Year: 2024

The Patient Preference Predictor: A Timely Boost for Personalized Medicine

Biller-Andorno, Nikola ; Ferrario, Andrea ; Biller, Armin

DOI: https://doi.org/10.1080/15265161.2024.2353029

THE AMERICAN JOURNAL OF BIOETHICS
2024, VOL. 24, NO. 7, 35–38
https://doi.org/10.1080/15265161.2024.2353029

Taylor & Francis
Taylor & Francis Group

OPEN PEER COMMENTARIES

The Patient Preference Predictor: A Timely Boost for Personalized Medicine

Nikola Biller-Andorno[a], Andrea Ferrario[a] and Armin Biller[b]

[a]University of Zurich; [b]Heidelberg University

The future of medicine will be predictive, preventive, personalized, and participatory. Recent technological advancements bolster the realization of this vision, particularly through innovations in genomics and biotechnology. While the domains of predictive, preventive, and participatory medicine have received substantial attention, personalization often remains underexplored beyond the molecular level, despite the longstanding recognition of patient-centeredness as a central tenet of biomedical ethics and core domain of healthcare quality.

The Patient Preference Predictor (PPP) is a great chance to capture the values and preferences that inform individuals' expectations regarding their future healthcare, strengthening personalized medicine in the full sense of the term. The PPP comes in different shapes and colors, ranging from population-based statistical models to the use of machine learning (ML) and generative AI "to extract patients' values or preferences from individual-level material" (the Personalized Patient Preference Predictor, or P4) (Earp et al. 2024, 15). Systems could also be combined, so that P4 data could be used to fine-tune individual predictions based on population data or vice versa (Biller-Andorno and Biller 2019).

These concepts and their possible combinations differ with a view to their challenges and limitations. Whereas in an exploratory phase it is quite appropriate to be thinking in many possible directions, a critical assessment of the technical feasibility and ethical desirability of the P4 will require clarity about specific features, such as the nature of the models (static vs. self-learning); their complexity and explainability; the kind of data these systems are based on and their sources; the level of automation and the role humans play in the development and deployment of the models.

Whereas population-based predictions struggle with the issue of stereotyping individuals, building predictive models on an individual's past decisions will have to deal with a lot of noise that might be confusing and misleading. On the other hand, if such a digital twin was trained on an individual's choices (and, ideally, their satisfaction with those choices) over a sustained period, maybe even a lifetime, it is not unlikely it could reliably predict an individual's healthcare choices as least as well as next of kin (Iqbal, Krauthammer, and Biller-Andorno 2022). The use case for such interactive "decision-making companions" could inspire and support individual reflection on advance care choices—a feature that could be integrated into digital advance directives and advance care planning (Biller-Andorno and Biller 2019; Biller-Andorno and Biller 2021; Biller and Biller-Andorno 2023).

AI-based individualized decision-support is part of our everyday lives. While it is most notable today in consumer tips (e.g., online shopping) and health-related recommendations (e.g., on sleep or nutrition), it is likely that it will soon extend to medical decisions. Bringing together the vast amount of health-related personal data that are available in electronic health records with individual preferences and choices documented in online services and social media would in principle seem to provide material for a P4. This system "could be queried in real-time as to the patient's most likely preferences for treatment in any given healthcare crisis" (Earp et al. 2024, 15).

There is considerable potential for personalized patient preference predictions to create value for patients and to contribute to better care. However, at this stage, a number of technical questions need be addressed. For example, for a P4, fine-tuning a large language model (LLM) is key to generate robust, individualized predictions. The recent Coral study, conducted on an expert-annotated oncological dataset, has demonstrated that while zero-shot performances

CONTACT Nikola Biller-Andorno biller-andorno@ibme.uzh.ch Institute of Biomedical Ethics and History of Medicine (IBME), University of Zurich Main Navigation, University of Zurich, Zurich, Switzerland.

of LLMs are impressive, their accuracy is not sufficient in clinical settings. It also revealed that even small changes to the prompt can have a significant impact on their performance (Sushil et al. 2024). The lack of robustness of LLMs is a problem in medical applications (Ferrario and Biller-Andorno 2024). How the fine-tuning of a robust P4 would look like remains to be clarified. Crucially, there is still no consensus on the procedures to evaluate the performance of these models.

There are also legal questions to be addressed, including liability issues and the potential classification of a P4 as medical device. The fine-tuning of individualized LLMs requires managing a daunting number of ML pipelines and poses serious risks related to data privacy. Further, if more traditional ML algorithms already suffer from limited explainability ("black box" problem), in the case of P4 this challenge is multiplied manifold by the use of LLMs. The complexity of their architecture makes simple explanations of their predictions particularly difficult. The development of P4 will need to be carefully monitored to ensure that these systems comply with applicable laws and standards.

Any further exploration of the feasibility and desirability of designing, developing, and deploying a PPP– and, a fortiori, a P4–model will require an intense interdisciplinary collaboration of ethicists, AI experts, healthcare providers, lawyers, and social scientists, with a strong patient and public involvement. The best way to discover and tackle relevant issues might be to try building such a system (Ferrario, Gloeckler, and Biller-Andorno 2023a, 2023b). Even well-documented failures and aborted attempts would be highly instructive.

When should such a system be considered good enough? It has been argued that a PPP would need to be only slightly better than chance, given that human surrogate decision-making seems to be no more accurate than that (Earp et al. 2024). Even if this skeptical assessment of surrogates' performance was true, we believe that if a P4 were to be used not only as a substitute for absent surrogates but also to question their judgements or resolve disputes among them, it should demonstrably exceed the performance of surrogates. That said, we believe that a P4 should primarily be used to enhance surrogates' decision-making process and facilitate consensus, rather than acting as a superior authority to adjudicate conflicts.

High standards for performance would be desirable not only because an AI replacing or overriding human judgment is a highly sensitive issue that requires a robust justification. If P4 were to become an established part of AI-based clinical decision-support systems, their output would likely carry considerable weight. For clinicians, to depart from what a P4 suggests would take considerable resolve, even if it would formally only be providing likelihoods of preferences, possibly followed by a recommendation. Given that professionals might be quite inclined to trust such a system, particularly if it has gone through a certification process, and that important decisions are at stake, high standards should apply regarding accuracy and reliability (Biller-Andorno and Biller 2021).

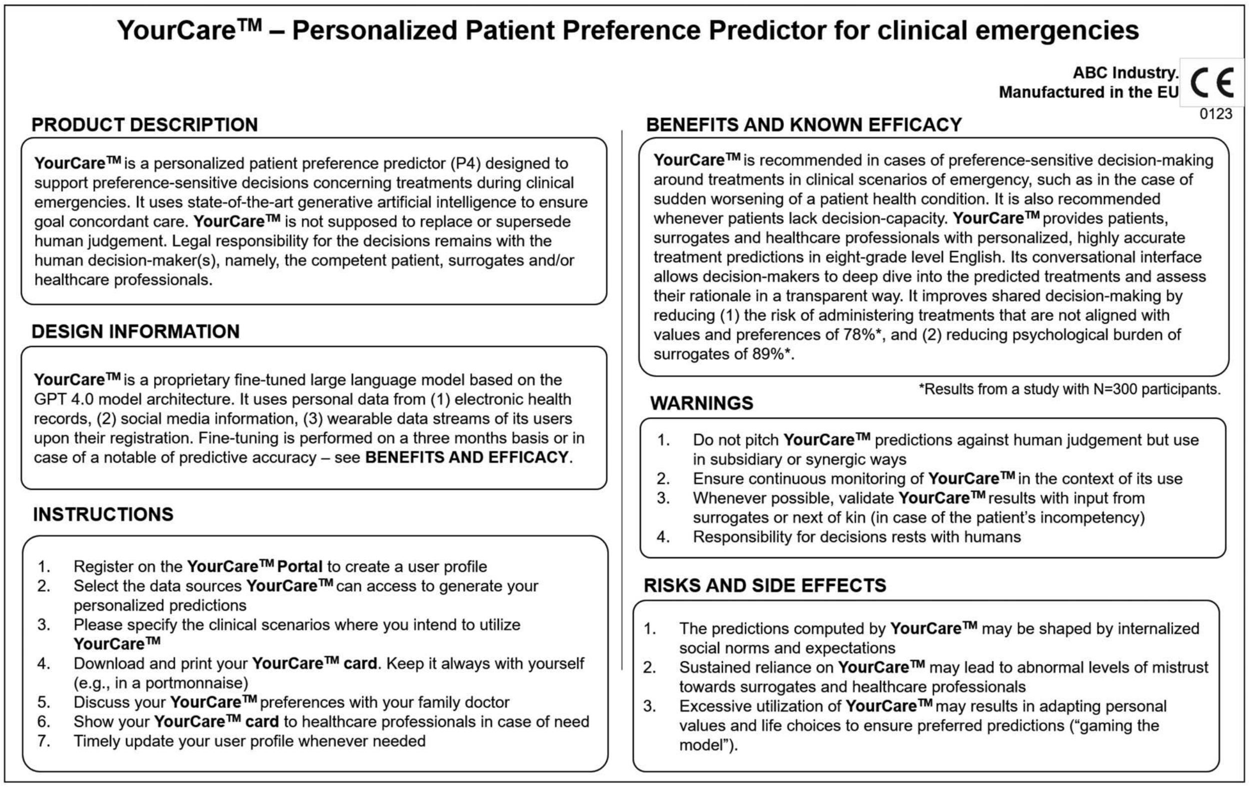
Predictive accuracy is, however, not the only performance measure that could be considered. In addition to convincing user experience in the respective target groups, it would be of interest to see if a P4 that was integrated into clinical routine could help improve goal-concordant care. An even more demanding measure would be a decline in decisional regret regarding treatment choices, as perceived by patients (if possible) or by surrogates. All these studies are methodologically challenging and will demand significant resources.

Let's assume we had a carefully developed and evaluated system. Users–healthcare professionals, surrogates and citizens completing their advance directive–would benefit from a "package insert" that explains how a P4 functions, what it can offer them, and what its risks and limitations are. Such a "package insert" would need to be formulated in a way that laypersons can grasp—see Figure 1 for an illustrative, non-exhaustive example.

In conclusion, the perspectives of P4 are as promising as they are concerning. Errors could have fatal consequences and bias may aggravate inequitable access to goal-concordant care. Thus, particular care will be required with regard to the design, development, implementation and continued evaluation of P4.

Precisely because such systems may come massively and fast, given the pace of AI uptake in medicine, there is a need for swift but careful scrutiny, if we do not want to leave this field to market-driven solutions. It may be wise to start out with a P4 as a conversational reflection prompter that allows users to interrogate their own preferences and values. This P4 conversational agent could feature in a decision aid for advance directives. If this worked to users' satisfaction, more challenging use cases probing the role of a P4 in surrogate decision-making could be addressed.

At this stage, we should encourage a rich pool of approaches for pilot testing. Such efforts would need to be accompanied by the development of evaluation standards and well-designed intervention studies focusing on the most promising examples. In



YourCare™ – Personalized Patient Preference Predictor for clinical emergencies

ABC Industry.
Manufactured in the EU CE
0123

PRODUCT DESCRIPTION

YourCare™ is a personalized patient preference predictor (P4) designed to support preference-sensitive decisions concerning treatments during clinical emergencies. It uses state-of-the-art generative artificial intelligence to ensure goal concordant care. **YourCare™** is not supposed to replace or supersede human judgement. Legal responsibility for the decisions remains with the human decision-maker(s), namely, the competent patient, surrogates and/or healthcare professionals.

DESIGN INFORMATION

YourCare™ is a proprietary fine-tuned large language model based on the GPT 4.0 model architecture. It uses personal data from (1) electronic health records, (2) social media information, (3) wearable data streams of its users upon their registration. Fine-tuning is performed on a three months basis or in case of a notable of predictive accuracy – see **BENEFITS AND EFFICACY**.

INSTRUCTIONS

1. Register on the **YourCare™ Portal** to create a user profile
2. Select the data sources **YourCare™** can access to generate your personalized predictions
3. Please specify the clinical scenarios where you intend to utilize **YourCare™**
4. Download and print your **YourCare™ card**. Keep it always with yourself (e.g., in a portmonnaise)
5. Discuss your **YourCare™** preferences with your family doctor
6. Show your **YourCare™ card** to healthcare professionals in case of need
7. Timely update your user profile whenever needed

BENEFITS AND KNOWN EFFICACY

YourCare™ is recommended in cases of preference-sensitive decision-making around treatments in clinical scenarios of emergency, such as in the case of sudden worsening of a patient health condition. It is also recommended whenever patients lack decision-capacity. **YourCare™** provides patients, surrogates and healthcare professionals with personalized, highly accurate treatment predictions in eight-grade level English. Its conversational interface allows decision-makers to deep dive into the predicted treatments and assess their rationale in a transparent way. It improves shared decision-making by reducing (1) the risk of administering treatments that are not aligned with values and preferences of 78%*, and (2) reducing psychological burden of surrogates of 89%*.

*Results from a study with N=300 participants.

WARNINGS

1. Do not pitch **YourCare™** predictions against human judgement but use in subsidiary or synergic ways
2. Ensure continuous monitoring of **YourCare™** in the context of its use
3. Whenever possible, validate **YourCare™** results with input from surrogates or next of kin (in case of the patient's incompetency)
4. Responsibility for decisions rests with humans

RISKS AND SIDE EFFECTS

1. The predictions computed by **YourCare™** may be shaped by internalized social norms and expectations
2. Sustained reliance on **YourCare™** may lead to abnormal levels of mistrust towards surrogates and healthcare professionals
3. Excessive utilization of **YourCare™** may results in adapting personal values and life choices to ensure preferred predictions ("gaming the model").

Figure 1. "Package insert" for a P4, called YourCare™, to be used in clinical emergencies.

addition, understanding how preferences are shaped in different societal contexts would help characterize biases, internalized social pressure or other factors that undermine the articulation of preferences as an expression of a person's autonomy.

Such efforts need to be appropriately funded. It is widely accepted that personalized medicine is an expensive but worthwhile undertaking. From an ethics perspective, it seems obvious that such efforts need to include individual treatment preferences and overall goals of care. This requires the allocating resources to move from the stimulating yet constrained phase of hypothesizing P4 outcomes to clearly determine what actually works.

DISCLOSURE STATEMENT

No potential conflict of interest was reported by the author(s).

FUNDING

The author(s) reported there is no funding associated with the work featured in this article.

ORCID

Nikola Biller-Andorno http://orcid.org/0000-0001-7661-1324
Andrea Ferrario http://orcid.org/0000-0001-9968-9474
Armin Biller http://orcid.org/0000-0001-6412-2670

REFERENCES

Biller, A., and N. Biller-Andorno. 2023. From text to interaction: The digital advance directive method for advance directives. *Digital Health* 9:20552076221147414. doi:10.1177/20552076221147414.

Biller-Andorno, N., and A. Biller. 2019. Algorithm-aided prediction of patient preferences—an ethics sneak peek. *The New England Journal of Medicine* 381 (15):1480–5. doi:10.1056/NEJMms1904869.

Biller-Andorno, N., and A. Biller. 2021. The advance care compass – a new mechanics for digitally transforming advance directives. *Frontiers Digital Health* 3:753747.

Biller-Andorno, N., A. Ferrario, S. Joebges, T. Krones, F. Massini, P. Barth, G. Arampatzis, and M. Krauthammer. 2022. AI support for ethical decision-making around resuscitation: Proceed with care. *Journal of Medical Ethics* 48 (3):175–83. doi:10.1136/medethics-2020-106786.

Earp, B. D., S. Porsdam Mann, J. Allen, S. Salloch, V. Suren, K. Jongsma, M. Braun, D. Wilkinson, W. Sinnott-Armstrong, A. Rid, et al. 2024. A personalized patient preference predictor for substituted judgments in healthcare: Technically feasible and ethically desirable. *The American Journal of Bioethics* 24 (7):13–26. doi:10.1080/15265161.2023.2296402.

Ferrario, A., and N. Biller-Andorno. 2024. Large language models in medical ethics: Useful but not expert. *Journal of Medical Ethics*. Advance online publication. doi:10.1136/jme-2023-109770.

Ferrario, A., S. Gloeckler, and N. Biller-Andorno. 2023a. AI knows best? Avoiding the traps of paternalism and other pitfalls of AI-based patient preference prediction. *Journal of Medical Ethics* 49 (3):185–6. doi:10.1136/jme-2023-108945.

Ferrario, A., S. Gloeckler, and N. Biller-Andorno. 2023b. Ethics of the algorithmic prediction of goal of care preferences: From theory to practice. *Journal of Medical Ethics* 49 (3):165–74. doi:10.1136/jme-2022-108371.

Iqbal, J., M. Krauthammer, and N. Biller-Andorno. 2022. The use and ethics of digital twins in medicine. *The Journal of Law, Medicine & Ethics: A Journal of the American Society of Law, Medicine & Ethics* 50 (3):583–96. doi:10.1017/jme.2022.97.

Sushil, M., V. E. Kennedy, D. Mandair, B. Y. Miao, T. Zack, and A. J. Butte. 2024. CORAL: Expert-curated oncology reports to advance language model inference. *NEJM AI* 1 (4). doi:10.1056/AIdbp2300110.

THE AMERICAN JOURNAL OF BIOETHICS
2024, VOL. 24, NO. 7, 38–40
https://doi.org/10.1080/15265161.2024.2353023

Taylor & Francis
Taylor & Francis Group

OPEN PEER COMMENTARIES

Weighing Patient Preferences: Lessons for a Patient Preferences Predictor

Ben Schwan

Case Western Reserve University

INTRODUCTION

A Patient Preference Predictor (PPP)—an algorithm capable of predicting, on the basis of demographic or more personalized data, what an incapacitated patient would prefer were they capacitated—is a promising tool for guiding the care of patients whose treatment preferences are not clear. But, as with any tool, it might be wielded well or poorly. In this commentary, I will briefly sketch an account of why and how patients' preferences matter, then draw on this account to illustrate both the potential perks and pitfalls of utilizing PPPs in practice.[1]

To this end, I will be setting aside many of the concerns that typically occupy those interested in the ethics of PPPs—about how relevant training data is acquired, about PPPs' expected accuracy, about prerequisites for their use, etc.[2] Instead, my goal is to scrutinize why we care about the thing a PPP predicts and consider how this should inform the way we *use* our PPP when in fact we use it.

WHY PATIENT PREFERENCES MATTER

Roughly, patient preferences matter because autonomy matters. There is a weighty reason to respect patient autonomy, and respecting a patient's autonomy requires deferring to their preferences—their expressed treatment preference when they have capacity and their hypothetical treatment preference when they lack it.[3] But this rough characterization is only roughly right because it fails to capture the complicated ways in which different, competing preferences are often relevant to respecting *non-ideal* expressions of patient autonomy.

Autonomy is often glossed in terms of an ideal. To be ideally autonomous is to have and form fitting, authentic beliefs and desires, to enjoy the social support necessary for good options, and to have the

CONTACT Ben Schwan ben.schwan@case.edu Department of Bioethics, Case Western Reserve University, 10900 Euclid Ave., Cleveland, OH 44106-4976, USA.

[1] As Earp et al. (2024) emphasize, there are many ways to build a PPP. My comments will apply broadly, so I will stick with this generic construal. But it is worth noting here at the outset that, if feasible, some version of Earp et al.'s "Personalized Patient Preference Predictor" (P4) will likely be best positioned to both realize the benefits and minimize the risks that I discuss in what follows.

[2] In setting these concerns aside I do not mean to suggest that they lack force; some highlight serious challenges for the development and general use of PPPs. Here, however, my focus is on risks and benefits of implementation that must be grappled with even assuming the more global challenges can be addressed.

[3] For a canonical discussion of what is often called the "substituted judgment standard," see Buchanan and Brock (1989).