and bidirectional relationship between the physical patient and the virtual model; with each iteration of this cycle, the model's prior assessment of the patient becomes increasingly difficult to challenge or revise. Thus, the epistemological challenges posed by HDTs cannot be solved by only providing better or more representative data, as they stem from the structure of algorithmic prediction itself.

Keywords

Digital twins, Epistemology, Algorithmic predictions, Healthcare

1. Introduction

Modern medicine is being reshaped by the rapid advancement of artificial intelligence and data-driven technologies, one of the most significant expressions of which is the ambition to move beyond population-based treatment protocols towards personalised care. Within this context, Health Digital Twins (HDTs) have emerged as a particularly promising innovation: conceived as continuously updated, patient-specific computational models, HDTs aim to create a dynamic virtual representation of a patient or of specific physiological systems. By promising to support personalised medicine through simulation and prediction, they raise questions not only about medical practice but also about the nature and limits of the knowledge they produce. This article first outlines the characteristics of digital twins and their applications in healthcare (Section 2), then examines their epistemological implications through the framework developed by Matsumi and Solove (section 3), addressing the fossilisation problem (3.1), the unfalsifiability problem (3.2), the preemptive intervention problem (3.3), and the self-fulfilling prophecy problem (3.4), before grasping them together (section 4) and drawing conclusions (section 5).

2. Digital twins and medical applications

The concept of digital twin was first articulated in the context of NASA's space program as a method for creating a high-fidelity virtual replica of a physical system, updated in real time from sensor data[1]. In the early 2000s, Michael Grieves introduced his "conceptual ideal for PLM"[2] at the University of Michigan, comprising all the elements that today define a digital twin, namely i) a physical entity

BEWARE-26 (Bias, Risk, Explainability and the role of Logic and Logic Programming), co-located with AIxIA 2026, October 8, 2026, Perugia, Italy

martina.baltuzzi2@unibo.it (M. Baltuzzi)
https://www.unibo.it/sitoweb/martina.baltuzzi2/en (M. Baltuzzi)
0009-0008-3932-2644 (M. Baltuzzi)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-11-17

CEUR-WS.org/Vol-4332/paper4.pdf

capable of performing tasks; ii) a virtual model of the physical entity; iii) an exchange of data and information that ties the real and virtual entities together[3]. Digital twins are accordingly understood as "a digital replica or representation of a physical object, process, or service"[4]. Their defining feature, which distinguishes them from earlier simulation models, is bidirectionality: data flows continuously from the physical system to the virtual replica, and the replica's outputs flow back to inform decisions about the physical system.
The translation of the digital twin into medicine has accelerated considerably in recent years, with the aim to construct patient-specific virtual models capable not only of representing physiological states but also of anticipating risks, disease trajectories and treatment outcomes[5]. In the EU context, within the EDITH initiative, a coordination and support action funded by the Digital Europe program, the Virtual Human Twin (VHT) has been defined as "a patient-specific virtual representation of real-world systems or processes [...] that can be used as a clinical decision-support system, a personal health forecasting tool or as a tool for the development and personalisation of medical products"[6].
In general, in the HDT, the physical entity corresponds to the patient or to parts of their body, while data collected from the patient, integrating diverse kinds of information, such as clinical and physiological data, feed the virtual replica and enables simulation. A key aspect of this potential is the possibility of testing clinical decisions in a virtual environment before applying them to the patient[7]. HDTs have found applications across diverse domains of medicine, from monitoring cardiovascular diseases and supporting surgical procedures, to shortening pharmaceutical processes[8], designing customised medical devices, and enabling precision oncology[7]. Across these applications, HDTs rely on algorithmic predictions, which raise the epistemological questions examined in the following section.

3. The epistemological lens: Matsumi and Solove's four problems

While traditional medical records document past and present conditions of patients, HDTs aim to use virtual representations to generate personalised predictions about future health trajectories[9], thus raising distinctive questions that stem from their predictive nature[10]. As HDTs may guide diagnoses and treatments and directly affect patients, epistemological questions arise regarding both the accuracy of their virtual representation and the reliability of algorithmic predictions. Moreover, once predictions are incorporated into clinical decision-making, they may influence the very outcomes they seek to forecast: in this sense, the risk of deploying predictive systems such as HDTs is that they may not merely observe reality, but actively participate in shaping it.
To analyse these issues, this paper adopts the framework developed by Matsumi and Solove in The Prediction Society[11], in which the authors identified and grouped the unique risks raised by algorithmic predictions into four problems: i) the fossilization problem; ii) the unfalsifiability problem; iii) the preemptive intervention problem; iv) the self-fulfilling prophecy problem. Although developed in the context of algorithmic prediction more broadly, these four problems provide a useful epistemological lens for examining the use of HDTs. The following sections develop this analysis.

3.1. The fossilization problem

Matsumi and Solove first discuss the fossilization problem, according to which individuals may be unequally treated on the basis of past patterns that predictive systems suppose will persist into the future. The authors define fossilization as the reification of past facts by casting them into the future, making the past harder for a person to escape[11]. Their central illustration is Amazon's hiring algorithm, trained on resumes of previously successful employees and therefore reproducing the gender imbalance among them rather than correcting it[12]. The example serves to highlight how the data available to predictive models is not a neutral record of the world: instead, what gets collected (and what does not) often reflects existing preferences and biases in society.
The same concern arises in healthcare: studies have shown that IBM's Watson for Oncology – a product designed to support physicians and patients by offering personalized cancer treatment options – performed less accurately when applied to non-Western patients, as it was trained primarily on data

from Western populations[13].
In the context of HDTs, fossilization may affect both the representation of the patient and the prediction of their future clinical trajectory. The "normal ranges" against which a patient's measurements are assessed, as well as the datasets used to calculate future risks, are indeed themselves products of prior decisions about which data to collect. These decisions can be shaped by specific priorities or prejudices, or structural factors such as unequal access to healthcare, whereby certain population groups generate more data than others. As a result, the twin may reproduce and continuously reinforce assumptions embedded in its training data, failing to account adequately for already marginalised groups. Fossilization therefore highlights how HDTs risk perpetuating forms of algorithmic discrimination, reinforcing existing inequalities under the appearance of objective prediction.

3.2. The unfalsifiability problem

The second problem identified by Matsumi and Solove, the unfalsifiability problem, concerns the impossibility of verifying the accuracy of an algorithmic prediction before the predicted event occurs. Because predictions make assertions about an uncertain future, individuals who wish to challenge them face a fundamental difficulty: there is no fact yet in existence against which the forecast can be tested. The authors illustrate this threat through the film Gattaca, in which a man is barred from a career because his genetic profile predicts a higher likelihood of heart disease, with no way for him to disprove the prediction. This highlights the fact that an algorithmic prediction is neither true nor false at the moment it is made, because the matter it asserts has not yet occurred[11].
This is, of all four problems, the one whose paradigmatic example is already drawn from healthcare. However, unlike the single prediction depicted in Gattaca, an HDT can produce several forecasts and – through the continuous bidirectional exchange of data between the physical patient and the virtual model – it can update and reassert them over time, presenting the clinician with new claims that cannot yet be checked against anything.

3.3. The preemptive intervention problem

The third issue identified by Matsumi and Solove is the preemptive intervention problem: this refers to situations in which decisions are made on the basis of a prediction before the predicted event occurs, which has the effect of destroying the evidence that would have been needed to assess whether the forecast was accurate or not. The authors illustrate this threat through the example of a bus company that uses a predictive algorithm to flag high-risk drivers and send them to additional safety training. If, after the training, a flagged driver does not have an accident, there is no way to determine whether the original prediction was mistaken or, on the contrary, whether it was correct and the training intervention was what prevented the accident[11].
Substitute a patient for the driver. A digital twin predicts that, left untreated, a patient's condition will deteriorate by a given date; his clinician intervenes immediately on that basis, and she does not deteriorate. Whether the intervention was necessary, or whether she would have recovered regardless, becomes impossible to verify, as the intervention removes the only test that could have distinguished the two possibilities.
In the clinical context, however, this problem carries an additional risk: a clinician who trusts the HDT's prediction may feel compelled to act even in cases where watchful waiting would have been appropriate, for fear that not intervening would expose the patient to the harm the model predicted. Through personalised virtual modelling of the specific patient being treated, the twin's prediction may appear more authoritative than the clinician's own judgment, and therefore it would intensify rather than reduce the tendency to treat preemptively. Patients may thus risk being subjected to procedures whose necessity can no longer be demonstrated once the decision is made, while the untreated trajectory the twin warned against can never be observed, as the only way to check it would be against the patient's actual future without intervention.

3.4. The self-fulfilling prophecy problem

Lastly, Matsumi and Solove identify the self-fulfilling prophecy problem, according to which algorithmic predictions have the power to shape the future in conformity with their forecast, regardless of their accuracy, as decisions made on their basis are more likely to bring about the predicted outcome[11]. The authors draw on the Pygmalion effect, which holds that individuals tend to perform in ways that reflect the expectations placed upon them. In Robert Rosenthal and Lenore Jacobson's study, teachers were informed that a randomly selected group of students had been identified as likely to experience significant intellectual development; despite the absence of any actual difference between these students and their peers, the selected students subsequently showed greater improvements in IQ scores, apparently because the teachers' raised expectations changed how they taught them[14].
The same mechanism may operate in clinical settings, where the consequences can be considerably more serious. In neuroprognostication, Mertens et al. illustrate this threat through the case of Hans, a patient who suffered a cardiac arrest and was assigned a poor neurological prognosis, on the basis of which life-sustaining treatment was withdrawn[15]. Had the prognosis been mistaken, the error could never have been discovered, since the clinical decision itself excluded any possibility of observing what would have happened otherwise. In this case, the prediction appears confirmed not because it was accurate, but because acting on it prevented any evidence to the contrary from emerging.
The use of HDTs risks intensifying this dynamic. A twin that flags a patient as a poor candidate for treatment can shift the urgency and attention she receives from her care team, thus contributing to the outcome it predicted and undermining the capacity to evaluate and improve the system in the case of errors. That outcome then becomes training data, making the next version of the model more confident in the same recommendation for the next similar patient. In this scenario, errors, rather than being corrected, are absorbed into the model and reproduced at scale[11].

4. Reading the four problems together

The four problems identified by Matsumi and Solove do not operate independently in the context of HDTs, instead, they form a self-reinforcing cycle: a prediction makes a claim about a future event that cannot yet be verified (unfalsifiability); acting on that prediction before the event occurs eliminates the evidence that would have allowed its accuracy to be assessed (preemptive intervention); the resulting outcome, partly shaped by the intervention itself, is then incorporated into the model as confirmation of the original forecast (self-fulfilling prophecy). That updated model then issues a new prediction, restarting the cycle. Fossilisation operates throughout the sequence: the historical patterns, biases, and assumptions embedded in the training data shape the initial prediction and are carried forward through each successive update, so that the cycle may progressively reinforce the very conditions that produced it. The resulting dynamic can be represented as:
Prediction → Intervention → Outcome → New Data → Updated Prediction → Prediction
The relevance of this pattern is highlighted by the nature of HDTs: while a one-time algorithmic prediction passes through this cycle once, a digital twin is continuously updated through the bidirectional flow of data between the physical patient and the virtual model, causing the cycle to repeat over time. With each iteration, the model's previous assessment of the patient can become increasingly difficult to challenge or revise. This dynamic cannot be fixed just through improvements in data quality or availability, as greater predictive accuracy does not overcome the fundamental issue that predictions can inform decisions which, in turn, alter the very future against which those predictions would otherwise be evaluated.

5. Conclusion

This article has drawn on Matsumi and Solove's epistemological framework to examine Health Digital Twins, showing how the four problems they identify become more pronounced when applied to a system

that continuously models a specific patient over time. Beyond questions of technical performance, future research should address the broader epistemological implications of Health Digital Twins through interdisciplinary collaboration. Empirical studies could investigate how clinicians interpret and rely on HDT-generated predictions under conditions of uncertainty, while conceptual research could explore mechanisms for improving the explainability and contestability of predictive outputs. Comparative analyses of different categories of HDTs may further clarify whether distinct applications generate distinct epistemic risks.

Declaration on Generative AI

During the preparation of this work, the author used Claude in order to grammar and spelling check. After using this tool, the author reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] R. Rosen, G. Von Wichert, G. Lo, K. D. Bettenhausen, About The Importance of Autonomy and Digital Twins for the Future of Manufacturing, IFAC-PapersOnLine 48 (2015) 567–572. URL: https://linkinghub.elsevier.com/retrieve/pii/S2405896315003808. doi:10.1016/j.ifacol.2015.06.141.

[2] M. Grieves, Origins of the Digital Twin Concept, Working Paper, 2016. doi:10.13140/RG.2.2.26367.61609.

[3] M. Grieves, Digital twin: Manufacturing excellence through virtual factory replication, Technical Report, 2014.

[4] M. N. Kamel Boulos, P. Zhang, Digital Twins: From Personalised Medicine to Precision Public Health, Journal of Personalized Medicine 11 (2021) 745. URL: https://www.mdpi.com/2075-4426/11/8/745. doi:10.3390/jpm11080745.

[5] M. Morrison, Personalised Medicine; A Critical Approach to Data-Driven Medicine, Routledge, 2026.

[6] European Virtual Human Twin, 2026. URL: https://www.edith-csa.eu/edith/.

[7] K. Zhang, H.-Y. Zhou, D. T. Baptista-Hon, Y. Gao, X. Liu, E. Oermann, S. Xu, S. Jin, J. Zhang, Z. Sun, Y. Yin, R. M. Razmi, A. Loupy, S. Beck, J. Qu, J. Wu, Concepts and applications of digital twins in healthcare and medicine, Patterns 5 (2024) 101028. URL: https://linkinghub.elsevier.com/retrieve/pii/S2666389924001612. doi:10.1016/j.patter.2024.101028.

[8] T. Sun, H. Xiwang, X. Song, L. Shu, Z. Li, The Digital Twin in Medicine: A Key to the Future of Healthcare?, Frontiers in Medicine 9 (2022) 907066. doi:10.3389/fmed.2022.907066.

[9] T. Erol, A. Mendi, D. Dogan, The Digital Twin Revolution in Healthcare, 2020. doi:10.1109/ISMSIT50672.2020.9255249, pages: 7.

[10] M. Durante, U. Pagallo (Eds.), The De Gruyter handbook on law and digital technologies, De Gruyter Handbooks of Digital Transformation, De Gruyter, Berlin ; Boston, 2025. URL: https://www.degruyter.com/document/doi/10.1515/9783111346632.

[11] H. Matsumi, D. J. Solove, The Prediction Society: AI and the Problems of Forecasting the Future, SSRN Electronic Journal (2023). URL: https://www.ssrn.com/abstract=4453869. doi:10.2139/ssrn.4453869.

[12] A. Aloisi, The Quantified Worker: Law and Technology in the Modern Workplace, by Ifeoma Ajunwa, International Labour Review 163 (2024) 163–166. URL: https://en.ilr-rit.org/article/pubid/163-5/. doi:10.1111/ilr.12410.

[13] C. Liu, X. Liu, F. Wu, M. Xie, Y. Feng, C. Hu, Using Artificial Intelligence (Watson for Oncology) for Treatment Recommendations Amongst Chinese Patients with Lung Cancer: Feasibility Study, Journal of Medical Internet Research 20 (2018) e11087. URL: http://www.jmir.org/2018/9/e11087/. doi:10.2196/11087.

[14] R. Rosenthal, L. Jacobson, Pygmalion in the Classroom: Teacher Expectation and Pupils' Intellectual Development, Holt, Rinehart and Winston, New York, 1968.

[15] M. Mertens, O. C. King, M. J. A. M. van Putten, M. Boenink, Can we learn from hidden mistakes? Self-fulfilling prophecy and responsible neuroprognostic innovation, Journal of Medical Ethics 48 (2022) 922–928. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC9626909/. doi:10.1136/medethics-2020-106636.