

Modelling in DOHaD

Challenges and Opportunities in the Era of Big Data

Julie Nihouarn Sigurdardottir and Salma Ayis

25.1 Introduction

Public health data available for research are booming with the expansion of Big Data sources, shifting the landscape of DOHaD research. These new forms of data offer ample opportunities to advance epidemiological modelling within the DOHaD framework. Big Data is often described by the ‘3 Vs’: high volume, high velocity, and wide variety and refers, for example, to the large volumes of Electronic Health Records (EHRs) now stored as many nations move towards the routine electronic recording and centralising of health data. The term Big Data also applies to data derived from wearable devices and phone applications, increasingly affordable technologies that allow for the collection of new kinds of data, in larger volumes, and almost in real time. Such technologies, along with improved data processing speed and advanced computing capacity, grant access to the lifestyle and health information of millions of individuals who can be followed through the lifespan.

However, within heterogeneous and dynamic socio-demographic contexts and a fast-moving technological landscape, these new forms of data raise a plethora of methodological challenges related to accurately characterising population health trajectories and biological mechanisms. In addition, while the current inferential potential of DOHaD research depends on which variables are collected, at what frequency, and at what time points, it is also closely shaped by the theoretical model(s) chosen for a given study: a framework implicating critical and sensitive windows in development shaped the early DOHaD literature, but other models were added such as the *accumulation of risk model*, the *chain of risk model*, and a hybrid of those [1]. These frameworks shape study designs, data collection practices, and the interpretation of results and set the scene for how Big Data is likely to be taken up in the field.

In this chapter, we provide an overview of DOHaD modelling methods and consider the emerging place of Big Data in investigating multidimensional research questions in the field. To do so, we discuss various methodological aspects of modelling, such as operationalisation, sampling, population representation, ethics, and the accuracy of tools used to acquire and analyse data. We also discuss the current landscape of artificial intelligence-derived methods, judging their utility against the validity of findings, and their potential when compared to ‘traditional’ empirical data sources and analytical approaches.

25.2 Current DOHaD Modelling and Methodological Challenges

A myriad of methodological challenges are present in DOHaD research even prior to Big Data, in particular, the issues of validity across time and space, characterising causal

links, and identifying sources of bias. Physiological processes are difficult to model because they are integrated into non-static systems. These refer to the less quantifiable and less predictable behavioural, lifestyle, environmental, and socio-economic systems interacting with biology. The moderating or mediating effects of culture, health inequalities, medical systems, and health policy on health outcomes must be clarified to assess the generalisability of any given model. Even in gold-standard birth cohort research designs, epidemiological models need to account for the fact that, for example, through societal restructuring and climate change, the properties of exposures can change over time and across generations [2]. As a result, it is difficult to produce predictive DOHaD models and interventions that remain valid and useful across time for a given population. Alternative designs such as the observational study design in humans cannot, however, capture the complexity of all important causal links.

A further challenge for aetiological and epidemiological models of DOHaD is how to define the sources of individual differences in health outcomes with a robust degree of certainty. Some examples include disentangling antenatal and postnatal exposure and their interactions [3]; accounting for sex and gender-based differences in biology and behaviour; inter-organ variation in adaptability to maternal ill health (e.g. the placenta response to stress) [4, 5]; and evaluating disparities in outcomes across diverse groups given that the bulk of data is from a small number of middle- to high-income and white-dominated contexts [6, 7].

Prediction models are also prone to confounding and collider bias [8] (a variable in a causal pathway that is a shared effect of more than one cause). For DOHaD research, the primary exposures studied in the developmental pathway are nutrition, parental physiological and psychological health, the environment and toxicants, and social and demographic determinants [7]. Intuitively it is easy to assume that many of these exposures can co-occur and may moderate one another. While confounding can often be resolved by taking these variables into account, residual confounding remains a risk when those influencing factors are unknown or unmeasurable. In the case of collider bias, this can also lead to counter-intuitive conclusions. Such counter-intuitive conclusions are exemplified by the 'birthweight paradox', where babies born with low birthweight (LBW) to smoking mothers (exposure) appear to have a lower risk of neonatal mortality (outcome) compared to those born with LBW to non-smoking mothers [9].

Here, tools such as directed acyclic graphs (DAGs) that portray causal relationships graphically can help a researcher explore a model's functional assumptions and conceptualise any mechanisms of causality. DAGs help recognise mediators, moderators, confounders, and colliders [10] and have helped to illustrate the collider role of LBW in the above paradox [9]; that is, when maternal smoking is absent, other unobserved causes (malnutrition and congenital defects) can lead to LBW and more severe health problems and thus higher infant mortality. Therefore, while the above research question appeared 'simple' initially and involves few measurable exposures (smoking or not) and outcomes (LBW and mortality), failing to incorporate inter-correlations between variables and confounding effects, conceptualised by theory and DAGs, is unlikely to provide reliable causal inference.

Taking another example, understanding the association between maternal stress and lower infant cognitive outcome [11] warrants pertinent exploration into the relative contributions of other exposures concomitant to maternal stress, such as under- or overnutrition, infections and toxicants, and their interrelationships. Here, modelling methods should integrate observed variables, latent variables (not directly observed but

derived from questionnaires or other observed variables, i.e. stress), and their measurement errors, alongside time indicators. Moreover, the mediating mechanisms proposed in the literature, such as epigenetic modulation, the microbiome, metabolism, and (offspring) endogenous immunity, would also have to be incorporated. In practice, a single model that simultaneously incorporates multiple predictive pathways and associations will more accurately capture the causal relationships of interest between the exposure and outcome of interest [12, 13]. Of translational value, such epidemiological modelling would eventually result in developing better targeted interventions.

25.2.1 Data: What Are We Collecting, What Are We Measuring?

High-quality data that are fit for purpose and meet the criteria of accuracy, validity, completeness, and consistency are a cornerstone of empirical science. Data quality may be affected at the stages of data collection, cleaning, or the numerical transformation that is often used to meet required assumptions such as the normal distribution in statistics. Measurement errors, whether systematic or random, are present in all observational studies. While these errors impact the validity and reliability of data and introduce biases, they are rarely acknowledged or accounted for in the epidemiological literature. Makin and de Xivry address common statistical mistakes [14], and Wagenmakers et al. [15] present guidelines on how to report statistical analyses transparently based on four scientific norms of ‘communalism, universalism, disinterestedness and organised skepticism’.

Since we allude above to the notion of variable choice and availability, next, we discuss the importance of clear terminology and data quality in DOHaD research methods. How we define and measure exposures and outcomes impacts inferences, findings, and subsequent interventions and policies. One example of this is the work of researchers who rely on clinically defined groupings based on dichotomisation, such as diabetes diagnosis, or the classification of body mass index (BMI) as obese/overweight/normal-weight/underweight. One obvious risk of using a strict classification of body morphology by BMI alone is undermining the field’s knowledge about fat distribution being a strong determinant for metabolism and cardiovascular health, especially fat within the abdomen (visceral adiposity). Without other markers to corroborate metabolic health risks (blood pressure, cholesterol, visceral fat mass, etc.), some individuals with normal weight, categorised as ‘controls’, may be metabolically unhealthy and ‘at-risk’ of physiological phenotypes. In fact, this group represents 35 per cent of normal-weight individuals [16]. This problem extends to gestational diabetes screening in pregnancy, which is provided to women meeting the BMI > 30 kg/m² criteria in the UK, while those under 30 are assumed to be void of any hyperglycaemia risks during pregnancy. The absence of evidence, however, is not evidence of absence. The consequence of such hidden (latent) subgroups of individuals within a ‘control’ or ‘normal-weight’ category is the introduction of bias into the statistical analyses that epidemiological models rely on for inference, thus leading to inaccurate conclusions.

Similarly, in psychology, the diagnosis of autism as present/absent is common, although autism spectrum disorder (ASD) is typically conceptualised by experts as a continuum (see also Azevedo et al. in this volume). Such a binary diagnosis of ASD ignores potential distinct mechanisms of importance for the DOHaD of autism subtypes, which could possibly relate to the timing of any ‘disruption’ in brain development and could be informative for mechanistic studies and prognosis [17].

Overall, what we suggest above is that DOHaD researchers must be conscious of relying on clinical data alone such as those retrieved from EHRs without clarifying sources of biases, the caveats of present/absent dichotomised diagnoses [18], and the local clinical guidelines from which they derive. We suggest, however, that some classification approaches are available to limit some of these caveats such as collating multiple variables and produce profiles based on similarities of exposure and/or outcomes at one time point (e.g. latent class modelling) or many time points over time to uncover trajectories (latent class growth analysis and piecewise modelling). This could mean, for example, retrieving glucose measures sampled throughout pregnancy and establishing the likely glycaemic status rather than relying only on a single GDM diagnosis. These approaches also help identify profiles of individual responses to interventions and can therefore improve tailored treatment allocation.

Additionally, missing data, either by design (e.g. unmeasured exposures/outcomes) or attrition, negatively impact data quality and challenge causal inference. This can be addressed by powerful analytical tools that recognise data complexity and the impact of missing data [19, 20]. Several methods are available to address missing data, including maximum likelihood, multiple imputation, and Bayesian methods. Complete case analysis leads to loss of data and statistical power but is widely used, while other complex but more justifiable methods are not often attempted [21]. Assumptions about the properties of missing data, whether these are missing completely at random or not, must be made. With the emergence of Big Data and EHRs that tend to have a high prevalence of missing information, appropriate techniques that deal with missing data need to be carefully applied.

25.3 Big Data: Challenges and Opportunities

As referenced above, Big Data refers to data large in volume, collected at high velocity, and comes in a variety of sources, formats, and dimensions, such as from birth cohort and longitudinal studies, medical records, or wearable/phone devices. Birth cohort studies, such as the Avon Longitudinal Study of Parents and Children study, have supported the DOHaD hypothesis by in-depth prospective sampling and large multidimensional data collection from human participants. While integral to the DOHaD evidence base, standard cohort studies are costly and may be of limited size. EHRs, a source of Big Data, can be obtained from centralised systems, while large omics data sets (genomics, transcriptomics, metabolomics, etc.) are often sourced from biobanks and can be added to increasingly available personal and external 'exposome' data (e.g. lifestyle and environmental). Among the benefits of EHRs is their level of comprehensiveness, and so with larger samples, this also improves the statistical power required to provide accurate estimates of effect size. The availability of such data means that if taken up in DOHaD research, the scope for such studies would no longer be limited by small sample sizes due to funding and/or the restrictive protocols of conventional longitudinal birth cohorts [22]. In practice already, linkage study designs join primary- and secondary-care databases or merge multiple EHR databases and registries, potentially offering new insight into disease pathways. For example, the UK-based CALIBER study drew on EHR sources to investigate the cumulative incidence and period prevalence of diseases over the lifecourse. Results were presented in the form of a chronological map of 308 physical and mental health conditions from four million individuals, from infants

to the elderly [23]. The role of universal medical coverage and centralised digital health records, such as the English National Health Service, in enabling such exploration in this particular study cannot be underestimated.

More recently, data acquired from real-time biosensors measuring pollution exposure, blood glucose levels, or heart rate from wearable devices have become available. Data involving behaviour and social networks can also be retrieved from open social media platforms at high speed. The past three years, especially during the COVID-19 pandemic, have seen a surge of software developments intended to meet the needs of monitoring health markers and well-being remotely. The uptake of telemedicine was enabled, for example, by digital platforms used by clinicians to manage antenatal hyperglycaemia [24] and the self-report of glucose levels by pregnant women on their phones [25]. These tools could be employed in future DOHaD studies. However, key ethical issues related to privacy, rights, and moral code of conduct when retrieving these data require careful considerations in this changing research landscape.

25.3.1 Limitations of Current Big Data Sources and Applications

One first potential caveat of relying on Big Data sources is that DOHaD researchers wanting to use Big Data may be obliged to formulate research questions based on data availability or including data not necessarily designed primarily for DOHaD research. These researchers will have less control over data quality because of the larger distance from data collection, that is the inputting user (clinician for EHR / hospitals, or user of a phone app), and from the decisions made in defining and measuring the variables in these data sets. Overall, sources of error need to be considered when evidence from Big Data is evaluated. Without researchers' involvement in data collection, it may be impossible to subsequently correct or even identify these errors.

The task of comparing and validating DOHaD models across populations may be further hindered by the heterogeneity in data architectures across national and international sources. Before the term Big Data surfaced, omics-derived data alone (e.g. genomics, transcriptomics, and proteomics) already inferred the outputs of millions of data points [3]. Formulating a cascading model of these omics layers, which follow biologically downstream from one another, is both necessary and extremely complex. Further, linking biologically derived material to clinical data of different formats, and based on a variety of measurements, including imaging, questionnaires, and diagnoses, requires technologies that facilitate multidimensional integration. Thereafter, powerful methods that support analysis are necessary.

Larger sample sizes improve the power to detect effects, and clearly the whole DOHaD framework requires both large samples and a comprehensive set of exposures and events to be modelled. However, the primary issue is that complex models are more difficult to explain and thus could complicate their practical translation into actionable policies. Users and clinicians equally need to be versed in their use and interpretation.

The use of Big Data also raises issues of data security and representation, particularly data obtained outside conventional academic institutions, in contexts where systems and resources are not fit for this purpose, such as in low- and middle-income countries (LMICs). Users of both healthcare services and digital platforms, such as social media, may represent distinct groups with possibly little overlap in demography, risks, and healthcare needs [22]. It is plausible to assume that LMICs are unlikely to have population-wide health records or

access to digital data collection and remote health monitoring from which DOHaD modelling could be done. Here, validation and reproducibility of DOHaD models are less feasible and highlight the lack of representation through their exclusion.

25.4 Artificial Intelligence: Challenges and Opportunities

When Big Data is considered, the word AI is not far behind. Pattern recognition, similarity profiling, and predictions are tasks for which AI methods such as machine learning (ML) have been developed, and these have potential applications to DOHaD research. It should be noted that ML and conventional statistical methods may be seen as a continuum since the algorithms behind ML, including linear and logistic regression, and several dimensionality reduction techniques have existed for decades. (For a contrast between ML and conventional statistics, see [26].) However, the real advantage of AI is that it supports the analysis of large data volumes alongside multidimensionality (i.e. where the number of variables is larger than the subjects).

AI is already being tested and implemented in the clinical domain, including to improve the efficiency of hospital administration. AI is also being used to predict medication side effects and patient outcomes from radiological imaging and thereby promote patient-tailored medicine and interventions. In DOHaD, ML approaches would subserve exploratory designs to identify biological pathways, which appear more frequently in the mosaic of data, in the form of associations, including from DNA sequences and omics data, and those obtained from EHRs [26]. Certain applications require user input from which the ML 'learns' to classify new data from sets of rules from previous data (supervised learning) or is completely unsupervised in detecting patterns. This is similar to the latent modelling techniques mentioned earlier that derive from the 'classical' statistics and the structural equation modelling framework. Other subfields of AI associated with ML include deep learning, rooted in multi-layered neural networks, which also allow computers to identify relations between concepts/features and characterise these associations from complex to simpler concepts [27]. ML methods to date could also assist in the processing of single modalities or data collection methods, such as magnetic resonance imaging of the brain, heart rate variability in the fetus, and DNA methylation patterns in disease [28], all of which are relevant for DOHaD research.

Despite the potential of ML for DOHaD described, very few ML studies have so far transitioned from single data type and scale to 'fusing' several dimensions, or towards the integration of additional outcome measures retrieved from EHRs, medical imaging, and biospecimens. Data harmonising and deployment of ML is an ongoing endeavour, but some attempts have been made in relation to cardiovascular medicine (reviewed by [29]). The issue to date is that these algorithms exploit two modalities at most (e.g. radiological imaging and free text from clinical reports). (For an in-depth review of the current landscape of AI for multi-modal integration, see [30]).

In our previous section, we discussed the necessity to characterise accurately DOHaD prediction models. It is at present the case that the several competing ML approaches and the rapidly evolving demands of AI have yet to produce a consensus regarding how to develop or validate a prediction model relying on these novel tools. For example, in predicting Type 2 diabetes and cardiovascular disease, Dalakleidi et al. reported the best performance to have been achieved by groups of artificial neural networks. However, the so-called decision trees, random forest algorithm, and support vector machines were said

to provide the best accuracy measures by Zheng et al. [31]. Furthermore, AI methods do not necessarily outperform conventional statistical regression applications and are not free of methodological biases [32]. Additionally, scholars warn that ML studies are often computationally demanding on resources (e.g. support vector machines, logistic regression, random forests, gradient-boosted machines, and neural networks) [33].

25.4.1 Issues of Interpretation and Reporting

Concerns are often raised about the scope, complexity, transparency, reproducibility across different scientific teams and different populations, and the interpretability of prediction models. While AI operates from a 'black box' within deep neural networks and unsupervised learning, the biological plausibility and meanings of the output are generated by researchers. Given that so far only a few published prediction models have found utility in clinical practice, the utility of AI when compared to conventional methods remains an open question. It is unclear whether AI can address the questions of causality most pertinent in DOHaD, when DOHaD draws from interpretability and theory and is moving towards the integration of social science and ethnography. (See Richardson, in this volume for a discussion of how DOHaD is characterised by 'cryptic causality'.)

Guidelines regarding AI are developing rapidly. For example, following a quality assessment of the conduct and reporting of multi-variable prediction models, a 22-item checklist (TRIPOD) was developed [34]. The risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on AI has also emerged to ensure that users have key information about the design, conduct, and analysis, alongside a robust standardised tool for bias evaluation that would allow a fair judgement on the utility of these models [31, 35, 36]. Guidelines should be consulted by authors, reviewers, and editors, to ensure reproducibility, reliability, and validity, and hence safe implementation. Again, even prior to applying AI-derived analytics, the uncertainty of measurements in Big Data and its sources of error must be accounted for and possibly identified systematically, and the data quality validated.

25.6 Ethical Questions about Big Data and AI

The rapid expansion of Big Data and AI raises a range of ethical concerns. First, the question of who audits and protects data including EHRs (which can also be used as testing data by commercial parties) is central to ensuring ethical research in the DOHaD field and is currently insufficiently addressed [22, 37]. The lack of standardisation of data protection laws between countries adds to this issue.

Additionally, the most powerful AI pipelines are deployed from within the very few corporations with the computational and financial resources. The commercialisation of both healthcare software or AI tools and their findings within the private sector is another challenge that research institutions must navigate if the potential of such data is to be realised. Such a feat would require more transparency and possibly a move to open sources of data. (For an example of Google's DeepMind approach to open data, see [38].)

Nevertheless, open and public data collection is also likely to introduce other ethical issues that need to be carefully considered [39]. Big Data collection and usage may move the position of the individual (the unique data provider) from the one fulfilling the 'social vision' of the healthcare system and science into the 'economic vision' of the commercial enterprise [22]. Participants recruited through academic institutions consciously engage

with the scientific community with consent and pledge their time voluntarily. This contrasts with the passive and often unwitting involvement in Big Data collection via the data generated by medical records and phone devices. The use of such data without consent raises concerns regarding data ownership, privacy, and the circulation of profits. Data protection in secondary research by academic institutions using public data is enforced by the institutions themselves through university and institutional ethics boards, but enforcing consent and data protection may be less clear when third-party commercial and private bodies are concerned.

Rarely mentioned in the discussion of AI-derived methods are the risks introduced or heightened to certain populations because of their implementation. Experts such as Professor Kate Crawford at the AI Now Institute are re-evaluating the societal burden of AI. She states AI is not 'artificial' since it requires the same earthly resources and labour to mine the power and hardware sustaining it. Consequently, this is also becoming a source of disparity and power imbalance on the ground, within and between populations who compete to mine these resources [40]. Such ramifications would be the real irony for an AI integration into DOHaD research and the long-term agenda of the scientific community.

25.7 Conclusion

There is a strong anticipation that in the future, DOHaD researchers will benefit from innovative methodological designs. This could build on the best of current biostatistical methods and soon include AI technologies where the multidimensionality of data sources and a longitudinal format can be integrated, and the outputs shown to be interpretable. Current and novel 'mega' projects may push this progress forward. An example is the protocol implemented in the EarlyCause project [41] that will explore the causal mechanisms between early-life adversity (antenatal and postnatal) and future psycho-cardio-metabolic multi-morbidity. It will involve the participation of 14 European institutions and include three complementary and sequential phases that integrate longitudinal population data sets (e.g. ALSPAC and UK BIOBANK), animal studies, and cellular models with analytical tools from structural equation modelling and machine learning. It also aims to offer a web-based platform for data access and information on research standards and best practices to support future study designs and exploration. Such a mix of granular data collection and Big Data sources in open access, along with AI and conventional statistical approaches, holds great potential for DOHaD research.

The DOHaD research community may look to other fields and consider how to train their own data and solution architects in the newest technologies and Big Data usage. Interdisciplinary teamwork will be crucial in ensuring both robust management and use of data as well as anticipating ethical and governance issues [42]. It is crucial to assess whether certain limitations are inevitable or can be remedied to create the necessary, transparent, and reliable evidence base. Collaborations in data collection could be expanded more frequently to 'crowdsourcing' in data analysis and interpretation. Of course, teamwork does not come without caveats when studying complex and dynamic modelling, such as leading to further heterogeneity in findings and conclusions [43]. Nevertheless, we reiterate that attention to operationalisation of exposures and outcomes, reducing bias in data collection and analysis, and the necessity for interpretability should be at the forefront of the DOHaD agenda in the era of Big Data and AI.

References

1. Johnson W, Kuh D, Hardy R. A life course perspective on body size and cardio-metabolic health. In: Burton-Jeangros C, Cullati S, Sacker A, Blane D, editors. *A Life Course Perspective on Health Trajectories and Transitions*. Cham: Springer; 2016.
2. Zolitschka KA, Razum O, Breckenkamp J, Sauzet O. Social mechanisms in epidemiological publications on small-area health inequalities – A scoping review [Internet]. *Front Public Health*. 2019; 7. Available from: <http://dx.doi.org/10.3389/fpubh.2019.00393>
3. Lapehn S, Paquette AG. The placental epigenome as a molecular link between prenatal exposures and fetal health outcomes through the DOHaD hypothesis. *Curr Environ Health Rep* [Internet]. 2022 29 Apr:490–50. Available from: <http://dx.doi.org/10.1007/s40572-022-00354-8>
4. Bowman CE, Arany Z, Wolfgang MJ. Regulation of maternal-fetal metabolic communication. *Cell Mol Life Sci*. 2021 Feb;78(4):1455–86.
5. Huynh J, Dawson D, Roberts D, Bentley-Lewis R. A systematic review of placental pathology in maternal diabetes mellitus. *Placenta*. 2015 Feb;36(2):101–14.
6. Brandlistuen RE, Ystrom E, Nulman I, Koren G, Nordeng H. Prenatal paracetamol exposure and child neurodevelopment: A sibling-controlled cohort study. *Int J Epidemiol*. 2013 Dec;42(6):1702–13.
7. Abdul-Hussein A, Kareem A, Tewari S, Bergeron J, Briollais L, Challis JRG, et al. Early life risk and resiliency factors and their influences on developmental outcomes and disease pathways: A rapid evidence review of systematic reviews and meta-analyses. *J Dev Orig Health Dis*. 2021 Jun;12(3):357–72.
8. Berkson J. Limitations of the application of fourfold table analysis to hospital data. *Int J Epidemiol*. 2014 Apr;43(2):511–15.
9. Whitcomb BW, Schisterman EF, Perkins NJ, Platt RW. Quantification of collider-stratification bias and the birthweight paradox. *Paediatr Perinat Epidemiol*. 2009 Sep;23(5):394–402.
10. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research [Internet]. *Epidemiology*. 1999; 10:37–48. Available from: <http://dx.doi.org/10.1097/00001648-199901000-00008>
11. Wu Y, Espinosa KM, Barnett SD, Kapse A, Quistorff JL, Lopez C, et al. Association of elevated maternal psychological distress, altered fetal brain, and offspring cognitive and social-emotional outcomes at 18 months. *JAMA Netw Open*. 2022 Apr 1;5(4):e229244.
12. Monk C, Fernández CR. Neuroscience advances and the developmental origins of health and disease research. *JAMA Netw Open*. 2022 1 Apr;5(4):e229251.
13. Sigurdardottir JN, White S, Flynn A, Singh C, Briley A, Rutherford M, et al. Longitudinal phenotyping of maternal antenatal depression in obese pregnant women supports multiple-hit hypothesis for fetal brain development, a secondary analysis of the UPBEAT study. *EClinicalMed*. 2022 Aug;50:101512.
14. Makin TR, de Xivry JJO. Ten common statistical mistakes to watch out for when writing or reviewing a manuscript [Internet]. *eLife*. 2019; 8:e48175. Available from: <http://dx.doi.org/10.7554/eLife.48175>
15. Wagenmakers EJ, Sarafoglou A, Aarts S, Albers C, Algermissen J, Bahnik Š, et al. Seven steps toward more transparency in statistical practice. *Nat Hum Behav*. 2021 Nov;5(11):1473–80.
16. Fan L, Qiu J, Zhao Y, Yin T, Li X, Wang Q, et al. The association between body composition and metabolically unhealthy profile of adults with normal weight in Northwest China. *PLoS One*. 2021 25 Mar;16(3):e0248782.
17. Lai MC, Kassee C, Besney R, Bonato S, Hull L, Mandy W, et al. Prevalence of co-occurring mental health diagnoses in the autism population: A systematic review and meta-analysis. *Lancet Psychiatry*. 2019 Oct;6(10):819–29.

18. Altman DG, Bland JM. Statistics notes: Absence of evidence is not evidence of absence [Internet]. *BMJ*. 1995; 311:485–485. Available from: <http://dx.doi.org/10.1136/bmj.311.7003.485>
19. Stavola BLD, De Stavola BL, Nitsch D, dos Santos Silva I, McCormack V, Hardy R, et al. Statistical issues in life course epidemiology [Internet]. *Am J Epidemiol*. 2006; 163: 84–96. Available from: <http://dx.doi.org/10.1093/aje/kwj003>
20. Johnson W. Analytical strategies in human growth research [Internet]. *Am J Hum Biol*. 2015; 27: 69–83. Available from: <http://dx.doi.org/10.1002/ajhb.22589>
21. Bell ML, Fiero M, Horton NJ, Hsu CH. Handling missing data in RCTs; a review of the top medical journals. *BMC Med Res Methodol*. 2014 19 Nov;14:118.
22. Delpierre C, Kelly-Irving M. Big data and the study of social inequalities in health: Expectations and issues [Internet]. *Front Public Health*. 2018; 6:1–5. Available from: <http://dx.doi.org/10.3389/fpubh.2018.00312>
23. Kuan V, Denaxas S, Gonzalez-Izquierdo A, Direk K, Bhatti O, Husain S, et al. A chronological map of 308 physical and mental health conditions from 4 million individuals in the English National Health Service. *Lancet Digit Health*. 2019 Jun;1 (2):e63–77.
24. Jardine J, Relf S, Magee LA, von Dadelszen P, Morris E, Ross-Davie M, et al. Maternity services in the UK during the coronavirus disease 2019 pandemic: A national survey of modifications to standard care. *BJOG*. 2021 Apr;128 (5):880–9.
25. Mackillop L, Hirst JE, Bartlett KJ, Birks JS, Clifton L, Farmer AJ, et al. Comparing the efficacy of a mobile phone-based blood glucose management system with standard clinic care in women with gestational diabetes: Randomized Controlled Trial. *JMIR Mhealth Uhealth*. 2018 20 Mar;6(3):e71.
26. Bzdok D, Altman N, Krzywinski M. Statistics versus machine learning [Internet]. *Nat Methods*. 2018; 15: 233–34. Available from: <http://dx.doi.org/10.1038/nmeth.4642>
27. Richards BA, Lillicrap TP, Beaudoin P, Bengio Y, Bogacz R, Christensen A, et al. A deep learning framework for neuroscience. *Nat Neurosci*. 2019 Nov;22(11):1761–70.
28. Rauschert S, Raubenheimer K, Melton PE, Huang RC. Machine learning and clinical epigenetics: A review of challenges for diagnosis and classification. *Clin Epigenetics*. 2020 3 Apr;12(1):1–11.
29. Amal S, Safarnejad L, Omiye JA, Ghanzouri I, Cabot JH, Ross EG. Use of multi-modal data and machine learning to improve cardiovascular disease care. *Front Cardiovasc Med* [Internet]. 2022 [cited 19 Oct 2022];9:1–11. Available from: <http://dx.doi.org/10.3389/fcvm.2022.840262>
30. Acosta JN, Falcone GJ, Rajpurkar P, Topol EJ. Multimodal biomedical AI. *Nat Med*. 2022 15 Sep;28(9):1773–84.
31. Zheng T, Xie W, Xu L, He X, Zhang Y, You M, et al. A machine learning-based framework to identify type 2 diabetes through electronic health records. *Int J Med Inform*. 2017 Jan;97:120–7.
32. Collins GS, Mallett S, Omar O, Yu LM. Developing risk prediction models for type 2 diabetes: A systematic review of methodology and reporting. *BMC Med*. 2011 8 Sep;9:103.
33. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J Clin Epidemiol*. 2019 Jun;110:12–22.
34. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD Statement [Internet]. *Eur Urol*. 2015; 67:1142–51. Available from: <http://dx.doi.org/10.1016/j.eururo.2014.11.025>
35. Sounderajah V, Ashrafian H, Golub RM, Shetty S, De Fauw J, Hooft L, et al.

- Developing a reporting guideline for artificial intelligence-centred diagnostic test accuracy studies: the STARD-AI protocol. *BMJ Open*. 2021 28 Jun;11(6):e047709.
36. Collins GS, Dhiman P, Andaur Navarro CL, Ma J, Hooft L, Reitsma JB, et al. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. *BMJ Open*. 2021 9 Jul;11(7):e048008.
 37. Ruckenstein M, Schüll ND. The datafication of health [Internet]. *Ann Rev Anthropol*. 2017; 46:261–78. Available from: <http://dx.doi.org/10.1146/annurev-anthro-102116-041244>
 38. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*. 2021 Aug;596(7873):583–9.
 39. de Laat PB. Algorithmic decision-making based on machine learning from Big Data: Can transparency restore accountability? *Philos Technol*. 2018;31(4):525–41.
 40. Campolo A, Crawford K. Enchanted determinism: Power without responsibility in artificial intelligence [Internet]. *Engaging Sci Technol Soc*. 2020; 6:1–19. Available from: <http://dx.doi.org/10.17351/ests2020.277>
 41. Mariani N, Borsini A, Cecil CAM, Felix JF, Sebert S, Cattaneo A, et al. Identifying causative mechanisms linking early-life stress to psycho-cardio-metabolic multi-morbidity: The EarlyCause project. *PLoS One*. 2021 21 Jan;16(1):e0245475.
 42. O'Doherty KC, Shabani M, Dove ES, Bentzen HB, Borry P, Burgess MM, et al. Toward better governance of human genomic data. *Nat Genet*. 2021 Jan;53(1):2–8.
 43. Silberzahn R, Uhlmann EL. Crowdsourced research: Many hands make tight work. *Nature*. 2015 8 Oct;526(7572):189–91.