

**Title:**

Actionability of Synthetic Data in a Heterogeneous and Rare Healthcare Demographic; Adolescents and Young Adults (AYAs) with Cancer

**Author list, including affiliations:**

J. (Joshi) Hogenboom<sup>1</sup>, A. (Ailara) Lobo Gomes<sup>1</sup>, A.L.A.J. (Andre) Dekker<sup>1</sup>, W.T.A. (Winette) Van Der Graaf<sup>2,3</sup>, O. (Olga) Husson<sup>2,4</sup>, L.Y.L. (Leonard) Wee<sup>1</sup>

1: Department of Radiation Oncology (Maastrro), GROW School for Oncology and Reproduction, Maastricht University Medical Centre+, Maastricht, the Netherlands

2: Department of Medical Oncology, Netherlands Cancer Institute, Amsterdam, the Netherlands.

3: Department of Medical Oncology, Erasmus MC Cancer Institute, Erasmus University Medical Centre, Rotterdam, the Netherlands.

4: Department of Surgical Oncology, Erasmus Medical University Centre, Rotterdam, the Netherlands

**Prior presentation:**

This study and its results are original and have not yet been presented elsewhere.

**Support:**

J. Hogenboom, A.L.A.J. Dekker, W.T.A. Van Der Graaf, O. Husson, and L.Y.L. Wee are supported by the European Union's Horizon 2020 research and innovation programme through The STRONG-AYA Initiative (Grant agreement ID: 101057482). A. Lobo Gomes is supported by Innovative Medicines Initiative (IMI), Digital Oncology Network for Europe (DigiONE), and the European Regional Development Fund (ERDF). O. Husson is also supported by the Netherlands Organization for Scientific Research through a Vidi grant (ID: 198.007). L.Y.L. Wee is also supported by ZonMW and Stichting Hanarth Fonds

**Acknowledgements:**

All individuals that have contributed to this study are included in the author list. Experiments were made possible using the Data Science Research Infrastructure (<https://dsri.maastrichtuniversity.nl/>) hosted at Maastricht University.

**Code availability:**

The code and software versioning used for this work are available on GitHub with a working example, see: <https://github.com/MaastrichtU-CDS/AYA-synthetic-data>.

**Conflict of interests:**

All authors declare to lack any competing conflicts of interest.

**NOTE:** This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

## **Abstract**

**Purpose:** Research on rare diseases and atypical healthcare demographics is often slowed by high inter-subject heterogeneity and overall scarcity of data. Synthetic data (SD) has been proposed as means for data sharing, enlargement, and diversification, by artificially generating ‘real’ phenomena while obscuring the ‘real’ subject data. The utility of SD is actively scrutinised in healthcare research, but the role of sample size for actionability of SD is insufficiently explored. We aim to understand the interplay of actionability and sample size by generating SD sets of varying sizes from gradually diminishing amounts of real subjects’ data. We evaluate the actionability of SD in a highly heterogeneous and rare demographic: adolescents and young adults (AYAs) with cancer.

**Methodology:** A population-based cross-sectional cohort study of 3735 AYAs was sub-sampled at random to produce 13 training datasets of varying sample sizes. We studied four distinct generator architectures built on the open-source Synthetic Data Vault library. Each architecture was used to generate SD of varying sizes based on each aforementioned training subsets. SD actionability was assessed by comparing the resulting SD to its respective ‘real’ data against three metrics – veracity, utility, and privacy concealment.

**Results:** All examined generator architectures yielded actionable data when generating SD with sizes similar to the ‘real’ data. Large SD sample size increased veracity but generally increased privacy risks. Using fewer training subjects led to faster convergence in veracity, but partially exacerbated privacy concealment issues.

**Conclusion:** SD is a potentially promising option for data sharing and data augmentation, yet sample size plays a significant role in its actionability. SD generation should go hand-in-hand with consistent scrutiny and sample size should be carefully considered in this process.

## **Introduction**

Healthcare data for rare diseases in heterogeneous populations – such as adolescents and young adults (AYAs) with cancer – are generally difficult to acquire, and sizable high-quality datasets on these subjects are not widely available. Research on this important group has thus been hampered by data availability, resulting in lack of standards for age-specific cancer care<sup>1</sup>. Sharing of datasets is regularly constrained by administrative burden and privacy-related challenges. However, it is exactly the research into such diseases that would benefit most from greater access to larger volumes of high-quality data.

Synthetic data (SD) has been suggested as a potential surrogate that addresses data availability and data acquisition<sup>2-7</sup>, since it has means to augment both sample size and data diversity. Here, SD is defined as artificially generated person-level health data that resembles – to some degree of authenticity – ‘real’ subject health data, whilst implying the absence of identifiable information of any actual persons. In this case, the goal of SD is to catalyse knowledge generation in healthcare, but without exposing real human subjects’ data. While use of SD within healthcare is still in development, ethicists, policy makers, and researchers are trying to establish how SD can be used appropriately<sup>5,8</sup>. Meanwhile, various techniques<sup>9-12</sup> are now openly available which researchers can use to train generative models and thus create SD for various purposes<sup>13-22</sup>.

Constructing a SD generator always needs some initial form of real data for its training. A trained generative network (i.e., generator) is then used to produce any arbitrary size of SD. The sample size of training data is therefore an important question that also depends on the intended use of SD. Likewise, it is not universally known – for a given generator – if there are upper and lower sample size limits of SD that ought to be produced. The interplay of actionability and sample size (both for training and generation) therefore warrants careful examination.

What designates SD as “actionable” remains a disputed topic<sup>2,23,24</sup>; in this work, we propose three evaluation domains that may be considered as minimally necessary for actionable SD. (1) Veracity, meaning that the generated SD is statistically similar to the training data in terms of distribution of values. (2) Utility, meaning that SD replicates the inter-variable associations within the training data, for example, the coefficients in a logistic regression model. (3) Privacy concealment, in terms of obscuring individual identity and obscuring their observable attributes.

To elucidate the dependence of veracity, utility, and privacy concealment on sample size, we ran training and SD-generating experiments using real AYA cancer data (persons aged 18 to 39 years at first cancer diagnosis). Cancer incidence is low in this demographic (only 1 million new cases annually worldwide), making this group difficult to study by conventional methods<sup>1</sup>. Though this group experiences the burden of cancer in a starkly different way compared to paediatric and elderly patients<sup>25,26</sup>, AYAs are predominantly treated in either paediatric or adult care institutes. Evidence to inform personalised age-specific patterns of cancer care is very limited for AYAs, hence our interest in understanding how SD might potentially be used to address AYA health-related hypotheses.

## **Methods**

Data from the SURVAYA study<sup>25,26</sup> (clinicaltrials.gov NCT05379387) was re-used with permission from the sponsor. SURVAYA was a population-based cross-sectional cohort study amongst AYA cancer patients registered in the Netherlands Cancer Registry (NCR). SURVAYA included AYAs that had been treated either in a Dutch academic medical centre or in the Netherlands Cancer Institute, and were registered in the NCR between 1999 to 2015; Amsterdam AMC, Netherlands Cancer Institute, and Utrecht UMC supplied treatment-related data up to 2014. The principal instruments in SURVAYA were validated health-related Quality of Life (QoL) questionnaires<sup>27-30</sup>, which were linked to clinical attributes from the NCR. Details are provided in the original study publications<sup>25,26</sup>. In total, there were 950 variables, some of which contained unstructured text.

We limited ourselves only to variables selected by Saris et al. (2022) for logistic regression models of negative body image among AYAs<sup>25</sup>. As per the original research, participants that had filled in less than half of the body image-related questionnaire items were excluded. We imputed missing values using multiple imputation by chained equations. Age at diagnosis was left in as a privacy concealment challenge for attribute inference, since certain cancer diagnoses are widely known to be exceptionally rare among younger AYAs. The pre-processed training data, hereafter referred to as the original dataset (OD), consisted of 3735 AYAs with 21 variables (37 after a one-hot encoding of categories) – see Supplementary material S1.

Training and generating experiments are illustrated in Figure 1. The OD was sampled without replacement to produce smaller training datasets of sizes 3600, 3350, ..., down to 600 subjects, resulting in a total of 14 distinct training sets.

We tested four SD generator architectures from the Synthetic Data Vault (SDV)<sup>9</sup> Python library. Two of these were machine learning (ML)-based architectures that used strictly parametric statistical models. The first was a speed-optimised Gaussian copula generator (hereafter “Fast ML”) and the other was a classical Gaussian copula generator (hereafter “Gaussian Copula”). Two non-parametric architectures were based on deep-learning (DL) neural networks; a Conditional Tabular Generative Adversarial Network (CTGAN)<sup>12</sup> and a Conditional Generative Adversarial Network incorporating Differential Privacy (DP-CGAN)<sup>10</sup>. The DP-CGAN was not part of SDV’s current library, but was built on top of existing SDV-library modules.

In the Fast ML architecture, each of the aforementioned subsets of the OD were used for training, thus resulting in 14 separately trained Fast ML generators. This process was then repeated entirely for the Gaussian Copula, CTGAN, and DP-CGAN architectures, thus always resulting in 14 separately trained generators within each of the four selected architectures.

Each architecture used default (hyper-)parameters for training specified by its developers. We limited our experiments only to the default hyper-parameters, without fine-tuning and no grid search for these experiments. All architectures were supplied with the metadata of the OD as required in the documentation<sup>10,31</sup>. Hyper-parameters per architecture and variable metadata are provided in supplementary material S2 and S3, respectively.

From each of the separately trained generators (thus 4 architectures x 14 training OD subsets per architecture = 56 unique generators), we generated SD of varying samples sizes ranging from 100 up to 39100 artificial subjects, in increments of 1000. Thus, for each uniquely trained generator, we derived 40 distinct sets of SD (56 generators x 40 sets of SD per generator = 2240 SD sets in total).

Actionability was assessed by comparing a selected SD set, from a given generator, to the corresponding OD training set used to train that generator. For instance, a SD set of 5100 samples generated with a Fast ML generator that was trained using a set of 3735 real subjects was compared to those specific 3735 real subjects.

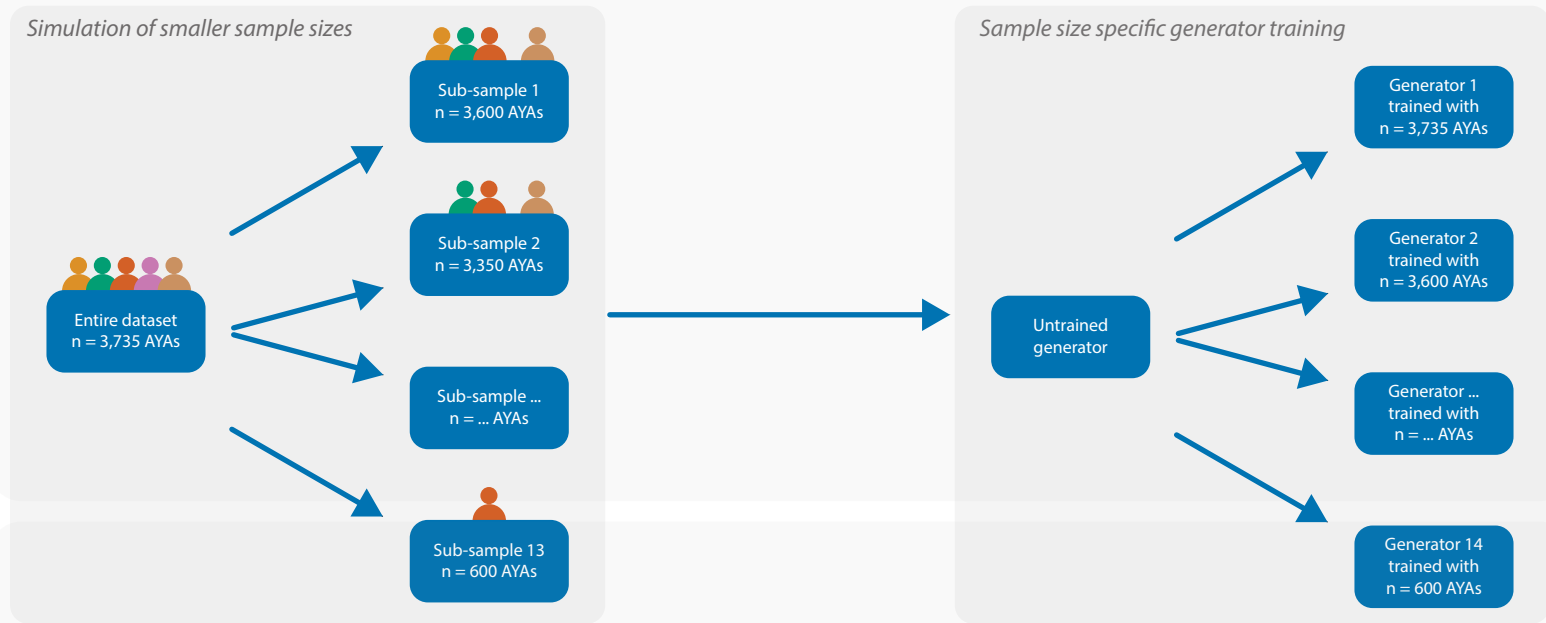
Veracity of SD was quantified in terms of precision, density, recall, and coverage<sup>32-34</sup>. The precision metric quantifies the proportion of individual subjects in the SD that fall within a minimum number of neighbouring subjects ( $k$ ) found in the OD. The density corrects for outliers in precision by scoring the number of samples inside the densely overlapping regions. Recall describes the proportion of OD samples that lay close to the SD's samples. Finally, coverage corrects for outliers in recall by building the radii with the OD's samples, rather than the SD's samples. A schematic overview of the veracity metrics' mechanism is provided in supplementary material S4. The metrics were computed using the *prdc* Python library with a constant  $k = 5$ .

Utility was examined by checking if a given SD reproduced the same logistic regression coefficients as the original paper by Saris et al.<sup>25</sup>, for generative architectures that had been trained using all of 3735 real subjects. For each generator, we selected an SD output size of 3100 for the comparison of odds ratios (ORs) with Saris et al. The regression analyses were conducted using Python libraries; *Statsmodels* for regression analysis and *scikit-learn* for an optimism-adjusted (through random under-sampling) concordance-statistic as area-under receiver-operator curve score.

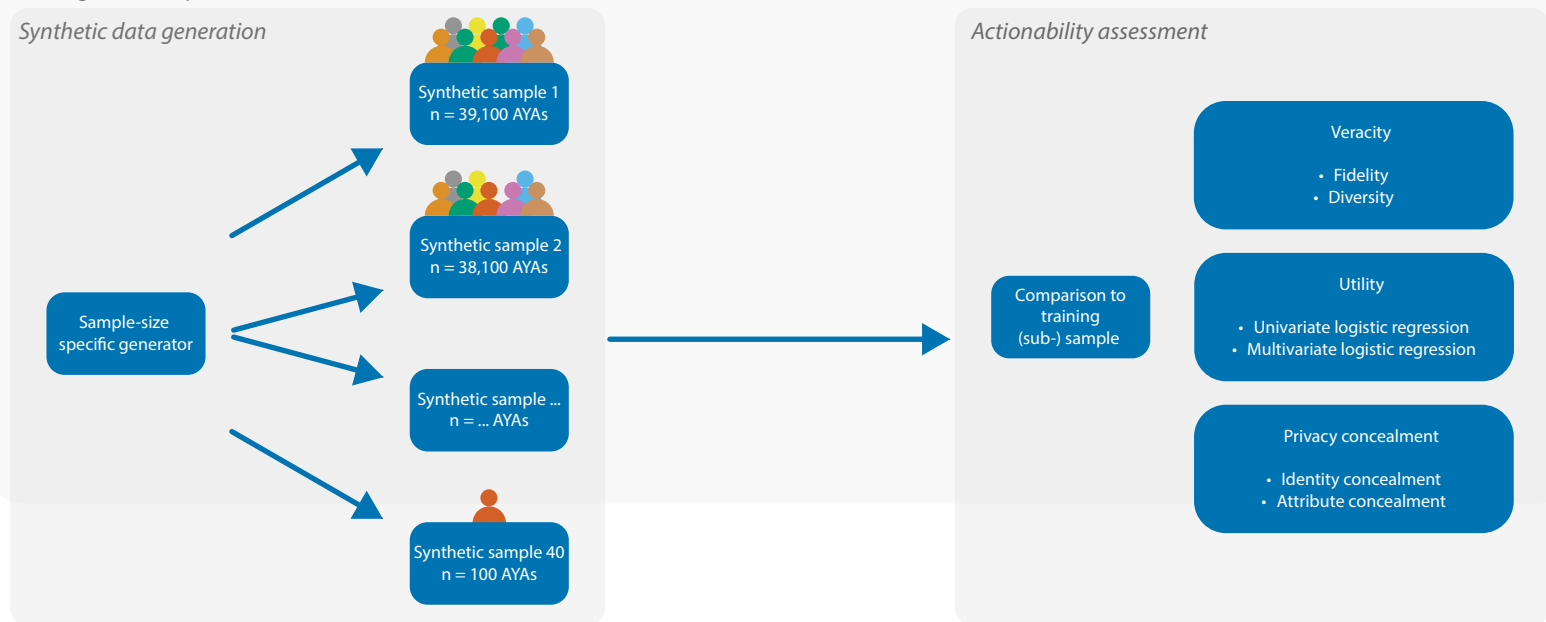
Privacy concealment was assessed with two metrics. First, a measure of the minimum Hamming distance<sup>10</sup> between any 'synthetic' subject in a given SD to any 'real' subject in the OD used to train the corresponding generator. Second, we estimated an attribute inference probability using seemingly innocent information that might be obtained by through public channels. We limited the test variables to age in nearest whole year, type of cancer and romantic partnership status, which we assumed would be easily gleaned from social media. We then assumed that an attacker would infer a sensitive and not easily known attribute, i.e., an AYA's self-perception of sexual attractiveness (a primary outcome in SURVAYA). The attribute inference attack was simulated using a Correct Attribution Probability algorithm (CAP)<sup>35</sup>.

Figure 1: Overview that describes the experimental setup to evaluate the role of OD and SD sample size on SD actionability.

I. Generator training process



II. Data generation process



Overview



## **Results**

### Veracity

Figure 2 shows a sensitivity of coverage towards training and SD sample size. For the generator architectures we tested, reasonable coverage ( $> 0.75$ ) was obtained for a wide range of training sample sizes, as long as more than 3000 synthetic subjects were generated. Among the four architectures, CTGAN overall yielded lower coverage for combinations of training and SD sample sizes. When the original data was relatively small in size and lacking in diversity, the chance of finding synthetic subjects within the radii of 5 closest real subjects was relatively high, hence the coverage seems counterintuitively better for smaller training set size.

Density was for the most part independent of training and SD sample size, as shown in Figure 2. However, CTGAN shows a stepwise improvement in the density metric from 1100 to 3100 training sample size. Similarly, the precision and recall scores were acceptably high and were independent of either training or SD sample size; the precision and recall curves are provided in supplementary material S5 and minimum scores were 0.86, 0.90, 0.75, and 0.93 for Fast ML, Gaussian Copula, CTGAN, and DP-CGAN respectively for both precision and recall.

### Utility

The overall picture for the ORs in univariable and multivariable logistic regressions against negative body image was quite mixed. There was a moderate degree of overlapping associations among the variables, where the confidence interval of the OR estimated from the SD included the same OR estimated in the OD. However, the number of overlapping associations between the SD and OD varied between different generator architectures.

Further, ORs that were not statistically significant in the OD ( $p < 0.05$ ) had become newly significant in the SD. Additionally, some of the ORs from the SD had been perturbed so much that its confidence interval no longer contained the OR derived from the OD, and other ORs had shifted effects.

The number of overlapping, shifted, and newly significant effects generally increased in multivariable regression compared to univariable regression. Table 1 summarises the number of perturbations in the ORs observed from univariable and multivariable regressions, in terms of overlap or shifted OR estimates in the SD relative to the OD, and the number of statistically significant variables in the OD and SD.

The adjusted c-statistics for multiple regression over all variables on negative body image were 0.87, 0.88, 0.86, and 0.86, for Fast ML, Gaussian Copula, CTGAN, and DP-CGAN generators, respectively.

Forest plots of the ORs (see supplementary material S6 to S13) show that the effect estimates from the OD had not been consistently reproduced in the SD, irrespective of generator architecture.

Figure 2: Veracity in terms of coverage (on the left side of the figure) and density (on the right side of the figure), per colour-coded training data sample size. The veracity score relates to the prdc's outcome, where a value of 1 is best, and 0 is worst. Each horizontal row of figures depict the respective scores of the included generator architectures.

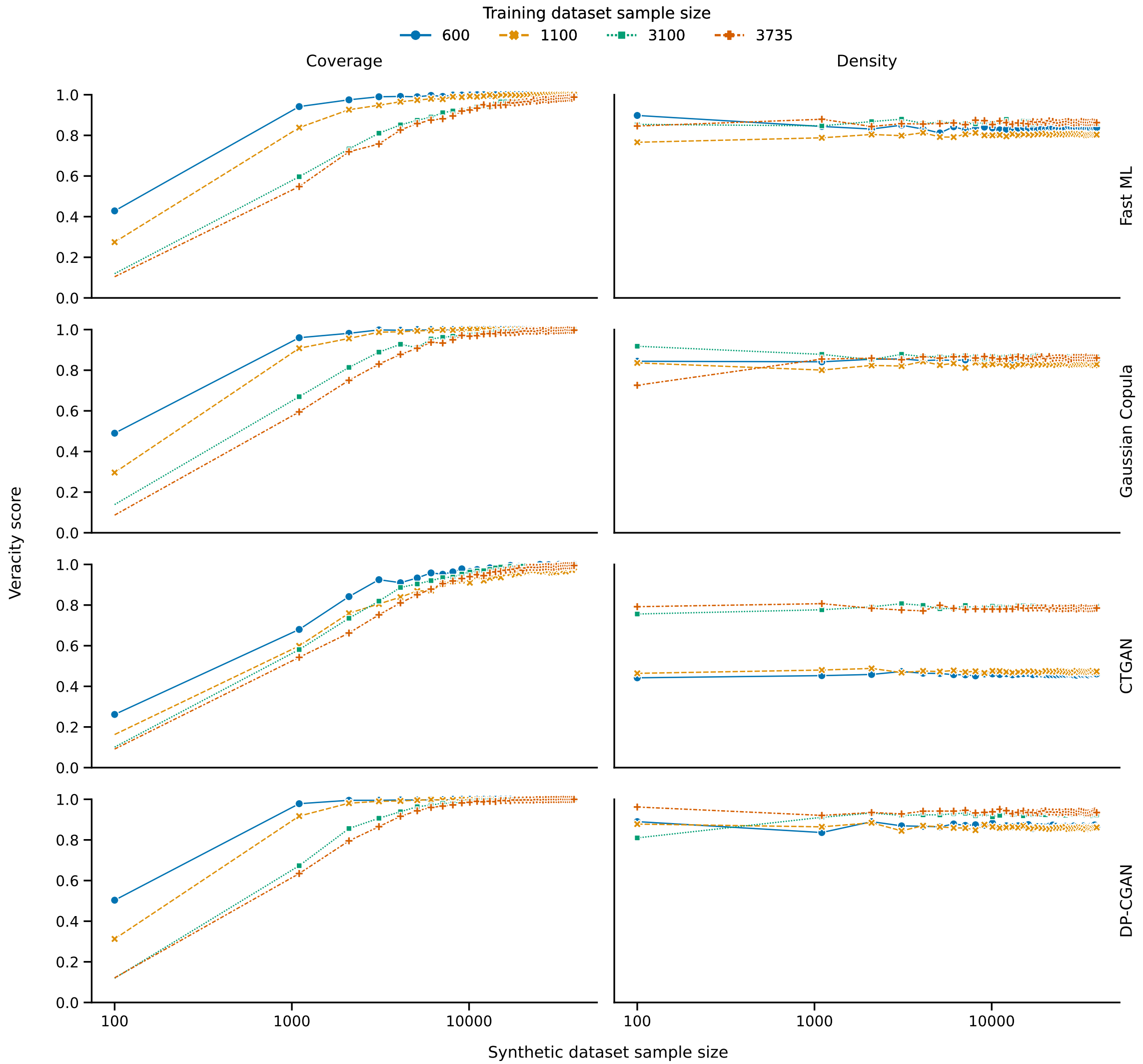




Table 1: Summary of the number of overlapping, shifted and statistically significant associations (after one-hot encoding) comparing a synthetic dataset of 3100 subjects to the original dataset of 3735 subjects.

<i>Architecture</i>		<i>Overlapping associations</i>	<i>Shifted associations</i>	<i>Significant in SD</i>	<i>Significant in OD</i>
<i>Fast ML</i>	Univariable	16 / 37	3 / 37	18 / 37	25 / 37
	Multivariable	26 / 37	1 / 37	14 / 37	18 / 37
<i>Gaussian Copula</i>	Univariable	13 / 37	11 / 37	17 / 37	25 / 37
	Multivariable	18 / 37	9 / 37	12 / 37	18 / 37
<i>CTGAN</i>	Univariable	17 / 37	6 / 37	31 / 37	25 / 37
	Multivariable	20 / 37	7 / 37	13 / 37	18 / 37
<i>DP-CGAN</i>	Univariable	22 / 37	5 / 37	26 / 37	25 / 37
	Multivariable	23 / 37	9 / 37	14 / 37	18 / 37

## Privacy concealment

For a given training sample, as one generates ever-larger sample sizes of SD, for all generators *except* DP-CGAN, there comes a point where a ‘real’ subject’s information becomes replicated into the SD. This would correspond to a minimum Hamming distance in given SD equalling to zero.

Figure 3 illustrates how the probability of a real patient being replicated in an SD dataset generally tends to increase for all architectures except for DP-CGAN. The figure shows the proportion of identical subjects in training and SD set, normalised to the size of the training set. The DP-CGAN appears entirely insensitive to training and generation sample size, as we were consistently able to generate up to 39100 SD samples without replicating any real subjects.

For Fast ML and Gaussian Copula, when generating 39100 subjects in the SD, there will be approximately 5% of synthetic subjects perfectly resembling ‘real’ subjects, and this occurred independently of the training sample size. The number of replicated subjects in the SD decreases roughly linearly towards zero when (obviously) the SD size is zero.

The training sample size of the CTGAN is notable; for 1100 training subjects and fewer, the number of replicated subjects is very low, barely just above that of the DP-CGAN which was always zero. However, when a training sample size of 3100 or more is used, the CTGAN switched to a state where it was much more likely than Fast ML or Gaussian Copula to replicate a real subject in the SD. This consistently reaches up to 15% of synthetic subjects perfectly resembling ‘real’ subjects, when 39100 synthetic subjects are generated. Note that this is the same range of training sample dependence where the veracity metric of density – and precision – switches from poor (~0.50) to acceptable (~0.70) for the CTGAN, in Figure 2. However, the training and generation loss curves (see supplementary material S14 and S15) did not specifically indicate a clear sign of ‘model collapse’ for the CTGAN in this range of training sample sizes.

Even if real subjects’ values are not fully replicated in the SD, it might still be possible to identify sensitive information through attribute inference attacks. Figure 4 plots the attribute inference score from the CAP algorithm. Higher scores imply better concealment, therefore presenting lower risk of exposing sensitive information (here, self-perception of attractiveness) via an attribute inference attack using only the SD. Concealment score for Fast ML and Gaussian Copula were insensitive to training sample size and to SD sample size; median scores were 0.41 and 0.46, respectively. Concealment in CTGAN and DP-CGAN was dependent on training sample size, but not dependent on SD output size; range of scores were 0.32 to 0.40 and 0.34 to 0.42, respectively. However, Figure 4 shows that attribute inference is not directly related to replication of real subjects in the SD.

Figure 3: Identity concealment in relation to varying training and synthetic data sample size, per colour-coded generator architecture. Each horizontal row depicts the respective scores per training dataset sample size.

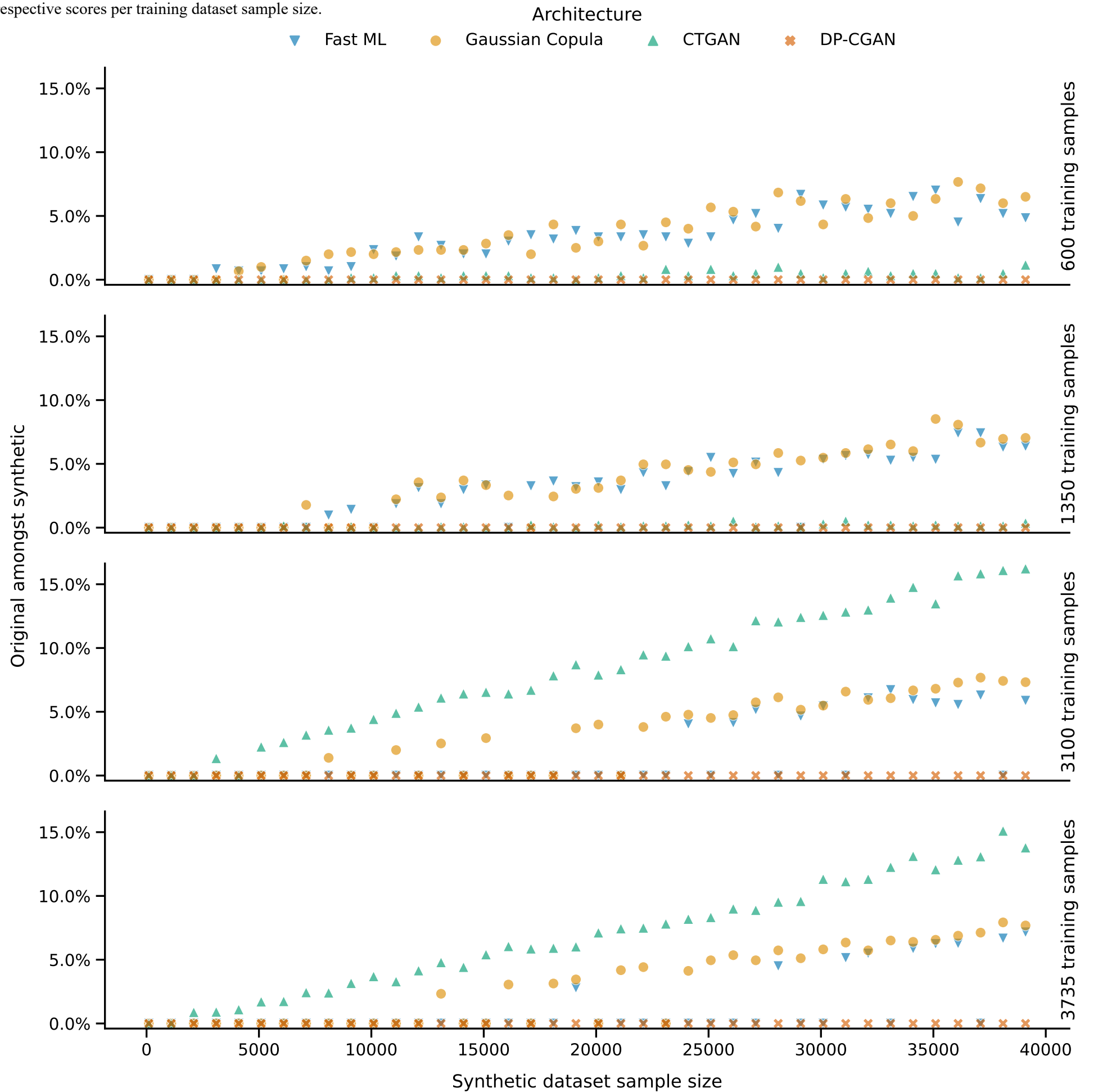
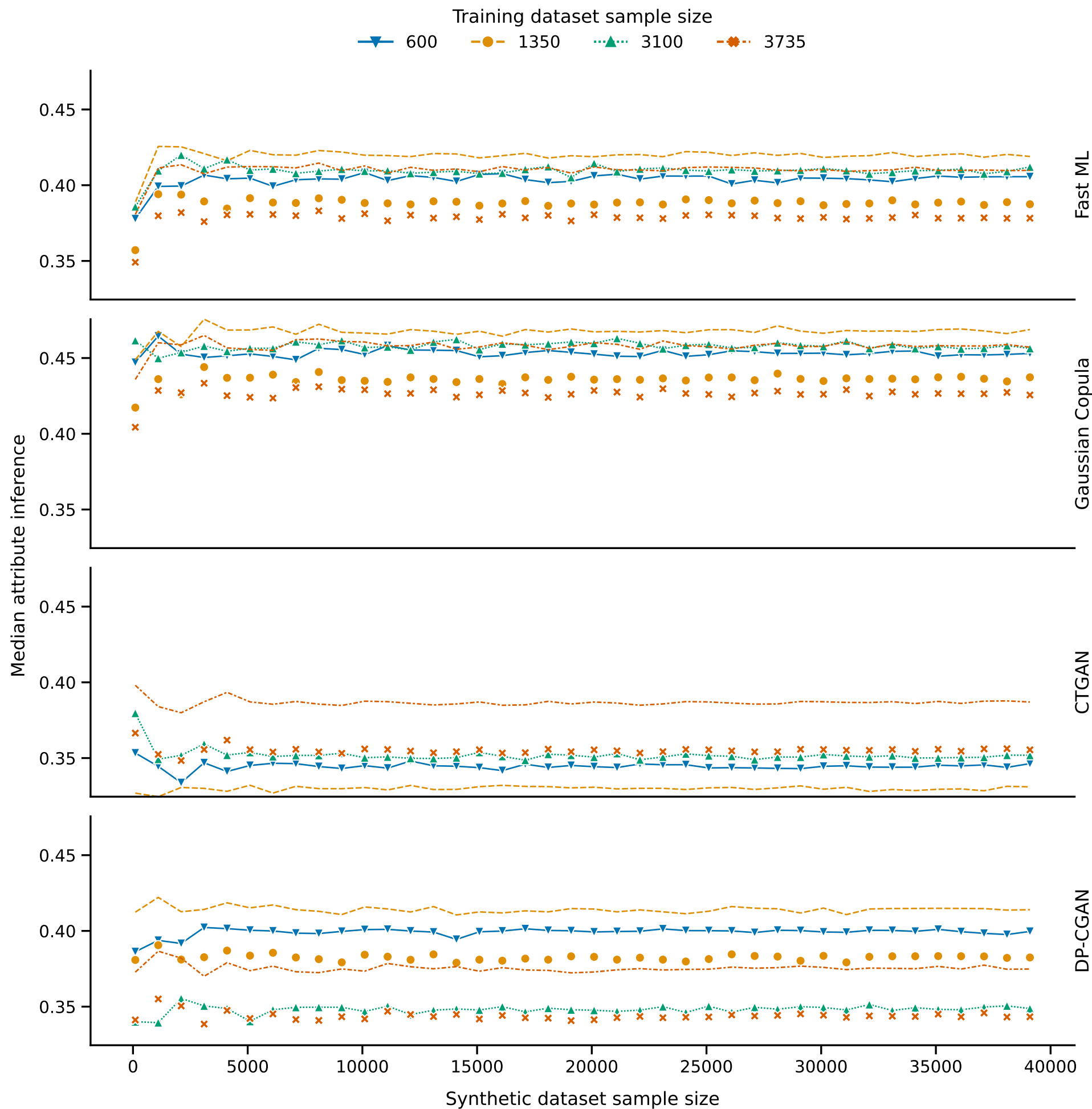


Figure 4: Attribute concealment in relation to varying synthetic data sample size, per colour-coded training data sample size. The median attribute inference score relates to the SDMetric's CAP-algorithm outcome, where a value of 1 is best, and 0 is worst. Each horizontal row depicts the respective scores of the included generator architectures.



## **Discussion**

Sample size played an important role in SD actionability. Training and SD sample size influenced veracity, as high diversity was only obtained through an appropriate balance between training and SD sample size regardless of generator architecture; the CTGAN architecture did however require proportionally larger SD sizes to obtain equal diversity. Utility was only assessed with a SD sample size similar to the training data, and we found that inter-variable associations were not uniformly reproducible in the SD. Privacy concealment was sensitive to sample size. The Fast ML and Gaussian Copula architectures had a distinct decrease in identity concealment that was proportional to both training and synthetic sample sizes. CTGAN was able to achieve up to a factor 65 SD enlargement with smaller training datasets, but was susceptible to identity disclosure when exceeding a certain training sample sizes. The DP-CGAN architecture was the best performer in terms of identity concealment as in none of our experiments it had included any training data in the SD. All SD had inferable attributes, irrespective of generator architecture, which carried finite risk of disclosing sensitive real information.

Arora and Arora already investigated SD sample size and found it less influential, but their sample size ranges were narrow, and they acknowledged the need for broader investigation<sup>15</sup>. To the best of our knowledge, our combination of varying training and SD size is unique, whereas other work consistently focusses on balancing representativeness and privacy for a given training and SD sample size, hindering further comparison to our findings. The minimum Hamming distance for identity concealment indeed presents the most conservative scenario. However, the experiment was informative insofar as demonstrating the increasing risk of disclosure as SD size increases, and the susceptibility thereof among generator architectures. Inescapably, failed identity concealment can potentially result in re-identification with an unfortunate combination of identifiable attributes that act as ‘real’ data held by a bad actor<sup>36</sup>.

We assume most clinical SD creators have limited computational resources and know-how to perform architecture adaptation and exhaustive tuning. Thus, we chose four accessible and easy-to-use generator architectures with default settings. It is however conceivable that actionability would improve after adaptation and tuning<sup>37</sup> – despite apparent over or under-fitting – yet we conceive this to be feasible only for a fraction of creators.

The implications can differ per use-case and accordingly require appropriate considerations. For instance, when aiming to augment a dataset, certain architectures appear to have an ‘upper’ enlargement limit, but also when sharing SD with a small sample size, for instance to prototype certain applications, the SD should at least be sufficiently large to capture the ‘real’ diversity. Thus, it is advisable that SD creators consistently scrutinise the suitability of their product for a certain purpose.

The (in-)dispensability of consistent SD scrutiny – in light of sample size – might substantially be influenced by ‘real’ subjects’ characteristics and sample size’s role requires further evaluation. In part, the necessity of consistent scrutiny reiterates the need for established methods to assess SD’s actionability<sup>2,23,24</sup>. Regardless, SD can technically be made actionable, and exuberant innovation can not only simplify this process and appraisal thereof but can incorporate here unmentioned domains such as SD fairness<sup>38</sup>, yet for true SD actionability in healthcare established ethical and legal consensus is paramount as these should dictate what actionability actually implies<sup>5,8</sup>.

Overall, SD is indeed able to resemble ‘real’ data to a moderate degree. In our experiments with AYA cancer data, we found that sample size has various roles in defining SD’s actionability. Typically, SD sample size had to be sufficiently large to encapsulate the entire veracity, yet SD sample size also should not be too large, as it might exacerbate flaws in utility and increases privacy risks. The training sample size dictated this balance between veracity, privacy, and SD sample size, in which smaller training sample sizes generally appear best suited with smaller SD sample sizes.

## References

1. Collaborators GBDAYAC: The global burden of adolescent and young adult cancer in 2019: a systematic analysis for the Global Burden of Disease Study 2019. *Lancet Oncol* 23:27-52, 2022
2. Rajotte JF, Bergen R, Buckeridge DL, et al: Synthetic data as an enabler for machine learning applications in medicine. *iScience* 25:105331, 2022
3. Goncalves A, Ray P, Soper B, et al: Generation and evaluation of synthetic patient data. *BMC Med Res Methodol* 20:108, 2020
4. Rodriguez-Almeida AJ, Fabelo H, Ortega S, et al: Synthetic Patient Data Generation and Evaluation in Disease Prediction Using Small and Imbalanced Datasets. *IEEE J Biomed Health Inform* 27:2670-2680, 2023
5. Jacobsen BN: Machine learning and the politics of synthetic data. *Big Data & Society* 10, 2023
6. Gonzales A, Guruswamy G, Smith SR: Synthetic data in health care: A narrative review. *PLOS Digit Health* 2:e0000082, 2023
7. Jacobs F, D'Amico S, Benvenuti C, et al: Opportunities and Challenges of Synthetic Data Generation in Oncology. *JCO Clin Cancer Inform* 7:e2300045, 2023
8. Shanley D, Hogenboom J, Lysen F, et al: Getting real about synthetic data ethics : Are AI ethics principles a good starting point for synthetic data ethics? *EMBO Rep* 0:1-4, 2024
9. Montanez A, Massachusetts Institute of Technology. Department of Electrical Engineering and Computer Science.: SDV : an open source library for synthetic data generation, 2018
10. Sun C, van Soest J, Dumontier M: Generating synthetic personal health data using conditional generative adversarial networks combining with differential privacy. *J Biomed Inform* 143:104404, 2023
11. Walonoski J, Kramer M, Nichols J, et al: Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. *J Am Med Inform Assoc* 25:230-238, 2018
12. Xu L, Skoularidou M, Cuesta-Infante A, et al: Modeling tabular data using conditional GAN, *NeurIPS2019 - Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2019, pp 7335–7345
13. Azizi Z, Zheng C, Mosquera L, et al: Can synthetic data be a proxy for real clinical trial data? A validation study. *BMJ Open* 11:e043497, 2021
14. Foraker RE, Yu SC, Gupta A, et al: Spot the difference: comparing results of analyses from real patient data and synthetic derivatives. *JAMIA Open* 3:557-566, 2020
15. Arora A, Arora A: Machine learning models trained on synthetic datasets of multiple sample sizes for the use of predicting blood pressure from clinical data in a national dataset. *PLoS One* 18:e0283094, 2023
16. Smith A, Lambert PC, Rutherford MJ: Generating high-fidelity synthetic time-to-event datasets to improve data transparency and accessibility. *BMC Med Res Methodol* 22:176, 2022
17. Greenberg JK, Landman JM, Kelly MP, et al: Leveraging Artificial Intelligence and Synthetic Data Derivatives for Spine Surgery Research. *Global Spine J* 13:2409-2421, 2023
18. Reiner Benaim A, Almog R, Gorelik Y, et al: Analyzing Medical Research Results Based on Synthetic Data and Their Relation to Real Data Results: Systematic Comparison From Five Observational Studies. *JMIR Med Inform* 8:e16492, 2020

19. D'Amico S, Dall'Olio D, Sala C, et al: Synthetic Data Generation by Artificial Intelligence to Accelerate Research and Precision Medicine in Hematology. *JCO Clin Cancer Inform* 7:e2300021, 2023
20. Kim H, Jang WS, Sim WS, et al: Synthetic Data Improve Survival Status Prediction Models in Early-Onset Colorectal Cancer. *JCO Clin Cancer Inform* 8:e2300201, 2024
21. Zhang Y, Wang ZX, Zhang Z, et al: GAN-based one dimensional medical data augmentation. *Soft Computing* 27:10481-10491, 2023
22. El Kababji S, Mitsakakis N, Fang X, et al: Evaluating the Utility and Privacy of Synthetic Breast Cancer Clinical Trial Data Sets. *JCO Clin Cancer Inform* 7:e2300116, 2023
23. Yan C, Yan Y, Wan Z, et al: A Multifaceted benchmarking of synthetic electronic health record generation models. *Nat Commun* 13:7609, 2022
24. Hernandez M, Epelde G, Alberdi A, et al: Synthetic data generation for tabular health records: A systematic review. *Neurocomputing* 493:28-45, 2022
25. Saris LMH, Vlooswijk C, Kaal SEJ, et al: A Negative Body Image among Adolescent and Young Adult (AYA) Cancer Survivors: Results from the Population-Based SURVAYA Study. *Cancers (Basel)* 14, 2022
26. Vlooswijk C, Poll-Franse LVV, Janssen SHM, et al: Recruiting Adolescent and Young Adult Cancer Survivors for Patient-Reported Outcome Research: Experiences and Sample Characteristics of the SURVAYA Study. *Curr Oncol* 29:5407-5425, 2022
27. Pols MA, Peeters PH, Ocke MC, et al: Estimation of reproducibility and relative validity of the questions included in the EPIC Physical Activity Questionnaire. *Int J Epidemiol* 26 Suppl 1:S181-9, 1997
28. Betegon E, Rodriguez-Medina J, Del-Valle M, et al: Emotion Regulation in Adolescents: Evidence of the Validity and Factor Structure of the Cognitive Emotion Regulation Questionnaire (CERQ). *Int J Environ Res Public Health* 19:3602, 2022
29. van Leeuwen M, Kieffer JM, Young TE, et al: Phase III study of the European Organisation for Research and Treatment of Cancer Quality of Life cancer survivorship core questionnaire. *J Cancer Surviv* 17:1111-1130, 2023
30. Aaronson NK, Ahmedzai S, Bergman B, et al: The European Organization for Research and Treatment of Cancer QLQ-C30: a quality-of-life instrument for use in international clinical trials in oncology. *J Natl Cancer Inst* 85:365-76, 1993
31. DataCebo: Single Table Metadata JSON, 2023
32. Sajjadi MSM, Bachem O, Lucic M, et al: Assessing Generative Models via Precision and Recall. *Advances in Neural Information Processing Systems* 31 (Nips 2018) 31, 2018
33. Kynkäänniemi T, Karras T, Laine S, et al: Improved precision and recall metric for assessing generative models, *NeurIPS2019 - Proceedings of the 33rd International Conference on Neural Information Processing Systems, 2019*, pp 3927–3936
34. Naem M, Oh S, Uh Y, et al: Reliable Fidelity and Diversity Metrics for Generative Models, *Proceedings of Machine Learning Research - Proceedings of the 37th International Conference on Machine Learning, 2020*, pp 7176-7185
35. CategoricalCAP - SDMetrics, 2022
36. El Emam K, Mosquera L, Bass J: Evaluating Identity Disclosure Risk in Fully Synthetic Health Data: Model Development and Validation. *J Med Internet Res* 22:e23139, 2020
37. Yeom S, Giacomelli I, Fredrikson M, et al: Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting, *2018 IEEE 31st Computer Security Foundations Symposium (CSF), 2018*, pp 268-282



38. Bhanot K, Qi M, Erickson JS, et al: The Problem of Fairness in Synthetic Healthcare Data. Entropy (Basel) 23, 2021

Variables of the original dataset, variables with an asterisk (\*) are summed questionnaire scores and were considered categorical to avoid generation of impossible outcomes i.e., scores that are technically not possible when determining a score.

<b>CATEGORICAL</b>	<b>CONTINUOUS</b>	<b>DICHOTOMOUS</b>
Tumour type	Age	Biological sex
Tumour stage		Romantic partnership
Educational level		Chemotherapy
Time since diagnosis		Radiotherapy
Body mass index		Hormone therapy
Maladaptive coping style*		Targeted therapy
Physical functioning*		Sexual attractiveness
Role functioning*		Negative body image
Emotional functioning*		
Cognitive functioning*		
Social functioning*		
Global health status*		

## The Fast ML generator does not have modifiable parameters

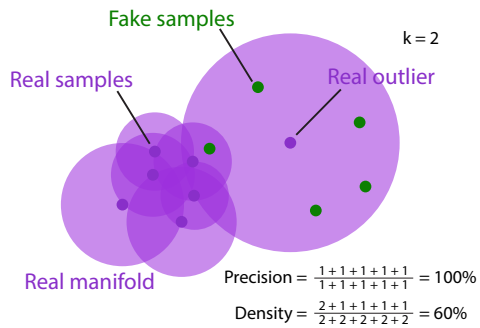
<i>Default (hyper)parameters</i>	<i>Fast ML</i>	<i>Gaussian Copula</i>	<i>CTGAN</i>	<i>DP-CGAN</i>
<i>Synthetic Data Vault version</i>	0.18.0	0.18.0	0.18.0	0.14.0
<i>Field distribution</i>		None		
<i>Default distribution</i>		Truncated Gaussian		
		Deep learning parameters		
<i>Epochs</i>			300	2000
<i>Batch size</i>			500	100
<i>Log frequency</i>			True	True
<i>Embedding dimension</i>			128	128
<i>Generator dimension</i>			(256, 256)	(128, 128, 128)
<i>Compress dimension</i>				
<i>Decompress dimension</i>				
<i>Discriminator dimension</i>			(256, 256)	(128, 128, 128)
<i>Generator learning rate</i>			$2 * 10^{-4}$	$1 * 10^{-4}$
<i>Generator decay</i>			$1 * 10^{-6}$	$1 * 10^{-6}$
<i>Discriminator learning rate</i>			$2 * 10^{-4}$	$1 * 10^{-4}$
<i>Discriminator decay</i>			$1 * 10^{-6}$	$1 * 10^{-6}$
<i>Discriminator steps</i>			1	1
<i>Verbose</i>			True	True
<i>CUDA</i>			True	True
<i>Private</i>				False
<i>L2scale</i>				
<i>Loss factor</i>				

The variable information (or metadata) was provided to Synthetic Data Vault (SDV) generators as a Python dictionary stored in JSON format as necessary for SDV's specific version.

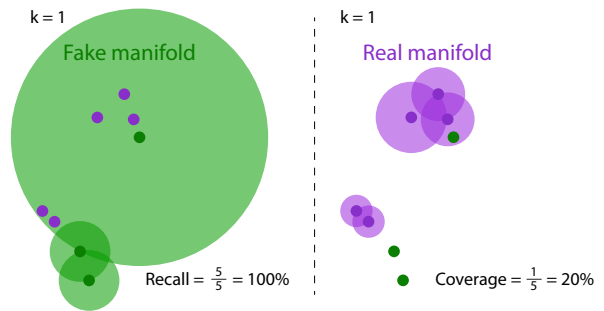
Field names were adjusted for publication.

```
{
  "fields": {
    "age": {"type": "numerical", "subtype": "integer"},
    "biological_sex": {"type": "boolean"},
    "tumour_stage": {"type": "categorical"},
    "partnership_status": {"type": "boolean"},
    "educational_level": {"type": "categorical"},
    "physical_functioning": {"type": "categorical"},
    "role_functioning": {"type": "categorical"},
    "emotional_functioning": {"type": "categorical"},
    "cognitive_functioning": {"type": "categorical"},
    "social_functioning": {"type": "categorical"},
    "global_health_status": {"type": "categorical"},
    "maladaptive_coping": {"type": "categorical"},
    "chemotherapy": {"type": "boolean"},
    "radiotherapy": {"type": "boolean"},
    "hormonotherapy": {"type": "boolean"},
    "targettedtherapy": {"type": "boolean"},
    "body_mass_index": {"type": "categorical"},
    "time_since_diagnosis": {"type": "categorical"},
    "sexual_attractiveness": {"type": "boolean"},
    "negative_body_image": {"type": "boolean"},
    "cancer_type": {"type": "categorical"}
  }
}
```

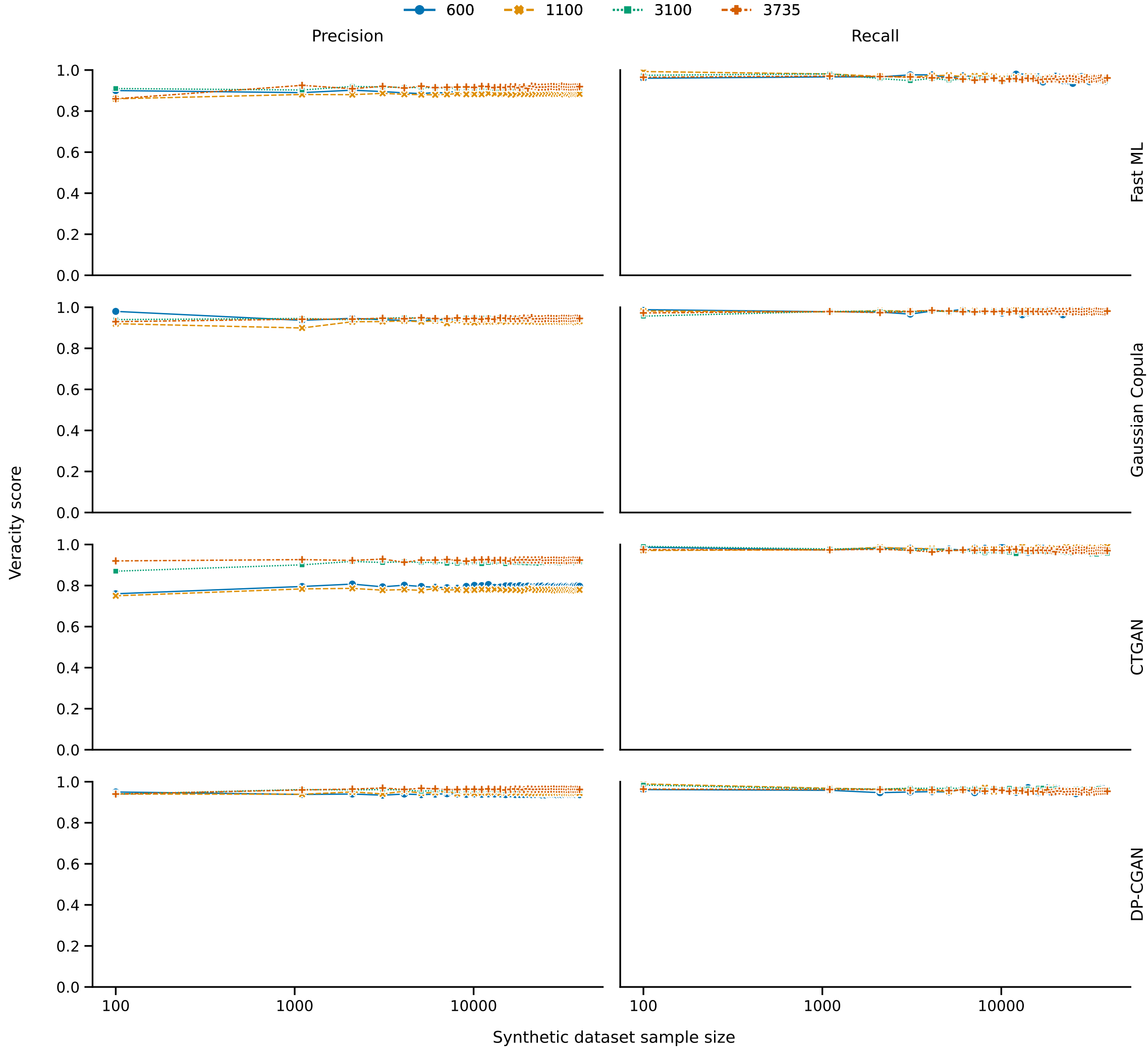
Supplementary material S4: Overview of metrics. Two example scenarios for illustrating the advantage of density over precision and coverage over recall. Note that for recall versus coverage figure, the real and fake samples are identical across left and right. [Taken from the original article (DOI: 10.5555/3524938.3525603) with consent of the corresponding author].



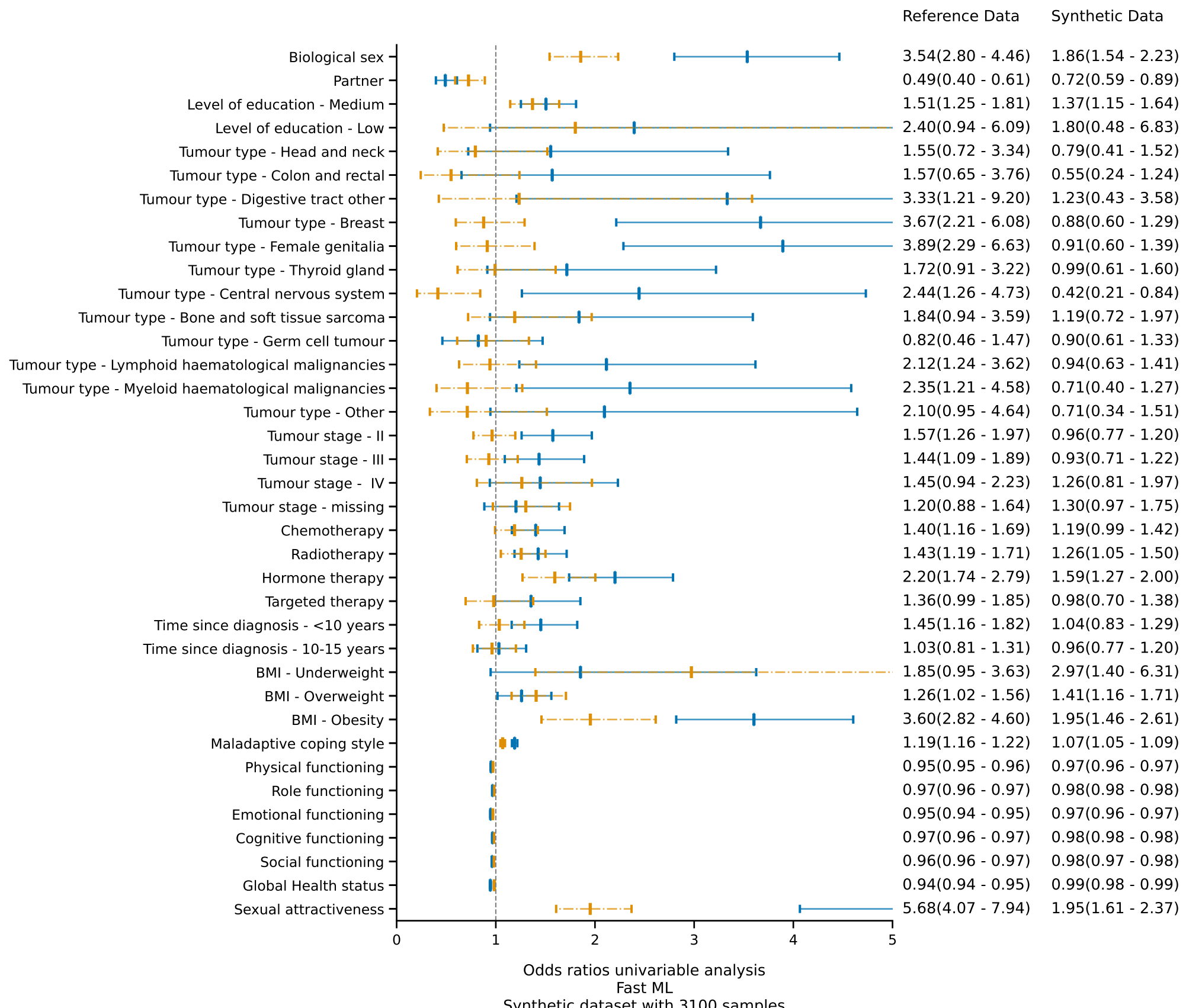
(a) Precision versus density.



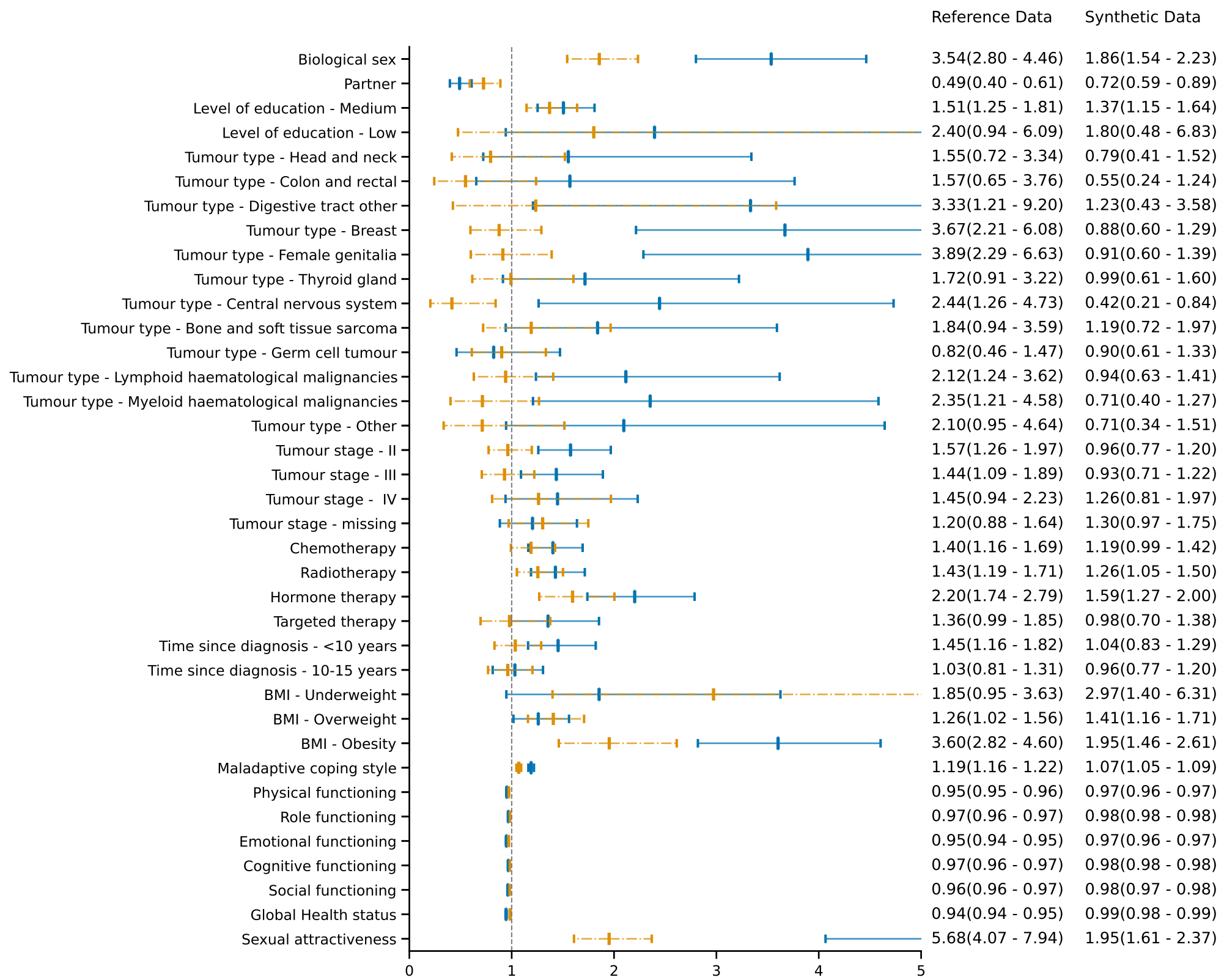
(b) Recall versus coverage.



Supplementary Material S6: Effect measure plot for the univariable logistic regression analysis on negative body image using datasets produced by Fast ML. The blue graphics represent the odds ratios and confidence intervals of the original (training) data, whilst the orange graphics represents those belonging to the synthetic data.



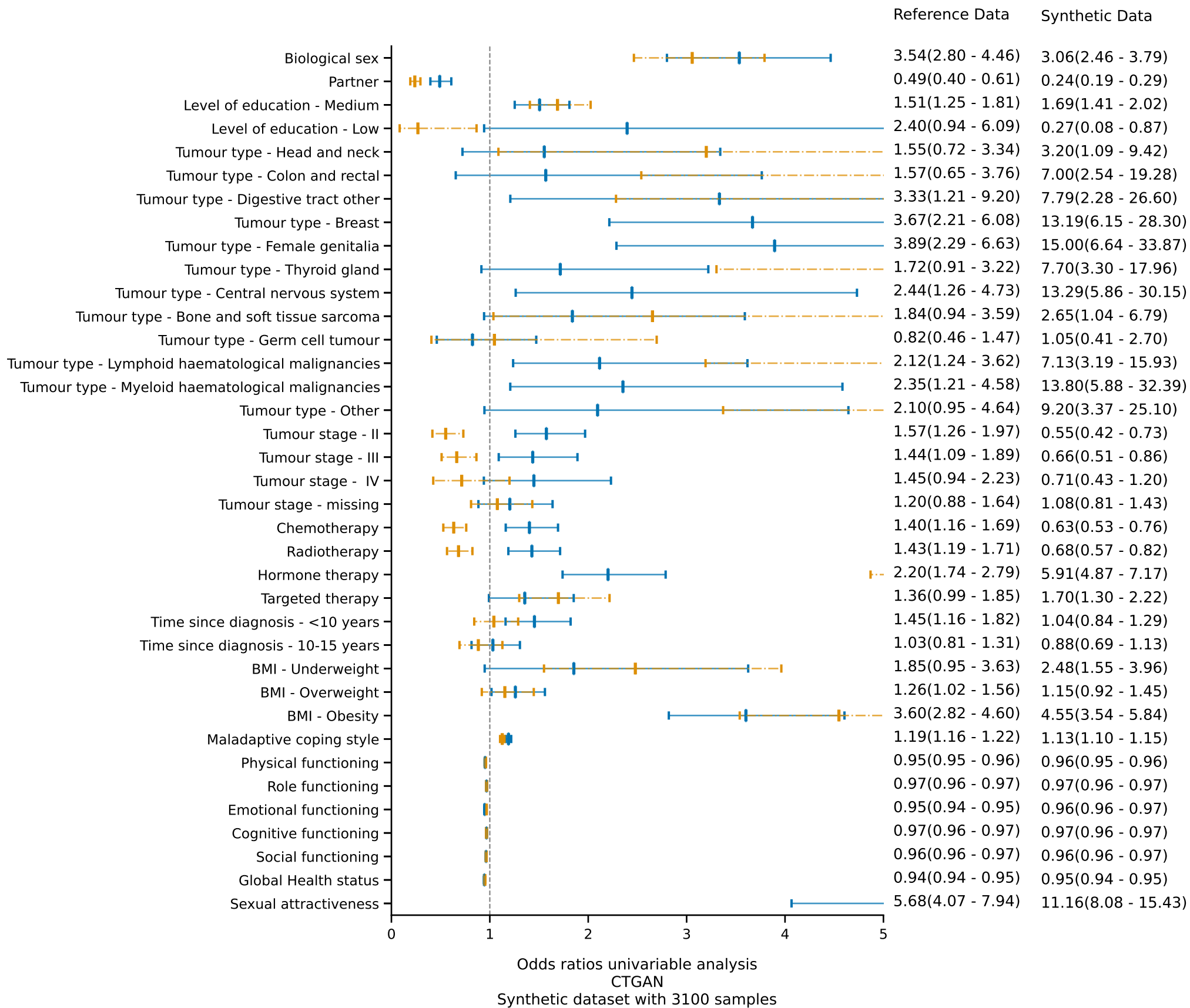
Supplementary Material S7: Effect measure plot for the univariable logistic regression analysis on negative body image using datasets produced by Gaussian Copula. The blue graphics represent the odds ratios and confidence intervals of the original (training) data, whilst the orange graphics represents those belonging to the synthetic data.



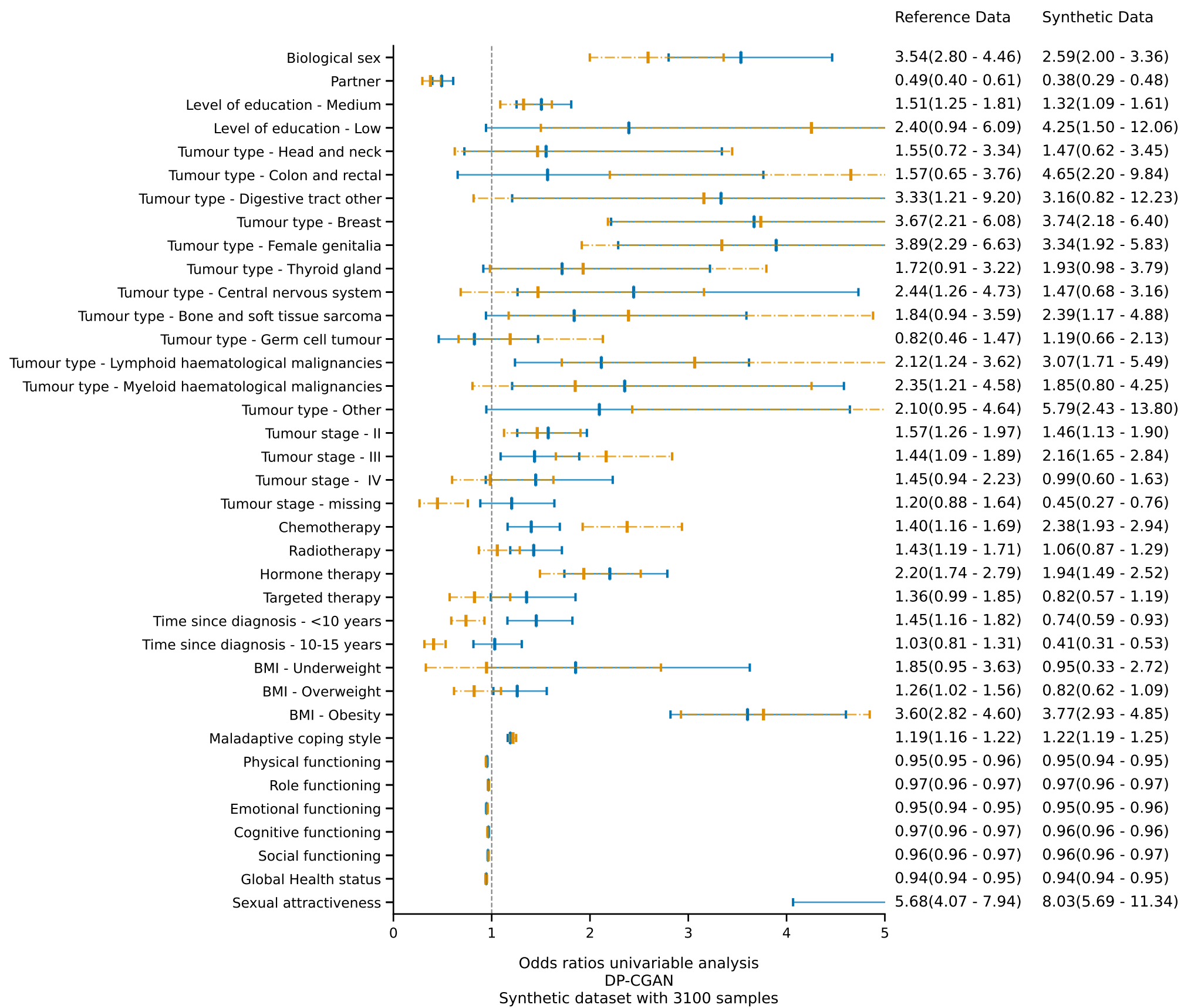
Odds ratios univariable analysis  
Gaussian Copula  
Synthetic dataset with 3100 samples



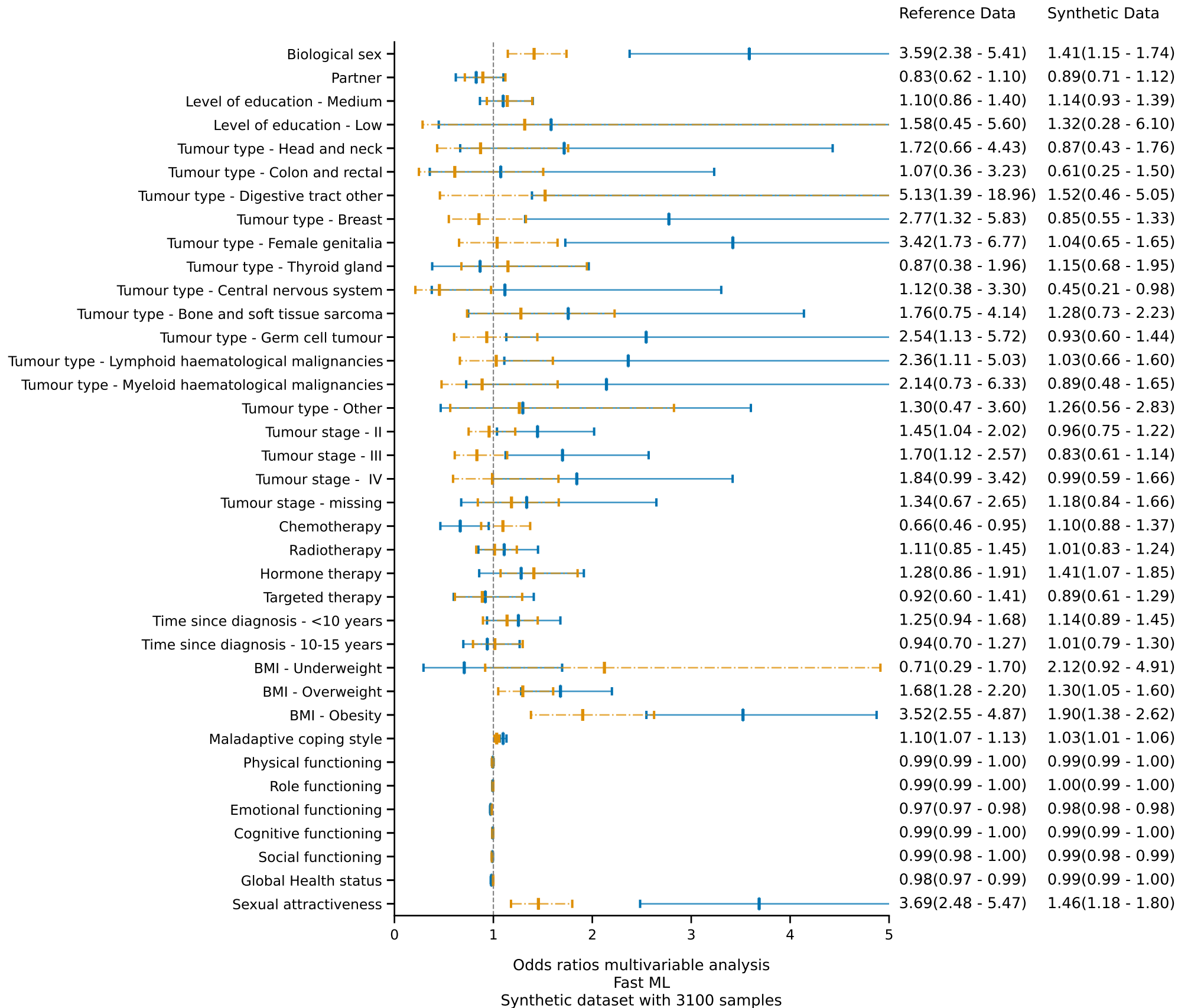
Supplementary Material S8: Effect measure plot for the univariable logistic regression analysis on negative body image using datasets produced by CTGAN. The blue graphics represent the odds ratios and confidence intervals of the original (training) data, whilst the orange graphics represents those belonging to the synthetic data.



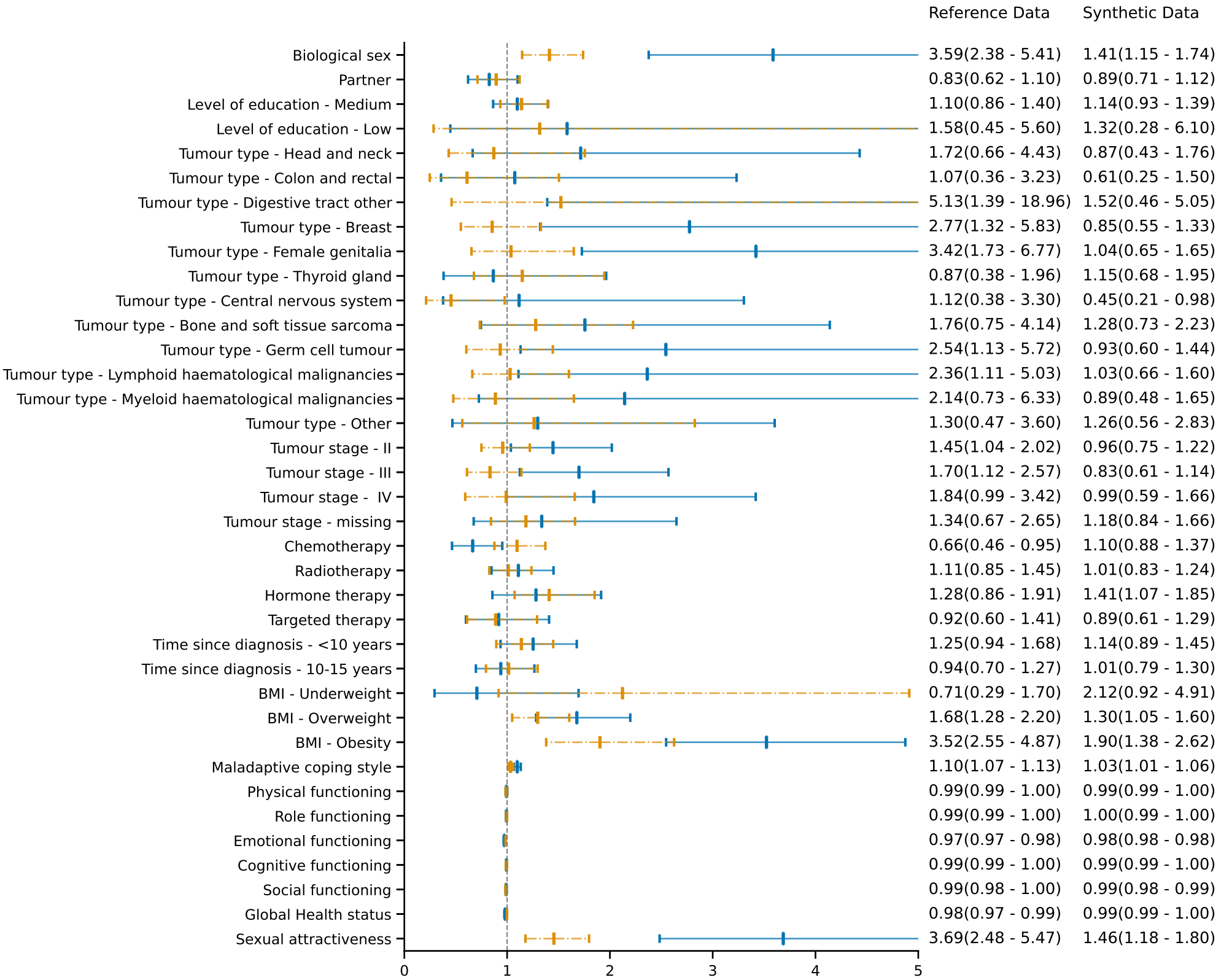
Supplementary Material S9: Effect measure plot for the univariable logistic regression analysis on negative body image using datasets produced by DP-CGAN. The blue graphics represent the odds ratios and confidence intervals of the original (training) data, whilst the orange graphics represents those belonging to the synthetic data.



Supplementary Material S10: Effect measure plot for the multivariable logistic regression analysis on negative body image using datasets produced by Fast ML. The blue graphics represent the odds ratios and confidence intervals of the original (training) data, whilst the orange graphics represents those belonging to the synthetic data.

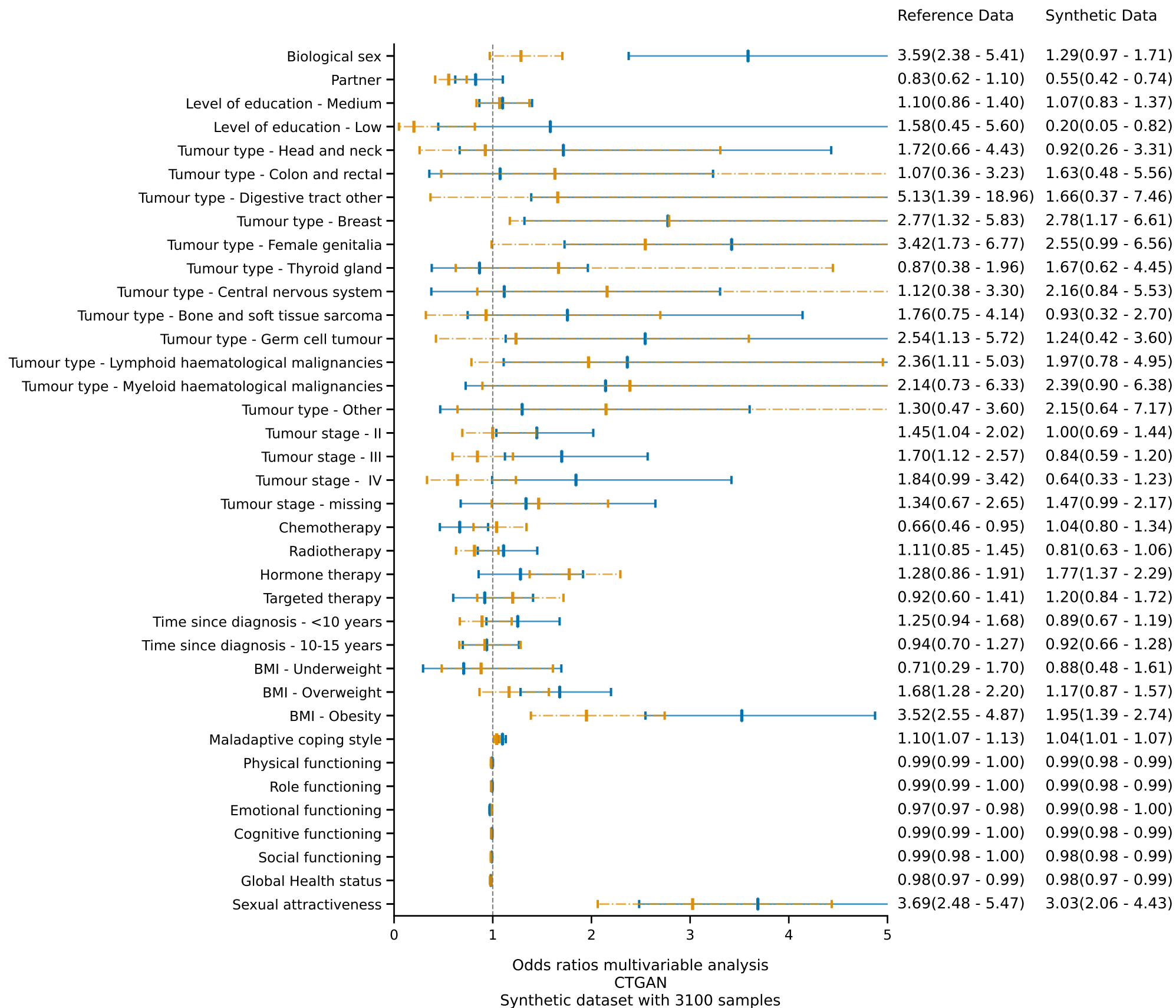


Supplementary Material S11: Effect measure plot for the multivariable logistic regression analysis on negative body image using datasets produced by Gaussian Copula. The blue graphics represent the odds ratios and confidence intervals of the original (training) data, whilst the orange graphics represents those belonging to the synthetic data

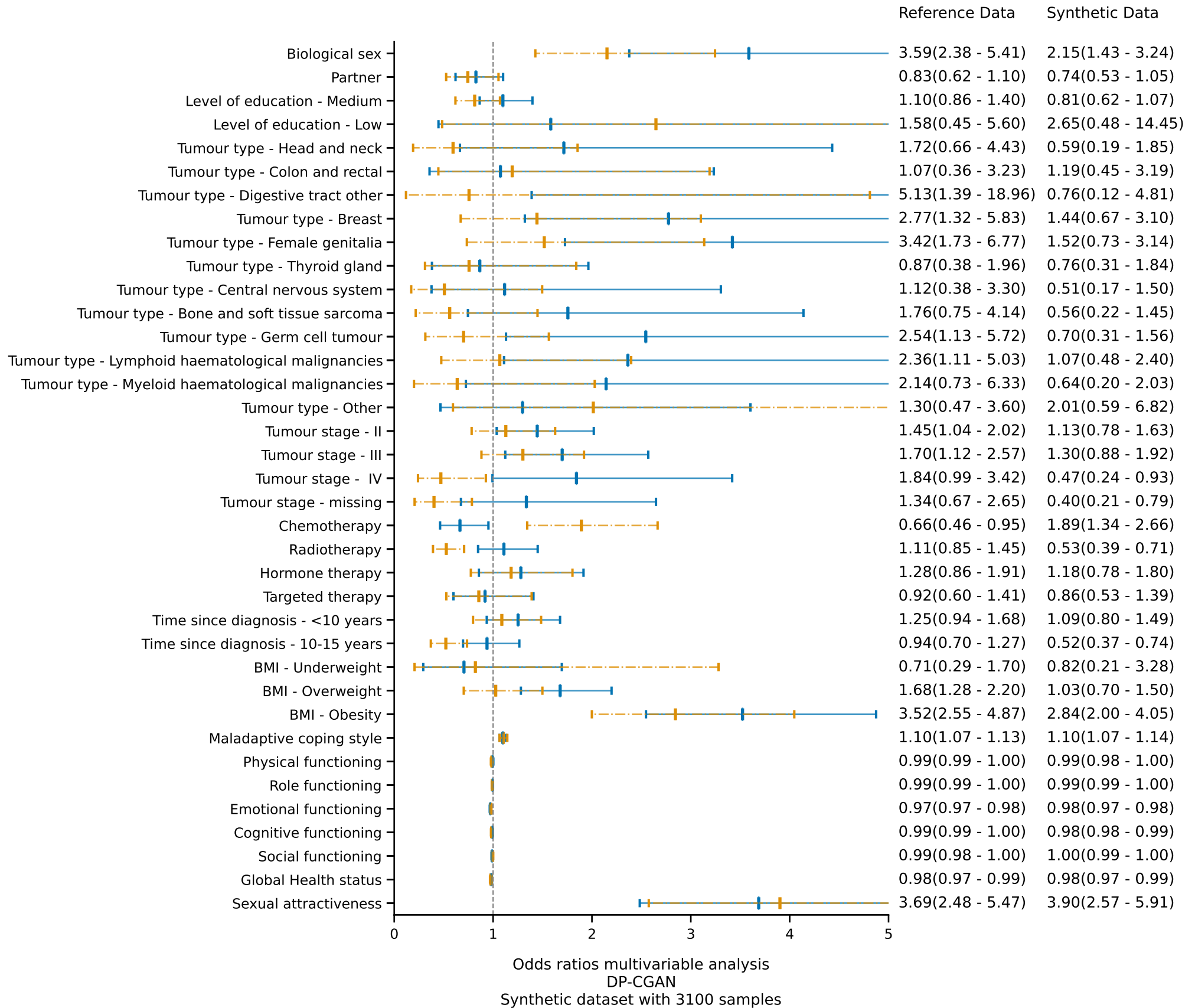


Odds ratios multivariable analysis  
Gaussian Copula  
Synthetic dataset with 3100 samples

Supplementary Material S12: Effect measure plot for the multivariable logistic regression analysis on negative body image using datasets produced by CTGAN. The blue graphics represent the odds ratios and confidence intervals of the original (training) data, whilst the orange graphics represents those belonging to the synthetic data.



Supplementary Material S13: Effect measure plot for the multivariable logistic regression analysis on negative body image using datasets produced by DP-CGAN. The blue graphics represent the odds ratios and confidence intervals of the original (training) data, whilst the orange graphics represents those belonging to the synthetic data.



Supplementary material S14: Generator and discriminator loss of a CTGAN architecture with default hyper-parameters trained on 3735 OD samples.

