Enabling the Nonhypothesis-Driven Approach: On Data Minimalization, Bias, and the Integration of Data Science in Medical Research and Practice

Safarlou CW[1], van Smeden, M[2,3], Vermeulen R[1,4], Jongsma KR[1]

1: Department of Global Public Health and Bioethics, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, The Netherlands

2: Department of Epidemiology and Health Economics, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, The Netherlands

3: Department of Data Science and Biostatistics, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, The Netherlands

4: Department of Population Health Sciences, Utrecht University, Utrecht, The Netherlands

Note: This is an Accepted Manuscript of an article published by Taylor & Francis in The American Journal of Bioethics on August 30th 2023, available at: https://doi.org/10.1080/15265161.2023.2237452. Please cite the published version.

Introduction

Cho and Martinez-Martin provide a wide-ranging analysis of what they label 'digital simulacra' – which are in essence data-driven AI-based simulation models such as digital twins or models used for *in silico* trials – that explores many ways in which 'digital simulacra' could affect certain (perceived) ethical and epistemic values (Cho and Martinez-Martin 2022). Their analysis outlines challenges and limitations of incorporating 'digital simulacra' in healthcare, such that potential harms can be mitigated (Cho and Martinez-Martin 2022). While their analysis provides a starting point for understanding the ethical implications of such models, it is our central contention that their analysis misses an identification of *the way in which data are selected* through the 'data-first' or nonhypothesis-driven approach.[1] Instead of drawing on data that are determined to be relevant on the basis of prior hypotheses or theory, a nonhypothesis-driven approach ideally requires all data that one can possibly gather on a target system, in order to subsequently generate a model (of that system) with statistical or AI-based tools that is determined by mathematical and statistical standards. In the following, we argue that, once one recognizes this core element of the nonhypothesis-driven approach as understood in the context of statistical/AI-generated models, it leads to different conclusions than those of Cho and Martinez-Martin on the topics of data

[1] To make the central point of our paper, our preferred term is nonhypothesis-driven approach. But such an approach can go by many other names: data-driven, big data, agnostic, unbiased, untargeted, hypothesis-free, or holistic. When intended to (later on) generate hypotheses and/or identify causal connections, it can also be called: discovery-based, discovery-driven, exploratory, or hypothesis-generating.

minimalization, bias, and the perceived conflict between data science and clinical medicine. Furthermore, we argue that these conclusions each actively *enable*, rather than impede, the ethical and epistemic value of the further development of data-driven statistical/AI-based models as a crucial emerging technology for biomedical research and innovation.

The nonhypothesis-driven approach and the principle of data minimalization

While Cho and Martinez-Martin recognize ethical benefits of 'digital simulacra', such as speeding innovation, lowering the costs of innovation, and minimizing risks to human beings via in silico drug testing, they claim that data minimalization is a looming barrier (Cho and Martinez-Martin 2022). They argue that "regulations that oblige researchers to collect on the minimum necessary protected health information conflict with the analytic needs of digital simulacra developers" because these developers "attempt to collect as much data as possible" (Cho and Martinez-Martin 2022, 11).

However, if the value of a nonhypothesis-driven approach lies in its potential to analyze all possibly gatherable data on a system (as opposed to a set of data determined per hypothesis), then the 'minimum necessary protected health information' is simply equal to the maximum amount of data that one can possibly gather.[2] Thus, there is no conflict between such an approach and the principle of data minimalization. In other words, a nonhypothesis-driven approach does not conflict with a (properly articulated) principle of data minimalization if the standard of 'minimally necessary data' is fulfilled by the maximally high 'analytic needs' of the approach (which signify part of the approach's value).

Furthermore, we have found in the literature that, if a particular formulation of the principle of data minimalization includes the need for a 'particular research question' that one attempts to answer via one's research approach, then there is a possibility for conflict between the two (Safarlou et al. 2023). This possibility depends on whether the granularity of the word 'particular' implies a research question being 'informed by a prior hypothesis that determines which data should be gathered'. In such a case, however, one's particular formulation of the principle of data minimalization oversteps its purpose and subsequently amounts to a burden that stands in the way of the distinctive value that nonhypothesis-driven research can create. A proper principle of data minimalization is not a causeless duty for scientists that falls from the sky; the purpose of such a principle is to stimulate purposeful thinking when choosing which data to gather, to help guard against the consequences of data leaks and to make it easier to share and reuse data (among other things). If a specific formulation of the principle of data minimalization does not allow for nonhypothesis-driven research, then it needs to be reformulated. Also, researchers doing nonhypothesis-driven research should then look into other measures for guarding against (the harmful effects of) data leaks and for making it easy to share and reuse data, such as those recommended by the FAIR principles (Safarlou et al. 2023).

Still, one could object that one's reasons for having a restrictive version of the principle outweigh the risk of using a nonhypothesis-driven approach. However, we believe that such 'weighty' reasons would then present a separate argument against the risk of such an approach because such reasons

[2] Arguably, collecting as much data as possible could lower the total amount of data generated in the long run (and save research costs) if such data can be repurposed for different studies. Relatedly, on the other hand, if bodily materials or exposure samplers such as silicone wristbands are gathered and stored in biobanks, then they can be retested with targeted high-resolution mass spectrometry methods for more specific data (Chung et al. 2021).

do not fit with the concerns that generally give rise to the formulation of a meso-level principle like that of data minimalization (also known as the principle of data minimization).

The bias-reducing potential of the nonhypothesis-driven approach

In effect, Cho and Martinez-Martin deny that the nonhypothesis-driven approach has the potential to reduce or eliminate biases that originate from researchers (Cho and Martinez-Martin 2022). They write that this approach is "depicted" as being able to do so, but argue that "In practice, however, features of digital simulacra have the potential to increase bias, obscuring values and inequities that are embedded in the decisions made throughout the design process" (Cho and Martinez-Martin 2022, 8). They specify that the approach's "purported" potential to reduce human bias lies in its potential to "detect unexpected patterns in data" (Cho and Martinez-Martin 2022, 11). They proceed to deny that the nonhypothesis-driven approach possesses this potential: "simulation models can only detect patterns from the data that they are given, which is determined by the scientist, and is therefore prone to human bias and the limits of human knowledge" (Cho and Martinez-Martin 2022, 11). The authors substantiate this claim by arguing that representations of complex systems are "necessarily highly simplified in digital simulacra" as simplification "requires scientists to make decisions prior to modeling about what features are important" (Cho and Martinez-Martin 2022, 11).

First, we believe that their analysis treats 'human biases' too monolithically: the nonhypothesis-driven approach can reduce some biases and increase other biases. The 'bias' reduced or eliminated by the nonhypothesis-aspect of the nonhypothesis-driven approach is the error of excluding relevant data on the basis of hypotheses (if there actually is data being wrongfully excluded).[3] For example, such an approach can ignore historical decision-making about the safety status of chemical compounds to allow for a more rigorous evaluation of the effects of classes of chemicals on specific perturbed biological pathways (Vermeulen et al. 2020).[4]

Second, due to the fact that the manual creation and/or operation of a big-data model is cognitively complex and often costs an immense amount of time, automatized (AI-based) software can *expand* the limit of human knowledge and ability, and allow for the reduction/elimination of errors that the human mind can make when creating and using highly dynamic and complex models.

Thus, a nonhypothesis-driven approach does not attempt to reduce or remove *all* human error that results from scientific decision making, just particular ones. Nonetheless, the nonhypothesis-driven approach introduces other potential errors and other human decisions that affect scientific modeling. A typical example concerns (the effects of) the decision to use dimension-reduction techniques (Chung et al. 2021). Note, however, that such techniques do not necessitate the type of simplification that Cho and Martinez-Martin describe because "key feature" selection is not hypothesis-driven (Cho and Martinez-Martin 2022, 11). Subsequently, contrary to their claims, such

[3] For other (arguably derivative) biases/errors that can be reduced/eliminated by such an approach, see the discussion of false positives, publication bias, and more, in (Chung et al. 2021). Also, with respect to cognitive bias, such an approach affords the elimination of confirmation bias (and the identification of confounders) to the extent to which confirmation bias affects variable selection. For an overview of discussions of statistical bias, 'normative' bias, and how these two interact with each other, see (Safarlou et al. 2023).

[4] Note that we do not exhaustively discuss all positive (or potentially negative) ethical and epistemic aspects of the nonhypothesis-driven approach in this commentary. For example, we leave aside the tradeoff between coverage and sensitivity/specificity when choosing to gather data via untargeted instead of targeted high-resolution mass spectrometry (Chung et al. 2021).

models retain the potential to detect unexpected patterns in data or generate surprising results (Stingone et al. 2021; Chung et al. 2021).

Or take another example: the extent to which nonhypothesis-driven approaches are actually holistic/agnostic/untargeted and thus unaffected by prior hypotheses. The idea behind such data gathering is that there is a finite amount of data gatherable on a system, that a subset of that amount allows us to fully describe how the system works, and that the more of its superset we gather, the more relevant information statistical or AI-based tools have for performing well and the less room exists for wrongfully omitting relevant variables. Naturally, one would then also capture more irrelevant correlations, and this is where the value of data reduction strategies and the bias-variance tradeoff comes in (Chung et al. 2021). At the same time, one would also run the risk of including colliders and intermediates, and of generating illogical correlations such as death influencing events earlier in life. Such factors need to be taken into account for (subsequent) exploratory research (such as by bringing in prior structure, which again could be a source of bias).

Moreover, researchers attempting to use a nonhypothesis-driven approach might not always (be able to) draw on data that is gathered agnostically. For example, they might draw on data from health registries that have gathered data on the basis of existing theories about health-relevant data.[5] Similarly, López-Cervantes et al. 2021 report that not all cohorts use untargeted high-resolution mass spectrometry to measure exogenous and endogenous compounds, due to perceived risk communication liability. In other words, gathering 'all data available' does not necessitate that one's approach is truly nonhypothesis-driven, as available data could have been gathered through, or more broadly affected by, previous hypothesis-driven investigations. Nonhypothesis-driven research that *does* use such data would then incur what may be coined as a variable pre-selection bias. See also the related discussion of reporting bias for 'the dark matter of the exposome' in Chung et al. 2021.

By explicitly recognizing these benefits (and potential downsides) of the nonhypothesis-driven approach, we are better positioned to explicitly *leverage* its benefits (and account for its downsides) when considering to use or implementing a nonhypothesis-driven approach, and when using the models that it generates.

The nonhypothesis-driven approach and the scientific method

At several points in their paper, Cho and Martinez-Martin juxtapose the 'data-first'/nonhypothesis-driven approach against the "traditional biomedical scientific methods and the logic of clinical reasoning" in a way that anticipates conflict (Cho and Martinez-Martin 2022, 5). For example, they claim that the "worldview" of the former represents a "shift away" from the latter, and they question whether "the epistemic standards of data scientists [should] be allowed to supplant those of traditional biomedical and clinical researchers" (Cho and Martinez-Martin 2022, 5; 12). However, we believe that there is common ground to these two approaches, and that this common ground affords a normative vantage point from which the data science approach can be integrated into clinical research and practice without facing irreconcilable epistemic standards or culture clashes. First of all, let us note that the scientific method is an inductive method that, at its most fundamental level, starts with observing the world in order to understand it. This is a step that 'traditional biomedical scientific methods and the logic of clinical reasoning' share with the 'data-first' approach utilized by 'digital simulacra'. The best way to proceed from this step, however,

[5] Cho and Martinez-Martin make a related point when mentioning "convenience sampling" and "convenience samples" (Cho and Martinez-Martin 2022, 13).

differs depending on one's context and purpose. For example, organizing randomized control trials is not always possible, and observational research can provide helpful discovery- and population-based information for clinical practice. 'Traditional' hypothesis-driven biomedical and clinical researchers have long recognized this fact, and the field of clinical epidemiology has been incorporating observational methods, predictive modeling, and population-to-individual inferences, into clinical medicine for almost a century (Paul 1938; Grobbee and Hoes 2014).

Furthermore, although Cho and Martinez-Martin phrase it as an open question whether there will be "attempts to force a merger" between the nonhypothesis-driven approach and the 'traditional biomedical scientific methods and the logic of clinical reasoning', there already exist bodies of work that smoothly merge the two (Cho and Martinez-Martin 2022, 12). Two examples in this respect concern discussions of explanatory artificial intelligence in medicine and healthcare, and the discovery-based aspects of the exposome approach (Durán, Sand, and Jongsma 2022; Chung et al. 2021).

In conclusion, Cho and Martinez-Martin should not unjustly accuse data scientists of "epistemic hubris" by ascribing to data scientists "the assumption of superiority of one's expertise (or a whole field's way of knowing) over others' or false inferences about the limits of their knowledge" (Cho and Martinez-Martin 2022, 13–14). Instead, we should encourage the epistemic ambitiousness of data scientists through the integration of their innovative approaches via the established and developing methods and standards of clinical epidemiology (Gorlin 2023; Grobbee and Hoes 2014; Chung et al. 2021). Doing so provides an avenue for data scientists from outside clinical medicine to constructively integrate the ethical and epistemic value that they wish to create into medical research and practice, without any *fundamentally* irreconcilable epistemic standards or culture clashes.[6]

Bibliography

Cho, Mildred K, and Nicole Martinez-Martin. 2022. "Epistemic Rights and Responsibilities of Digital Simulacra for Biomedicine." *The American Journal of Bioethics*, 1–12. doi:10.1080/15265161.2022.2146785.

Chung, Ming Kei, Stephen M. Rappaport, Craig E. Wheelock, Vy Kim Nguyen, Thomas P. van der Meer, Gary W. Miller, Roel Vermeulen, and Chirag J. Patel. 2021. "Utilizing a Biology-Driven Approach to Map the Exposome in Health and Disease: An Essential Investment to Drive the next Generation of Environmental Discovery." *Environmental Health Perspectives* 129 (8). doi:10.1289/EHP8327.

Durán, Juan M, Martin Sand, and Karin Jongsma. 2022. "The Ethics and Epistemology of Explanatory AI in Medicine and Healthcare." *Ethics and Information Technology* 24 (4): 42. doi:10.1007/s10676-022-09666-7.

Gorlin, Gena. 2023. "'Intellectual Humility' Is a Copout: Why Builders Need to Raise, Not Lower, Their Epistemic Bar." *Building the Builders*. https://genagorlin.substack.com/p/intellectual-humility-is-a-copout.

Grobbee, Diederick E, and Arno W Hoes. 2014. *Clinical Epidemiology: Principles, Methods, and Applications for Clinical Research*. 2nd ed. Burlington: Jones & Bartlett Learning.

[6] We would like to thank Stefan Gaillard, Salome Kakhaia, and our colleagues from Bioethics & Health Humanities, for their constructive feedback.

López-Cervantes, J P, M Lønnebotn, N O Jogi, L Calciano, I N Kuiper, M G Darby, S C Dharmage, et al. 2021. “The Exposome Approach in Allergies and Lung Diseases: Is It Time to Define a Preconception Exposome?” *International Journal of Environmental Research and Public Health* 18 (23). doi:10.3390/ijerph182312684.

Paul, John R. 1938. “President’s Address Clinical Epidemiology.” *The Journal of Clinical Investigation* 17 (5): 539–41. doi:10.1172/JCI100978.

Safarlou, Caspar W, Karin R Jongsma, Roel Vermeulen, and Annelien L Bredenoord. 2023. “The Ethical Aspects of Exposome Research: A Systematic Review.” *Exposome* 3 (1): osad004. doi:10.1093/exposome/osad004.

Stingone, J A, S Triantafillou, A Larsen, J P Kitt, G M Shaw, and J Marsillach. 2021. “Interdisciplinary Data Science to Advance Environmental Health Research and Improve Birth Outcomes.” *Environmental Research* 197: 111019. doi:10.1016/j.envres.2021.111019.

Vermeulen, Roel, Emma L. Schymanski, Albert László Barabási, and Gary W. Miller. 2020. “The Exposome and Health: Where Chemistry Meets Biology.” *Science* 367 (6476): 392–96. doi:10.1126/science.aay3164.