Old biases in new data:

Inclusive pre-processing to create disability representation in synthetic datasets

AT2030 | Inquire Cluster www.at2030.com

Authors: Jamie Danemayer, Victoria Austin

1 Table of contents

Contents

2 Overview

Population-based data disaggregated by disability are essential for informed policymaking, especially for disability-inclusive development and the realisation of the rights of persons with disabilities. Both areas rely on accurate evidence and its efficient use, especially in the current global context of resource constriction. Disability inclusive data, and inclusive disaggregated data sets more widely can enable assessment of whether people with disabilities participate in society on equal terms with those without disabilities, as well as supporting difficult decision making about how and what to prioritise in a resource poor context.

The importance of inclusive data sets is growing in the context of population ageing, as governments must plan for increasing levels of support need. When individuals are not represented in the data used to design policy, they are less likely to benefit from it, perpetuating exclusionary practices in a phenomenon described in clinical health research as data poverty. Yet effectively representing disability in population data, and then interpreting what resources are needed when and by whom, is not straightforward.

Disability is a protected characteristic, similar to gender or sexual orientation, meaning that its collection and use pose risks around privacy, discrimination, and misuse. Combined with challenges with non-inclusive and inaccessible data collection, this contributes to the systematic underrepresentation of disabled people in many population datasets. Data poverty and evidence exclusion can then result in poor, ill-informed decision-making which can lead to disadvantaged outcomes for disabled people. However, rapid developments in data science mean new statistical methods can be drawn upon to improve existing datasets, including techniques to fill in missing information and mitigate bias. These methods warrant investigation, as they potentially offer new hope for evidence-based, disability-inclusive policymaking in data-sparse settings.

One advancement in data science is using artificially-generated datasets (known as synthetic data) to fill in the gaps where complete, representative data are not available.

Synthetic data creation involves generating a new set of datapoints with the same properties as the original dataset (i.e. preserving key statistical relationships between variable), while removing identifying characteristics. This process can also include steps to augment the data, for example by correcting historical biases and underrepresentation. Further, because individuals cannot be identified, synthetic datasets can be made more freely available to researchers. For these reasons, synthetic data are increasingly used to study relationships within data and to inform decision-making where access to sensitive data is restricted, or complete data is unavailable.

However, synthetic data are inherently dependent on the quality and representativeness of the original, real dataset upon which they are based. Without careful preprocessing, synthetic data risk reproducing and amplifying the representation gaps of the real dataset, into any evidence and policy they have informed.

Although the use of synthetic data in disability policy has not been widely documented, its application in adjacent areas of public health, such as clinical guidance and health system planning, suggests that its use will continue to expand. As population ageing also drives governments to revise public health policies, the appetite for population-based evidence (that factors in demographic- and age- structure) is rapidly increasing. It is highly likely that without clear guidance on fairness and bias mitigation, limited datasets will be used to generate synthetic data for policy-level decision-making in the near future. Subsequently, opportunities to improve disability representation in data—by protecting individuals' privacy and expanding access to quality data on which to base policy decision-making—will be missed. Researchers in this space can seek to test, develop and apply new and good practice for assessing and improving disability representation (among other protected characteristics) using new statistical methods. These inclusive efforts are most vital and impactful at the preprocessing stage, improving the quality of evidence and advancing towards a global coherence on the fair use of synthetic data.

To develop the idea of inclusive preprocessing, this article brings together three intersecting areas: the role and limitations of administrative disability data; the opportunities and risks associated with synthetic data for disability inclusion; and the future methodological work needed to ensure that new data practices do not reproduce old biases and is in fact useful for disability-inclusive policymaking, highlighting our research agenda. While appropriate methods to do this will vary by context and use case, there is a clear need for developed overarching guidance on the use of synthetic data and AI more broadly in structural health research, to enable analyses to delineate and address inequities.

3 Disability data

3.1What is meant by disability data?

In this article, 'disability data' refers to administrative survey data that can be disaggregated by binary disability status (e.g., disability registration), indicators of disability (e.g., functional difficulty), or the presence of conditions that qualify individuals for services (e.g., a clinically diagnosed health condition). These data are not limited to public health or social service datasets, but may span sectors such as education or employment. provided they are available at the general population level (i.e. also including non-disabled respondents, and representative of the population from which they were drawn). Population-based data that can be disaggregated by disability indicators are essential for informed and inclusive

policymaking,[1–3] monitoring societal participation inequities by disability status,[2] and the realization of the rights of persons with disabilities. Specifically, Sustainable Development Goal target 17.18 commits governments and other actors to increase the availability of high-quality, timely, and reliable data disaggregated by disability.[4]

3.2 Representation in data, representation in policy.

Disability data are needed to ensure the disabled population are not disenfranchised by data poverty. In clinical research, data poverty describes the effect of being 'locked out' of the benefits from emerging clinical guidance and innovations, due to poor representation in the evidence basis from which these advancements were developed.[5] A downstream effect is described by Saunders et al, who argue that the lack of public health policies supporting hearing healthcare, for example, can be attributed to a lack of data and evidence on the interventions that support hearing aid adoption at later-life stages.[1] Data poverty inhibits data-driven advancements in both technology and policy from being more widely realised, and can even lead to them causing harm.[5] Aitken et al examined methods for constructing disability indicators within linked administrative datasets, finding that diagnostic information is heavily relied upon due to the scarcity of disability-relevant variables, and emphasising the importance of rigorous validation to understand the strengths and limitations of derived indicators.[6] There is a pressing need to avoid creating a data divide that exacerbates existing inequalities. A review of disability in AI research specifically emphasises the urgent need for further research to ensure that AI benefits all members of society equitably and that future AI systems are designed with inclusivity and accessibility as core principles.[7]

3.3 Data collection and use limitations

Disability data are not straightforward to collect, reuse, or interpret. Definitions of 'disability' vary within societies and languages, and will also be shaped by the needs of survey administrators and funders. Beyond these, other factors will differentially influence the data collection outcomes. For instance, self-report of disability indicators is influenced by social desirability bias and privacy concerns, which all vary by normative cultural factors (i.e. stigma and perception of disabled people in a society) as well as individual-level factors and experiences. For these reasons, disability is often underrepresented in general population data, with low-income countries in particular more often reporting low national disability prevalences[8] compared to global average estimates of 15-20%.[9] Some of these challenges are common barriers to interpreting other protected characteristic data,[10] as well as data which secondarily contains information on protected characteristics, all of which are considered sensitive.[11] However, disability may be further underrepresented because of

accessibility barriers that limit participation in surveys and administrative data collection or assumptions about the value of participation. These challenges motivate disability-specific surveys that use targeted sampling, rather than general population-based methods, which can make it a challenge to compare disabled population-specific outcomes to general population outcomes. Methodological choices also have downstream effects that complicate the representation of disability in evidence. For example, disability estimates in many countries are solely based on the prevalence of disability benefits registration,[12] yet registration is often a prohibitively difficult process that excludes many individuals. This reliance can lead to incomplete or inconsistent identification of people with disabilities in data, particularly where registration is not accessible, where individuals do not or cannot effectively engage with public services. More broadly, research on artificial intelligence (AI) and disability has highlighted that many digital health and AI studies take a narrow view of disability risk (i.e. downplay the harms that AI systems can do to disabled users/participants), or omit this consideration entirely, essentially neglecting issues of bias, discrimination, and broader social implications for human rights and accessibility.[7] These omissions lead to similarly flawed evidence bases that shape guidance on AI fairness and use.

4 Synthetic data

4.1 What are synthetic datasets?

Synthetic data have been used on parallel areas of public health to address original dataset issues such as availability, privacy, incompleteness, and granularity. Synthetic data are artificially generated datasets that mimic the statistical characteristics, patterns, and relationships of real-world data while containing no personally identifiable information.[13] Researchers use algorithms, statistical models, and computational techniques to generate synthetic datasets that replicate distributions, correlations, and other properties of the original datasets at a larger scale. These datasets can then be used for tasks such as algorithm development, simulation, statistical modelling, and machine learning training, while still protecting individual privacy.[13] Synthetic data is of course heavily dependent on the quality and unbiased nature of the original dataset. At present, misrepresentation issues in a real-world dataset will be exacerbated when that dataset is used to generate synthetic data, which are then used in evidence generation. These issues are particularly salient when considering whether and how disabled people, and disability as a concept, are included in administrative data collection. To mitigate bias reproduction, data scientists must assess how inclusive the original dataset is and initiate inclusive preprocessing techniques. This process will require detailed and flexible guidance, which is a main component of the future work set out in this article.

4.2 Creating synthetic data

Various methods can be used to generate synthetic data. The UK Government's Digital Service[14] describes an especially popular method, the use of Gaussian (normal) distributions, which takes a similar approach to sampling for research purposes:

To reflect the population, we ensure that we know the feature distributions. We know what percentage of each feature is present as a proportion of the source data. From this we generate artificial datasets with these same percentage representations. We can also generate 'what if' scenarios if needed, such as altering the scale to emphasise lower populations, or to increase variability in subsets with high similarity.[14]

To improve the granularity of available data on health inequalities, Rice et al. developed a synthetic dataset for Great Britain at the electoral ward level, supporting analysis of economic inclusion and its links to population health.[15] The population was generated using simulated annealing, which iteratively selects and replaces individuals from a sample survey to ensure that aggregated characteristics match known area-level constraints. For each area, candidate populations were refined through repeated swaps of individuals, retaining changes that improve alignment with observed data until a specified tolerance is reached. The resulting SIPHER synthetic population can be viewed as a localised reweighting of a UK household survey (UKHLS) to small-area geographies. While direct validation is limited where benchmark data are unavailable, using SIPHER to fill such gaps is consistent with its intended applications and established practice in the literature.[15]

Methods can also be combined. Kaur et al further use Bayesian networks to incorporate prior understanding when generating synthetic data from health records, ideally enabling healthcare organisations to disseminate synthetic health data to researchers to generate hypotheses and develop analytic tools.[16] James et al describe a sequential analysis, which synthesises datasets variable by variable, wherein the more variables in a dataset, the more the sequence of the variables will need to be optimised.[17] Though these methods are promising, explicit use for generating synthetic disability data has not been documented. The complexities with this aim merit further consideration.

4.3 Public health policy use cases

More broadly, synthetic data offer valuable opportunities for public health policymaking, particularly in ageing policy, where decisions must reflect the population's age structure, demographics, and dynamic needs. By anonymising and making large-scale individual-level data available for evidence generation, synthetic datasets enable detailed analysis that would otherwise be restricted. More complete, detailed, flexible, and available data enables policymakers to collaborate with researchers to explore scenarios and test interventions in a

simulated population. Good quality synthetic data can accelerate data science and research, even contributing to data democratisation.[18]

Synthetic data have already been applied across public health research to attenuate common limitations of real-world data, including small sample sizes, non-representative sampling, and a lack of granular spatial resolution, to derive policy-level insights. Synthetic datasets can also support analysis and experimentation when access to sensitive data is restricted, or to examine the effects of limited representation in data. Synthetic population models have been developed to produce estimates at fine geographic scales, allowing policymakers to explore inequities and test the expected effects of interventions before implementation. For example, Wu et al developed a synthetic microdata population for individuals across Great Britain with detailed attributes enabling the modelling of health and socioeconomic outcomes at the small-area level, with both internal and external validation confirming that the simulated population adequately captured variation in health and income at this scale.[19] Saunders et al used a prototype data repository combining audiometric, real-world, and questionnaire data to model noise-induced hearing loss risk and inform hearing healthcare policy.[1] Rice et al created a dataset at the electoral ward level for Great Britain, focused on economic inclusion as a driver of improved population health and reduced health inequalities; the synthetic population was constructed using simulated annealing to match individual-level attributes to known area-level constraints.[15] Kraul and colleagues similarly constructed a synthetic population model to estimate excess cardiovascular disease risk, demonstrating the utility of synthetic data for modelling non-communicable disease burden and evaluating the likely outcomes of interventions in large populations.[20] It is essential to conduct and report validation measures for each of these applications.

4.4 Technical and ethical challenges

There are trade-offs when using synthetic data. While synthetic data retain high analytic value, they may smooth over rare or complex patterns and cannot always be fully validated against real-world benchmarks. Understanding both the strengths and limitations of synthetic data is essential to ensure they are used appropriately in informing dynamic, evidence-based public health policy.

Crucially, poorly generated synthetic data can introduce unrealistic patterns and distributions within the data; the model will learn these patterns and perform unsatisfactorally.[14] There are many technical challenges to effectively using synthetic data, which Jordon et al summarise:[18]

1. Synthetic data can still leak information about the data it was derived from, for example, if all key features are not removed.
2. Synthetic data that come with privacy guarantees are necessarily a distorted version of the real data, requiring any final tools to be deployed in the real world should be evaluated and fine-tuned on real data.
3. Outliers and low probability events are difficult to capture in a synthetic dataset in a private way—the synthetic data generator would either inaccurately replicate statistics, or reveal potentially private information about these individuals.
4. Evaluating the privacy of a single dataset can be problematic. Privacy arises from the mechanism generating the synthetic dataset and is not possible to rigorously evaluate by directly comparing it with real data.
5. Black box models can be opaque when it comes to generating synthetic data. The levels of accuracy and privacy of the datasets are hard to estimate and can vary significantly across the generated data points.

These technical challenges intersect with broader concerns about algorithmic bias,[6,7] where historical inequities and sampling errors embedded in source data are reproduced or amplified. James et al note the extent to which findings derived from synthetic data will be accepted within the scientific and medical communities remains unclear, with peer-reviewed research still needed to demonstrate results on real-world data until greater confidence in synthetic methods is established.[17]

These concerns persist amidst the limited oversight of AI systems generating such data,[13] and insufficient examination of their legal and ethical implications, particularly within the EU, where no coherent strategy for data collection currently exists.[21] In healthcare contexts, Nisevic et al highlight that while fully synthetic data may not constitute personal data, their downstream use in profiling and decision-making systems raises important concerns must be evaluated through the principles of autonomy, beneficence, non-maleficence, and justice, alongside the development of sector-specific standards and governance mechanisms.[22] There is also a growing recognition of the need to measure fairness in synthetic healthcare data,[23] particularly through ensuring equitable “resemblance” between real and synthetic data distributions across protected subgroups, while also addressing “representation” to identify over- or under-representation.[23] This requires the development of preprocessing guidance to enable equitable analysis of resultant datasets and the integration of fairness considerations across all stages of data generation. Specifically, Bhanot et al develop novel fairness metrics to address this aim.[23] Yet the very data limitations arising from sparsity also constrain the effective application of synthetic data, as real-world data remain essential for validating synthetic approaches. Consequently, these challenges are most acute in data-sparse contexts—precisely the settings that would benefit most from improved data availability.

4.5 Best practices for generation and use.

Using synthetic data effectively requires careful attention across the entire data use pipeline. Dankar et al emphasize that the utility of synthetic data depends on critical methodological choices.[24] Overall, researchers must ensure that the inclusion of artificially generated data has not unintentionally biased analytic models towards artificial patterns not reflected in real-world distributions.[14] To transparently address this concern, the UK Government's resource on synthetic data generation and use specifies that it is important to document and version control any synthetic datasets used in model development, so they may be evaluated and recreated if needed.[14] Particular care must be taken to identify and appropriately mask or transform unique or sensitive characteristics in the original data, with clear documentation to ensure transparency about any limitations this introduces.[11] Bias mitigation should be treated as an iterative process, with continuous validation against fairness metrics and the use of independent validation datasets or cross-validation techniques to detect unintended distortions.[14]

Yet best practices for fairness should begin even earlier in the preprocessing stage. Mitigating bias in synthetic data requires broader considerations of how representative the original dataset is, potentially using statistical and demographic methods to adjust the original dataset. For example, digital data collection for example will be impacted by differential access to smartphones and wearable sensors, as well as digital literacy, which often disproportionately affects participants who are lower-income, older, and/or disabled.[25] Differential levels of trust in data collection is important to consider when using and interpreting real and synthetic survey data, especially when concerning groups with disabilities. Individuals who have negative perceptions/experiences on interaction with health and social service systems, or extractive research practices, will be less willing to engage. Further, those who have experiences of personal data misuse, or historically-based concerns about surveillance, will also be less willing to participate fully. Public trust is therefore essential to build in advance of, and throughout, the research process. Researchers can facilitate this through transparent communication about how data are used and the societal benefits they enable. Indeed, studies show that the more people understand about how their data will be reused, the more willing they are to share their data for research.[5]

Various methods can also be incorporated to preserve participants' privacy during synthetic data generation. Removal of protected characteristic data is an important step during anonymisation, yet removing protected features can also remove key identifiers which qualify subsets of that data meaningfully. When removing these features, researchers must understand and evaluate the relationships between them and the outcomes, as well as the consequences of their removal.[14] Overall, there is a balance that needs to be considered when using synthetic data as to the quality of the data vs. its potential to be disclosive, which

is highly dependent on the intended use of the synthetic data.[11] “Noise injection” can also be used during the data generation process, wherein small, random variations are added to data so the resulting synthetic data are not an exact copy of the real observations.[14]

After generation and validation, transparent reporting of dataset composition, model assumptions, and limitations is also essential. For example, Kakampakou et al set out the role of the causal diagram informing data simulation in making their assumptions explicit and their process more interpretable. make their process more interpretable.[26] Responsibility also extends to beyond data producers: secondary users must critically assess the suitability and constraints of synthetic datasets for their intended applications.[11] Therefore, advancing fair use of synthetic data will require coherent efforts across user groups to establish standards and improve awareness that ensures synthetic data-driven systems across contexts.[5]

Many of the risks highlighted in the process of synthetic data generation disproportionately affect disabled people who are underrepresented or misrepresented in datasets. If these issues are not addressed, existing inequities will propagate through the pipeline of synthetic data usage. These risks also emphasise the role of inclusive data collection practices, early-stage bias mitigation for secondary data, and transparent documentation and reporting—all steps that are critical to ensure that disability is accurately and fairly reflected, trust is earned from disabled participants, and that findings are communicated inclusively.

If these steps can be achieved, synthetic data offers significant possibilities for disability research and policy. It can enable safer sharing of sensitive data, augment representation of small samples, and improve privacy while still allowing meaningful disaggregation of disability data in analyses. These data will support the development of fairer, more inclusive models and intervention, for example by helping to identify disparities in unmet needs for healthcare, assistive technology, and social services. When used responsibly, synthetic data use can lead to inclusive policy development that builds trust, enhances participation, and enables more equitable, data-driven interventions and planning.

5 The important research agenda

Despite the growing need to understand and respond to changing patterns of disability in ageing populations, best practices for disability data collection, representation, and use remain underdeveloped. This gap is most acute in low- and middle-income countries and other data-sparse settings, where the demand for evidence is high but available data are limited. Our wider research agenda seeking to create a fairer world includes addressing this imbalance, which requires the development and documentation of innovative, context-

sensitive data science methods that can mitigate data limitations while ensuring that disability is inclusively and appropriately represented.

Strengthening representation in population-based datasets is particularly urgent given the increasing use of synthetic data in public health research, alongside the current lack of systematic approaches to assessing and correcting disability misrepresentation prior to data synthesis. More broadly, the absence of coherent strategies for synthetic data use emphasises the need for clearer guidance and standards. Researchers in this space should contribute to a coherence on how to safely work with protected characteristics in synthetic data, with relevant guidance from/across other protected characteristics such as race, ethnicity, migration status, and sexuality.

A central priority of our research is the development of guidance and case studies on inclusive preprocessing—practical methods to assess and improve disability representation before and during synthetic data generation. By combining demographic methods and machine learning, we focus on approaches such as reweighting datasets using external population benchmarks, adjusting for differential non-response, and leveraging targeted surveys to model relationships that can be validated against general population data. More complex techniques, such as incorporating Bayesian inference and modelling,[27] will also be explored for imputing missing information. Additionally, we will test nested or linked data structures by embedding a representative disability sample within a broader population dataset, to enhance both marginal and conditional accuracy. However, such approaches introduce important challenges, including disclosure risks for small subgroups, the potential dominance of population-level structures over subgroup-specific patterns, and the need to make explicit normative decisions about which distributions to preserve. Addressing these challenges will be a throughline of our analytical work in this space.

An additional component to this work is exploring the role of qualitative research informing our work; in particular, when it focuses on how and why individuals engage with the data collection processes, build trust, and interact with research in the AI space. This evidence would highlight issues of accessibility, inclusivity, data protection, agency, and appropriateness that are not always visible in quantitative analysis and fairness research alone. Integrating qualitative insights with quantitative approaches represents a useful, emerging direction for improving both the measurement and interpretation of disability in data systems.

The progress of our research agenda and broader work to guide AI fairness will depend on methodological innovation as well as collective effort. Addressing the challenges of disability representation in synthetic data requires close collaboration across disciplines, sectors, and geographies, particularly between researchers, data producers, policymakers, and disabled communities. No single approach or setting will be sufficient to establish robust and transferable practices. We therefore see a clear need for greater coherence in how this work

is developed and applied. This work should involve the iterative development of shared principles, benchmarks, and evaluative frameworks; guidance on navigating risks for small or marginalised groups; and transparent, accessible case studies to guide secondary users. By working towards shared standards and a more unified approach, this research community can move beyond fragmented solutions in a rapidly changing data science landscape, towards a more inclusive and accountable research ecosystem that better supports equitable public health policy and ultimately, the populations it aims to serve.

6 References

1. Saunders, G. H. *et al.* Application of Big Data to Support Evidence-Based Public Health Policy Decision-Making for Hearing. *Ear Hear.* **41**, 1057 (2020).
2. Mont, D., Madans, J., Weeks, J. D. & Ullmann, H. Harmonizing Disability Data To Improve Disability Research And Policy: Commentary discusses harmonizing disability data to improve disability research and policy. *Health Aff. (Millwood)* **41**, 1442–1448 (2022).
3. Mitra, S., Chen, W., Hervé, J., Pirozzi, S. & Yap, J. *Invisible or Mainstream? Disability in Surveys and Censuses in Low- and Middle-Income Countries*. https://documents1.worldbank.org/curated/en/745481618324212396/pdf/Invisible-or-Mainstream-Disability-in-Surveys-and-Censuses-in-Low-and-Middle-Income-Countries.pdf (2021).
4. World Bank & Disability Data Initiative. *Disability Data Hub Methods Paper*. (2023).
5. Ibrahim, H., Liu, X., Zariffa, N., Morris, A. D. & Denniston, A. K. Health data poverty: an assailable barrier to equitable digital health care. *Lancet Digit. Health* **3**, e260–e265 (2021).
6. Aitken, Z., Walmsley, S., M Bishop, G., Badji, S. & Fortune, N. Methods used to construct disability indicators in linked administrative datasets: a systematic scoping review. *Popul. Health Metr.* **23**, 22 (2025).
7. El Morr, C. *et al.* AI and disability: A systematic scoping review. *Health Informatics J.* **30**, 14604582241285743 (2024).

8. Loeb, M. Disability statistics: an integral but missing (and misunderstood) component of development work. *Nord. J. Hum. Rights* **31**, 306–324 (2013).

9. World Health Organization & World Bank. *World Report on Disability*. https://www.who.int/teams/noncommunicable-diseases/sensory-functions-disability-and-rehabilitation/world-report-on-disability (2011).

10. Participation, E. Equality Act 2010. https://www.legislation.gov.uk/ukpga/2010/15/part/2/chapter/1.

11. Ethical considerations relating to the creation and use of synthetic data. *UK Statistics Authority* https://uksa.statisticsauthority.gov.uk/publication/ethical-considerations-relating-to-the-creation-and-use-of-synthetic-data/.

12. Hanass-Hancock, J. *et al. The Disability Data Report 2023*. https://disabilitydata.ace.fordham.edu/disability-data-report-2023/ (2023).

13. Ugarte, R. Synthetic Data and Health Equity. *Social Science Research Council* https://just-tech.ssrc.org/field-reviews/synthetic-data-and-health-equity/ (2024).

14. AI Insights: Synthetic Data. *GOV.UK* https://www.gov.uk/government/publications/ai-insights/ai-insights-synthetic-data-html.

15. Rice, H. P., Höhn, A., Meier, P., Heppenstall, A. & Lomax, N. An inclusive economy dataset for wards in Great Britain using administrative and synthetic data sources. *Sci. Data* **12**, (2025).

16. Kaur, D. *et al.* Application of Bayesian networks to generate synthetic health data. *J. Am. Med. Inform. Assoc. JAMIA* **28**, 801–811 (2021).

17. James, S., Harbron, C., Branson, J. & Sundler, M. Synthetic data use: exploring use cases to optimise data utility. *Discov. Artif. Intell.* **1**, 15 (2021).

18. Jordon, J. *et al.* Synthetic Data -- what, why and how? Preprint at https://doi.org/10.48550/arXiv.2205.03257 (2022).

19. Wu, G., Heppenstall, A., Meier, P., Purshouse, R. & Lomax, N. A synthetic population dataset for estimating small area health and socio-economic outcomes in Great Britain. *Sci. Data* **9**, 19 (2022).

20. Krauland, M. G. *et al.* Development of a Synthetic Population Model for Assessing Excess Risk for Cardiovascular Disease Death. *JAMA Netw. Open* **3**, e2015047 (2020).

21. Martins, Henrique. *EU Health Data Centre and a Common Data Strategy for Public Health*. (Publications Office of the European Union, 2021).

22. Nisevic, M., Milojevic, D. & Spajic, D. Synthetic data in medicine: Legal and ethical considerations for patient profiling. *Comput. Struct. Biotechnol. J.* **28**, 190–198 (2025).

23. Bhanot, K. *et al.* The Problem of Fairness in Synthetic Healthcare Data. *Entropy* **23**, (2021).

24. Dankar, F. K., Ibrahim, M., Dankar, F. K. & Ibrahim, M. Fake It Till You Make It: Guidelines for Effective Synthetic Data Generation. *Appl. Sci.* **11**, (2021).

25. Leung, W., Shi, L. & Jung, J. Are individuals with disabilities using wearable devices? A secondary data analysis of 2017 BRFSS. *Disabil. Rehabil. Assist. Technol.* **19**, 131–138 (2024).

26. Kakampakou, L. *et al*. Simulating hierarchical data to assess the utility of ecological versus multilevel analyses in obtaining individual-level causal effects. *BMC Med. Res. Methodol.* **25**, 79 (2025).

27. Sakamoto, M. & Kakuta, N. Bayesian modeling of underreported disabilities: Gender insights from the Bangladesh national household survey. *Disabil. Health J.* **19**, 101922 (2026).