Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 May 2023 doi:10.20944/preprints202305.1337.v1

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Exploratory report on data synchronising methods to develop machine learning-based prediction models for multimorbidity

Gayathri Delanerolle[1], **David Benfield[2], *Peter Phiri[1,3] **Yassine Bouchareb[12], **Kingshuk Majumder[11], **Heitor Cavalini[1] , *Jian Qing Shi[1,,8,9], *Om Kurmi[10], *Ashish Shetty[3,4], *Dharani K. Hapangama[6,7] and *Alain Zemkoho[2]

[1] Research & Innovation Department, Southern Health NHS Foundation Trust, Southampton, United Kingdom
[2]School of Mathematics, University of Southampton, United Kingdom
[3] School of Psychology, Faculty of Environmental and Life Sciences, University of Southampton, United Kingdom
[4]University College London, United Kingdom
[5]University College London Foundation Trust, United Kingdom
[6]University of Liverpool, Liverpool, United Kingdom
[7]Liverpool Women's Hospital, Liverpool, United Kingdom
[8]Southern University of Science and Technology, Shenzhen, 518055, China
[9]National Center for Applied Mathematics Shenzhen, 518000, China
[10]University of Coventry
[11]University of Manchester Foundation Hospitals
[12]Sultan Qaboos University, College of Medicine and Health Sciences, Muscat, Oman

Corresponding author: peter.phiri@southernhealth.nhs.uk

**Shared second author
*Shared last author

Abstract : Endometriosis is a complex chronic condition characteristic of chronic pelvic pain, dysmenorrhea, anxiety and fatigue. This can often lead to multimorbidity which is defined by the presence of two or more long term conditions. Delayed diagnosis of endometriosis is a crucial issue that leads to poor quality of life and clinical management. There are a variety of limitations linked to conducting endometriosis research including lack of dedicated funding. Additionally, accessing existing electronic healthcare records can be challenging due to governance and regulatory restrictions. Missing data issues are another concern that has been commonly identified among real-world studies. Considering these challenges, data science technique could provide a solution by way of using synthetic datasets that could be generated using known characteristics of endometriosis to explore the possibility of predicting multimorbidity. This study aimed to develop an exploratory machine learning model that can predict multimorbidity among women with endometriosis using real-world and synthetic data. A sample size of 1012 was used from two endometriosis specialized centres in the UK. In addition, 1000 synthetic data records per centre were generated using the widely used Synthetic Data Vault's Gaussian Copula model based on patients' records' characteristics. Three standard classification models, Logistic Regression (LR), Support Vector Machine (SVM), and Random Forest (RF), were used for classification. The average accuracies for all three models (LR, SVM and RF), given as "model accuracy-centre1: accuracy-centre2" were found to be: LR 64.26%:69.04%, SVM 67.35%:68.61%, and RF 58.67%:73.76% on real-world data, and LR 69.9%:72.29%, SVM 69.39%:70.13, and RF 68.88%:74.62 on synthetic data, respectively. The findings of this report show machine learning models trained on synthetic data performed better than models trained on real-world data. Our findings suggest synthetic data holds great promise for shows value to conduct clinical epidemiology and clinical trials that could devise better precision treatments and possibly reduce the burden of multimorbidity.

© 2023 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 May 2023 doi:10.20944/preprints202305.1337.v1

Background

Data science is a rapidly evolving research field that influences analytics, research methods, clinical practice and policies. Access to comprehensive real-world data and gathering life-course research data are primary challenges observed in many disease areas. Existing real-world data can be a rich source of information required to better characterise diseases, generate cohort specifications and understand clinical practice gaps to conduct more precision research that is value-based for healthcare systems. A common challenge linked to real-world and research data is a high rate of missingness. Historically, statistical methods were used to address missing data where possible, but advances in artificial intelligence techniques have provided improved and quicker methods for use. These methods could also be used for predicting disease outcomes, improving diagnostic accuracy and treatment suitability.

These methods can be particularly useful for women's health conditions, where the complex physical and mental health symptoms can give rise to insufficient understanding of disease pathophysiology and phenotype characteristics that play a vital role in diagnosis, treatment adherence and prevention of secondary or tertiary conditions. One such condition is endometriosis. Endometriosis is complex with an array of physical and psychological symptomatologies, often leading to multimorbidity [1]. Multimorbidity is defined by the presence of two or more conditions in any given individual and therefore could be prevented if the initial conditions are managed more effectively. The incidence of multimorbidity has increased with a rising ageing population, burden of non-communicable diseases in general and mental ill health which, is particularly important for women [2]. Another important aspect of multimorbidity is *disease sequalae,* where a physical manifestation could correlate with a mental health impact, and vice versa. The precise causation is complex to assess due to limitations in the current understanding of disease sequalae pathophysiology [3]. As such, multimorbidity could be deemed highly heterogeneous. Multimorbidity impacts people of all ages, although current evidence suggests it is more common among women than men, even though previously, multimorbidity was thought to have been more common in older adults with a high frailty index score [4]. Hence, multimorbidity is challenging to treat, and there remains a paucity of research available to better understand the basic science behind the complex mechanisms that could enable better diagnosis and management long-term [4].

This undercurrent of disease complexities linked to endometriosis that could lead to multimorbidity should be explored to support clinicians and healthcare organisations in future-proofing patient care [5]. In line with this, exploring machine learning as a technique in conjunction with synthetic data methods could demonstrate better predictions and offer a new solution to sample size challenges.

Methods

Our primary aim of the study was to develop an exploratory machine learning model that can predict multimorbidity among endometriosis women using both real-world and synthetic data. In certain instances, real-world data may present confidentiality issues, particularly in medical research where data often contains personal and sensitive information. Sharing such data for analysis can expose vulnerabilities. To develop these models, existing knowledge and symptomatology, comorbidities and demographic data were used. Anonymised data from an ethically approved study was provided from Manchester and Liverpool Endometriosis specialist centres in the UK. The data records used included symptoms, diseases, and conditions in women with a confirmed diagnosis of endometriosis. Data curation was completed for the entire sample size using the following steps;

1. Data pre-processing: the data was cleaned and prepared to manage missing values, encoding categorical variables, and standardizing or normalizing continuous variables.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 May 2023 doi:10.20944/preprints202305.1337.v1

2. Synthetic data generation: the synthetic data records were generated for each centre using a widely used synthetic Data Vault's Gaussian Copula model, based on the data characteristics from patients' records.
3. Model development: trained and implemented three standard classification models - Logistic Regression (LR), Support Vector Machine (SVM), and Random Forest (RF) - on both real-world and synthetic data. These models were used to predict multimorbidity among women with endometriosis.
4. Model evaluation: models were assessed the performance of the models by comparing their average accuracies on real-world and synthetic data. Multiple metrics' of accuracy, precision, recall, and F1-score were used to evaluate the models' performances.
5. Comparison and analysis: the results of the models trained on real-world data and synthetic data to determine if synthetic data could serve as a viable alternative for real-world data in predicting multimorbidity among women with endometriosis.

For all experiments, we trained one model on real-world data, and another on synthetic data. Both models were tested on the same test set which contains only real-world data because the overall population's true distribution for endometriosis is verified. The accuracies of these models can then provide better insight into whether the use of synthetic data affects the performance of machine learning models.

Ethics approval

Anonymous data used in this study was approved by the North of Scotland Research Ethics Committee 2 (LREC: 17/NS/0070) for the RLS study conducted at the University of Liverpool.

The model used age, height, symptoms, commodities and weight in a mathematical formulation. Let x_i be the vector containing these recordings for the i^{th} person and let $x = (x_1, \dots, x_n)$ be the matrix containing the data about all n people. As part of developing methodological rigour, we considered a working example to predict whether each person in the sample develops depression. Let $y = (y_1, \dots, y_n)$ be the vector of response variables where:

$$y_i = \begin{cases} 1 & \text{if patient } i \text{ develops a depression} \\ 0 & \text{if patient } i \text{ does not develop depression.} \end{cases}$$

In this example, s we collect data for $n = 3$ people and have $p = 3$ recordings for each person i, (i.e., age, height and weight). These are represented by x_{i1}, x_{i2} and x_{i3} respectively. The data can be summarised in Table 1 as follows:

PERSON #	AGE	HEIGHT (M)	WEIGHT (KG)	DEPRESSION
1	67	1.9	65	1
2	43	1.2	75	0
3	23	1.5	43	0

Table 1. Example Dataset For Predicting Depression.

We created a function, f_β with parameters β, that takes the age, height, and weight (x_{i1}, x_{i2}, x_{i3}) of the person i, as input and outputs a prediction of whether they will develop depression. Let y_i^* be the prediction of whether person i develops depression, then we say that

$$y_i^* = f_\beta(x_i).$$

The performance of parameters β can be tested through a loss function, defined as $\mathcal{L}(\beta)$ which measures the difference between the true values of y and the predictions, $y^* = (y_1^*, \dots, y_n^*)$. The loss function imposes a penalty when incorrect predictions are made. Hence, to find the best β, we solve the optimisation problem:

$$\beta^* = \underset{\beta}{\text{argmin}}\, \mathcal{L}(\beta, y, y^*).$$

The function f_{β^*} can then be used to make predictions for patients who haven't been tested for depression.

An initial observation was that our prediction function could become over-fitted to the data. This meant that the function captured the specific distribution between x and y very well, but if this data was not in a structured format representing the true distribution between symptoms and comorbidities, the prediction function would not be generalisable to other types of data.

The performance of the prediction function on unseen data can be estimated by separating the data into a training set, $(x^{\text{train}}, y^{\text{train}})$ and test set, $(x^{\text{test}}, y^{\text{test}})$. The optimal parameters are found using the training set and then the model's accuracy is tested on the test set. This accuracy is measured by the proportion of correctly classified data. This is measured by a confusion matrix, which records the frequencies of each possible outcome. Let C be the confusion matrix defined as:

$$C = \begin{pmatrix} C_{00} & C_{01} \\ C_{10} & C_{11} \end{pmatrix} \qquad \textit{(Equation 1)}$$

where C_{ij} is the number of times $y^{\text{test}} = i$ while $y^{\text{test}^*} = j$. The accuracy of our model is then

$$\text{Accuracy}(\%) = \frac{C_{00} + C_{11}}{C_{00} + C_{01} + C_{10} + C_{11}}. \qquad \textit{(Equation2)}$$

To summarise, the approach is broken down into the following three steps,
Solve optimisation problem

$$\beta^* = \underset{\beta}{\text{argmin}}\, \mathcal{L}\left(\beta, y^{\text{train}}, y^{\text{train}^*}\right)$$

on the training set, where the set of prediction values, y^{train^*}, is found by

$$y^{\text{train}^*} = f_\beta(x^{\text{train}}).$$

1) Make predictions on the test set using optimal weights β^*

$$y^{\text{test}^*} = f_{\beta^*}(x^{\text{test}}).$$

2) Construct confusion matrix, C as is defined in (1) and find the accuracy of the model on unseen data by equation (2).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 May 2023 doi:10.20944/preprints202305.1337.v1

Data preparation – Manchester

In the Manchester dataset, for each patient, the presence of various symptoms and multiple diagnoses among women with Endometriosis are recorded. These are summarised, with descriptions in Table 2. A total of $p = 15$ recordings are made for each person, and so we define $x_i = (x_{i1}, \dots, x_{ip})$ to be the vector containing the recordings for person i.

Feature	Data Type	Description
Age	Integer	Age of the Patient
Menorrhagia	Binary	Whether or not the patient has been diagnosed with menorrhagia
Dysmenorrhea	Binary	Whether or not the patient has been diagnosed with dysmenorrhea
Non menstrual Pelvic pain	Binary	Whether or not the patient experiences non-menstrual pelvic pain
Dysphasia	Binary	Whether or not the patient experiences dysphasia
Dyspareunia	Binary	Whether or not the patient experiences dyspareunia
other symptoms	Binary	Whether or not the patient has any other symptoms besides the ones recorded in other features
Infertility	Binary	Whether or not the patient is infertile
No of Endo symptoms	Binary	Whether or not the patient has more than 1 symptom
Year of diagnosis	Date	The year of the patient's diagnosis of endometriosis
Other surgery – Not related to endometriosis	Binary	Whether or not the patient received any surgeries not related to endometriosis
Discharged	Binary	Whether or not the patient was discharged
follow up	Binary	Follow up clinical appointments
Hormonal treatment Currently	Binary	Whether or not the patient is taking any hormonal treatment
No of hormonal treatment tried	Integer	The number of hormonal treatments the patient is taking

Table 2. Manchester Data Feature Variables.

Additionally, for each individual , three response variables are documented, which are summarised, along with their descriptions, in Table 3. These variables are defined as follows:

$$y_i^M = \begin{cases} 1 & \text{if patient } i \text{ develops a mental health condition} \\ 0 & \text{if patient } i \text{ does not develop any mental health condition} \end{cases},$$

$$y_i^I = \begin{cases} 1 & \text{if patient } i \text{ develops irritable bowel syndrome} \\ 0 & \text{if patient } i \text{ does not develop irritable bowel syndrome} \end{cases},$$

$$y_i^C = \begin{cases} 1 & \text{if patient } i \text{ develops at least one of various other comorbidities} \\ 0 & \text{if patient } i \text{ does not develop at least one of various other comorbidities} \end{cases}.$$

Variable	Name	Description
$\boldsymbol{y^M}$	Mental Health	The presence of at least one of various mental health conditions
$\boldsymbol{y^I}$	IBS	The presence of irritable bowel syndrome (IBS)
$\boldsymbol{y^C}$	Comorbidities (Other)	The presence of at least one other disease (Perhaps we have a list of these?).
$\boldsymbol{y^{\text{Comb}}}$	Combined	The presence of at least one of the above conditions.

Table 3. Manchester Data Response Variables.

We examined three models of fit, one for each response variable. We defined a fourth response variable, "Combined", as shown in the final row of table 3, which indicates the presence of at least one of the other three conditions. Formally, y^{Comb} is defined as:

$$y_i^{\text{Comb}} = \begin{cases} 1 & \text{if patient } i \text{ develops at least one of any of the conditions} \\ 0 & \text{if patient } i \text{ does not develop at least one of any of the conditions} \end{cases}.$$

We fitted a fourth model for this response variable.

We converted the binary variables, including our response variables of "Yes" and "No" to 1 and 0, respectively. There was no missing data in the Manchester dataset, and as such, we make use of all $n = 99$ observations.

Data preparation – Liverpool

The data from Liverpool had a sample size of 913 patients. The raw data defined 68 possible different symptoms which was considered as feature variables. A significant rate of missing data was identified. The complete list of features along with their percentage missing values can be found in Table 4.

To prepare the data, we first filtered by "Endometriosis = TRUE" which included patients with a diagnosis of endometriosis leaving a sample of 339 patients. Then we removed features with more than 1% of missing values with 29 final features. The feature "Endometriosis" is a binary identifier, which, after filtering, is always true, so we dropped this feature too. The final 27 features are summarised, with descriptions, in Table 5.

Feature	NaN (%)	Feature	NaN (%)	Feature	NaN (%)	Feature	NaN (%)
Sample ID	0.0	Age at diagnosis	98.5	Pain interferes with daily activities	0.0	Hormones	0.0
Age	0.1	Endometriosis symptoms	97.8	Dysmenorrhoea score	97.5	Other information	28.6
Ethnicity	96.7	Endometriosis stage	70.2	Non-menstrual pelvic pain	0.0	Previous ablation	0.0
Postcode	94.4	VAS	91.5	Analgesia for pain	0.0	Medications	85.9
Sample type	2.8	FH ENDO	98.1	Pain prevents daily activities	0.0	Endometrial cancer	0.0

Hair colour	96.7	Adenomyosis	0.0	Pelvic pain score	97.4	Metastatic lesion	0.0
Eye colour	96.7	Menorrhagia	0.0	Miscarriages	44.5	Metastatic lesion location	100.0
Height (m)	0.1	Fibroids	0.0	Polycystic ovary syndrome	0.0	Type of cancer	99.8
Weight (kg)	0.4	Reseason for surgery	18.7	Irregular cycles	0.0	Cancer comments	98.7
BMI	0.0	Previous history	84.7	Cu coil	0.0	Grade	100.0
Smoker	0.0	Gravidity	97.3	Menarche	97.2	Stage	99.8
Pack years	99.1	Parity	8.3	LMP	15.7	Pathology findings	99.8
Exercise	97.4	Deliveries	96.8	Menopause	100.0	Cancer staging	0.0
Alcohol	0.0	Infertility	0.0	Post-menopause	0.0	Dating by histology	64.3
Drinks per week	98.5	Dyspareunia	0.0	Cycle length	17.4	Hormonal dating	99.8
Endometriosis	0.0	Dysmenorrhoea	0.0	Days of bleeding	18.4	Agreement of date	0.0
Age first symptoms	98.6	Analgesia	0.0	Contraceptive/hormone treatment	59.9	Comments	70.1

Table 4. Liverpool Data Percentage Missing Data.

Feature	Data Type	Description
Age	Integer	Age of patient
Height (m)	Real	Height of patient in meters
Weight (kg)	Real	Weight of patient in kilograms
BMI	Real	BMI of patient
Smoker	Binary	Whether of not the patient smokes
Alcohol	Binary	Whether or not the patient consumes alcohol
Adenomyosis	Binary	Whether or not the patient has been diagnosed with Adenomyosis
Menorrhagia	Binary	Whether or not the patient has been diagnosed with Menorrhagia
Fibroids	Binary	Whether or not the patient has been diagnosed with Fibroids
Infertility	Binary	Whether or not the patient is infertile
Dyspareunia	Binary	Whether or not the patient has been diagnosed with Dyspareunia
Dysmenorrhoea	Binary	Whether or not the patient has been diagnosed with Dysmenorrhoea
Analgesia	Binary	Whether or not the patient takes analgesia
Pain interferes with daily activities	Binary	Whether or not the patient experiences pain with daily activities

Non-menstrual pelvic pain	Binary	Whether or not the patient experiences non-menstrual pelvic pain
Analgesia for pain	Binary	Whether or not the patient takes analgesia to relieve pain
Pain prevents daily activities	Binary	Whether or not the patient says that pain prevents them from performing daily activities
PCOS	Binary	Whether or not the patient has polycystic ovary syndrome
Irregular cycles	Binary	Whether or not the patient experiences irregular menstrual cycles
Cu coil	Binary	Whether the patient has ever had a CU coil
Post-menopausal	Binary	Whether or not the patient has had menopause
Hormones	Binary	Whether or not the patient is taking any hormonal replacement treatments
Previous ablation	Binary	Whether the patient has had a previous ablation
Endometrial cancer	Binary	Whether or the patient have or had endometrial cancer
Metastatic lesion	Binary	Whether or not the patient had any cancerous lesions
Cancer staging agreement with Pathology	Binary	Whether or not the patient had an existing involvement within the cancer pathway
Agreement of staging	Binary	Whether or not the patient had a staging agreement

Table 5. Liverpool Data Features with Less than 1% Missing Data.

A total of 4 patients were identified with missing values and subsequently removed them from the dataset, resulting in a final sample size of $n = 310$ patients. We selected two diseases as our response variables for prediction (Table 6). Given our ultimate objective of predicting multimorbidity in patients, we constructed a final response variable, "Combined", as a binary variable representing the presence of at least one of the other two response variables, akin to the data from Manchester. Their formal definitions of these response variables are as follows:

$$y_i^A = \begin{cases} 1 & \text{if patient } i \text{ develops Adenomyosis} \\ 0 & \text{if patient } i \text{ does not develop Adenomyosis} \end{cases},$$

$$y_i^I = \begin{cases} 1 & \text{if patient } i \text{ develops Menorrhagia} \\ 0 & \text{if patient } i \text{ does not develop Menorrhagia} \end{cases},$$

$$y_i^C = \begin{cases} 1 & \text{if patient } i \text{ develops at least one of any of the conditions} \\ 0 & \text{if patient } i \text{ does not develop at least one of any of the conditions} \end{cases}.$$

Variable	Name	Description
$\boldsymbol{y^A}$	Adenomyosis	Whether the patient has been diagnosed with Adenomyosis
$\boldsymbol{y^M}$	Menorrhagia	Whether the patient has been diagnosed with Menorrhagia
$\boldsymbol{y^{\mathrm{Comb}}}$	Combined	The presence of at least one of the above conditions.

Table 6. Liverpool Data – Response Variables.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 May 2023 doi:10.20944/preprints202305.1337.v1

Synthetic Data

To address this concern, we employed the Synthetic Data Vault (SDV) package in Python to create synthetic data as a substitute and assessed its similarity to the real data. By leveraging other sampling techniques, such as random simulation, the synthetic data could generate a dataset with an expanded sample size that more accurately represents the entire population.

During our data preparation, we eliminated numerous observations due to missing data. Synthetic data could potentially serve as a replacement for these missing values. However, in our analysis, we opted to generate entirely new observations, rather than filling in the gaps.

We utilised SDV's Gaussian Copula model, which constructs a distribution over the unit cube $[0,1]^p$ from a multivariate normal distribution over R^p by using the probability integral transform. The Gaussian Copula characterises the joint distribution of the random variables representing each feature by analysing the dependencies between their marginal distributions. Once the model is fitted to our data, it can be used to sample additional instances of data.

Manchester Data

We initiated our analysis with the Manchester data, and after fitting the Gaussian Copula to our 99 samples, we generated an additional 1000 samples.

By employing SDV's SD Metrics library, we were able to evaluate the similarity between the real and synthetic data. We examined how closely the synthetic data relates to the real data in order to determine whether we have adequately captured the true distribution. This assessment involved comparing the distribution similarities across each feature, and we adopted two approaches for this evaluation.

Initially, we measured the similarities across each feature by comparing the shapes of their frequency plots, as illustrated in Figure 1. This comparison was conducted based on the "age" distribution for both the real and synthetic data.

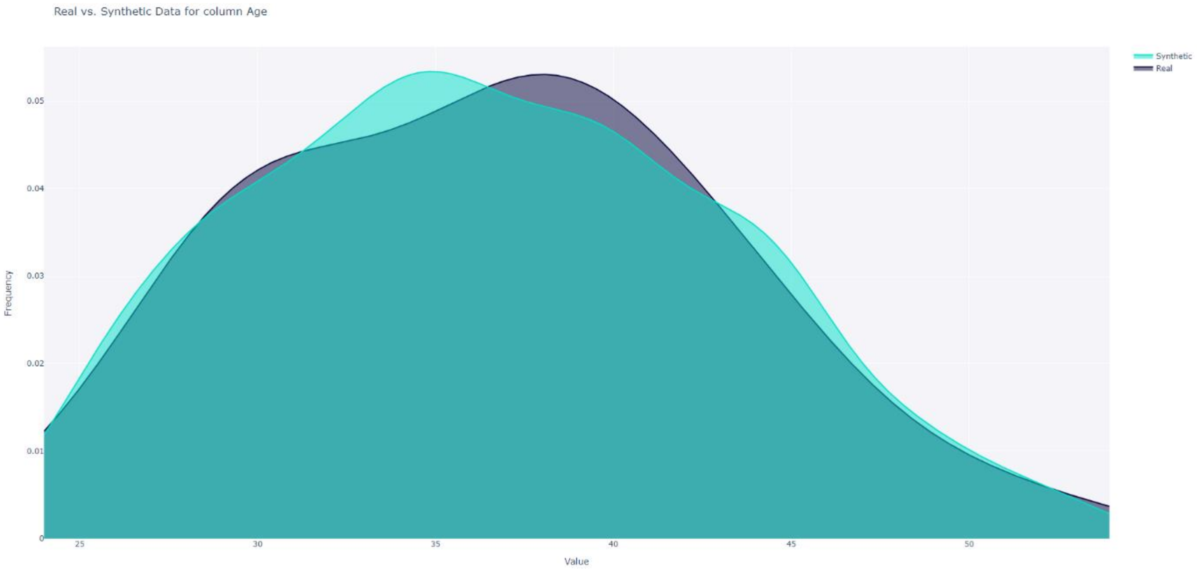

Figure 1. Age distribution shape comparison.

For numerical data, SDV computed the Kolmogorov-Smirnov (KS) statistic, which represents the maximum difference between the cumulative distribution functions. The value of this distance ranges between 0 and 1, with SDV converting it to a score by:

$$\text{Score} = 1 - \text{KS-statistic}.$$

For Boolean data, SDV calculates the Total Variation Distance (TVD) between the real and synthetic data. We determined the frequency of each category value and represented it as a probability. The TVD statistic compares the differences in probabilities, as given by:

$$\delta(R,S) = \frac{1}{2}\sum_{\omega\in\Omega} |R_\omega - S_\omega|$$

where Ω represents the set of possible categories, and R_ω and S_ω are the frequencies of category ω in the real and synthetic datasets respectively. The similarity score is then given by:

$$\text{score} = 1 - \delta(R,S).$$

The score for each feature is summarised in Figure 2, and we obtained an average similarity score of 0.92.

For the second measure of similarity, we constructed a heatmap to compare the distribution across all possible combinations of categorical data. This was accomplished by calculating a score for each combination of categories. To initiate this process, two normalised contingency tables were constructed; one for the real-world data and one for the synthetic data. Let α and β be two features, the contingency tables describe the proportion of rows that have each combination of categories in α and β, thereby illustrating the joint distributions of these categories across the two datasets.

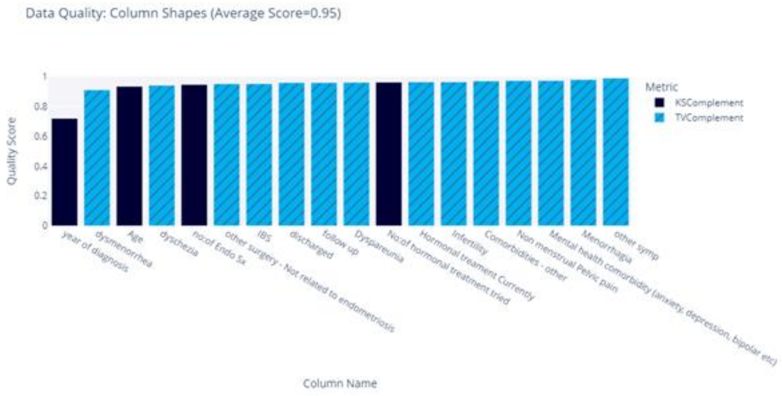

Figure 2. Feature Distribution Shape Comparison.

To compare the distributions, SDV computes the difference between the contingency tables using Total Variation Distance. This distance is subsequently subtracted from 1, implying that a higher score denotes greater similarity. Let A and B represent the set of categories in features α and β respectively; the score between features α and β are calculated as follows:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 May 2023 doi:10.20944/preprints202305.1337.v1

$$score = 1 - \frac{1}{2}\sum_{a\in A}\sum_{b\in B}|S_{a,b} - R_{a,b}|, \qquad \text{(Equation 3)}$$

where $S_{a,b}$ and $R_{a,b}$ represent the proportions of categories a and b occurring simultaneously, as derived from the contingency tables for the synthetic and real data, respectively. It is important to note that we did not employ a measure of association between features, such as Cramer's V, since it does not measure the direction of the bias and may consequently yield misleading results.

A score of 1 indicates that the contingency table was identical between the two datasets, while a score of 0 indicates that the two datasets were as dissimilar as possible. These scores for all combinations of features are depicted as a heatmap (Figure 3). It is worth noting that continuous features, such as "Age", were discretized in utilise Equation (3) in determining a score.

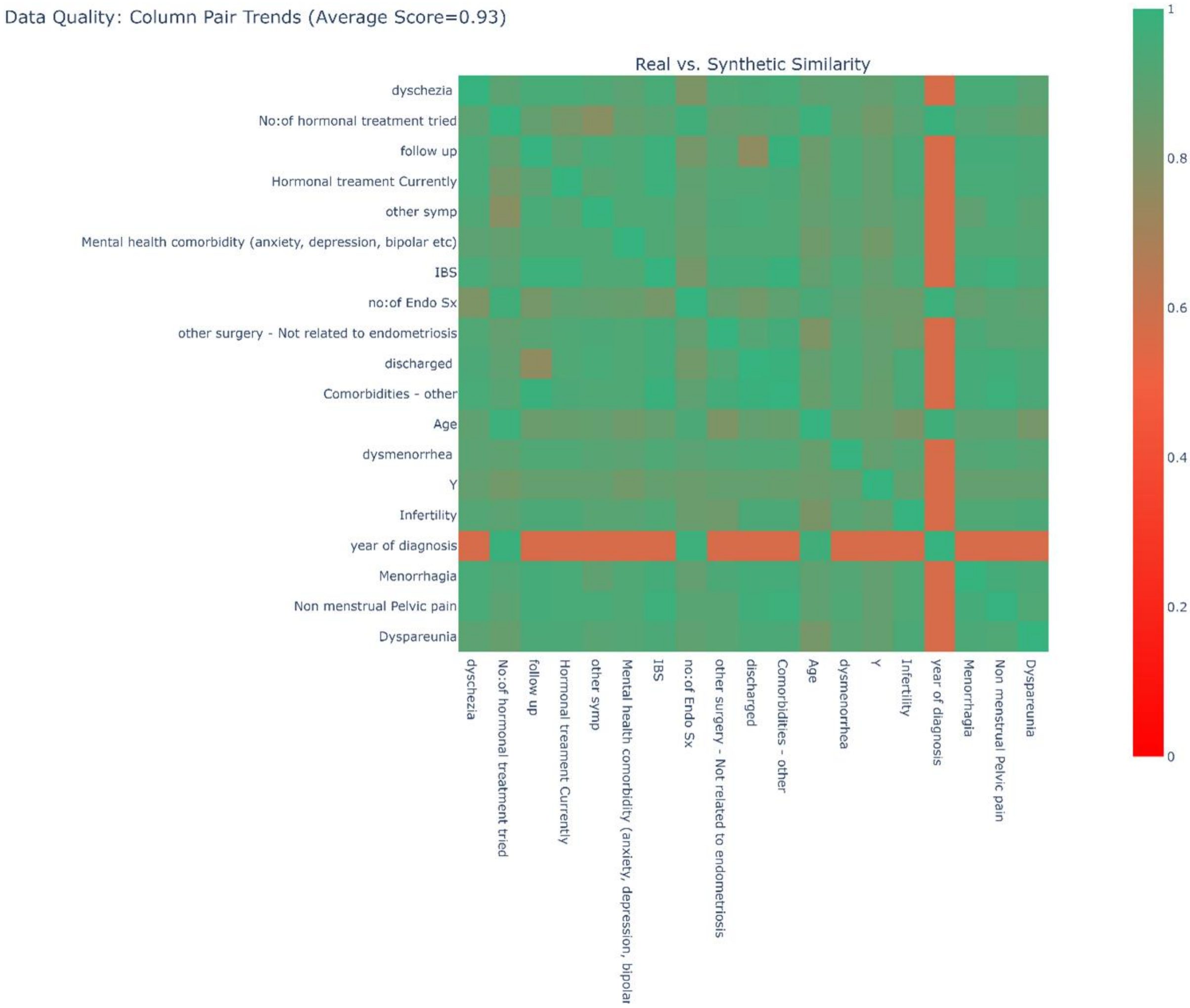

Figure 3. Distribution Comparison Heatmap.

The heatmap suggests that most features exhibit a strikingly similar distribution across the two datasets, with the exception for "Year of Diagnosis". This discrepancy could potentially be attributed to the feature's inherent nature as a date, despite being treated as an integer in the model. This issue merits further investigation.

Based on these metrics, we confidently concluded, that the new data closely adhered to the distribution of the original data.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 May 2023 doi:10.20944/preprints202305.1337.v1

Liverpool Data

To generate synthetic data, we adhered to the same procedure as with the Manchester data. We produced 1000 additional samples from a Gaussian copula fitted to the 311 real samples and combined them with the real data to create a new dataset. Using contingency tables, we developed a heatmap by applying the formula in Equation (3) to generate scores; this heatmap is displayed in Figure 4. A score of 1 implies that the contingency table was identical between the two datasets, whereas a score of 0 indicates that the two datasets were as distinct as possible. Our analysis revealed an average similarity of 0.94.

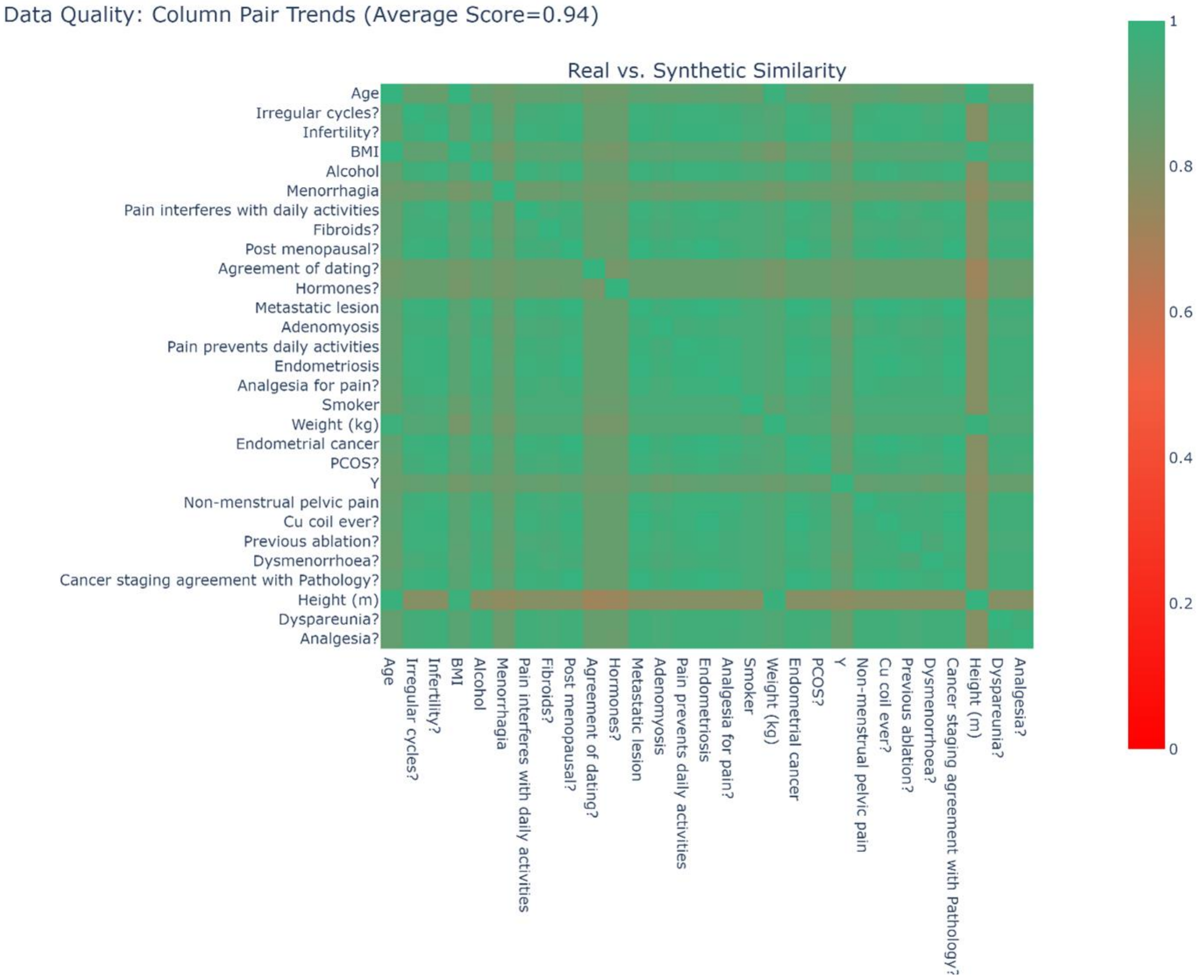

Figure 4. Real Vs Synthetic Data Distribution Heatmap (Liverpool Data).

We compared the shape of the distributions for each feature; for instance, the distributions for the "Height" feature are illustrated in Figure 5. We observed that the distributions were dissimilar. To calculate similarity scores, we employed the KS statistic for numerical features and Total Variation Distance for Boolean features. These scores are summarised in Figure 6. We found that the distributions of "Height" and "Weight" were not similar; however, the distributions of the remaining features exhibited similarity. With an average similarity of 0.75, we concluded that the data distributions were, on average similar. The distributions of all categorical features were accurately captured, but two of the continuous features were not.

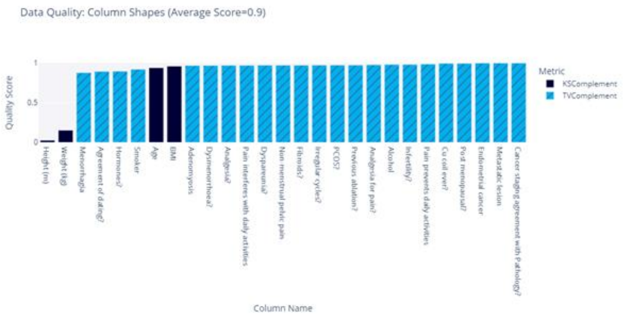

Figure 5. Height Distribution Shape Comparison (Liverpool).

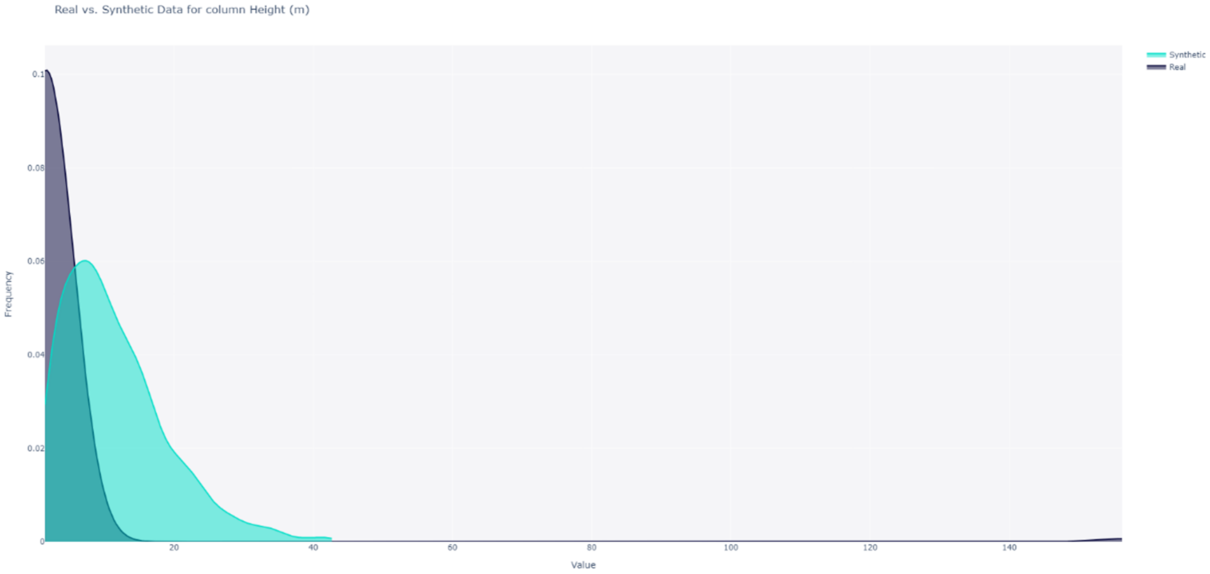

Figure 6. Feature Distribution Shape Comparison Between Real and Synthetic Data (Liverpool).

Models

We evaluated three standard classification models to predict the response variables; Logistic regression (LR), Support Vector Machines (SVM), and Random Forest (RF), as they employ distinct methods data separation and provide unique insights.

Logistic regression enables us to determine the likelihood of each class occurring. It offers straightforward interpretability of the model's coefficients, allowing us conduct statistical tests on these coefficients to discern which features significantly impact the response variable's value. While logistic regression adopts a more statistical approach by maximising the conditional likelihood of the training data, SVMs take a more geometric approach, maximising the distance between the hyperplanes that separate the data. We fitted both logistic regression and SVMs to compare the performance of these approaches.

In contrast to SVMs and logistic regression, which attempt to separate the data using a single decision boundary, random forest employ decision trees that partition the decision space into smaller regions using multiple decision boundaries.

The performance of these varies depending on the nature of the data's separability. Consequently, we fitted all three models and compared their accuracies to assess the useability of the synthetic data.

Logistic Regression

Let $y = (y_1, \dots, y_n)$ to be the general vector of response variables and let $x_i = (x_{i1}, \dots, x_{ip})$ be the corresponding vector of features for patient i. We defined the function:

$$\sigma_\beta(x_i) = P(y_i = 1) = \frac{1}{1 + e^{-\beta x_i}}$$

as be the probability of patient i developing the condition corresponding to y, where $\beta = (\beta_1, \dots, \beta_p)$ are some weights. The prediction function is then defined to be:

$$f_\beta(x_i) = \begin{cases} 0 & \text{if} \quad \sigma_\beta(x_i) < 0.5 \\ 1 & \text{if} \quad \sigma_\beta(x_i) \geq 0.5 \end{cases}$$

We determined the optimal weights by solving the optimisation problem:

$$\min_\beta \mathcal{L}(\beta)$$

where, for logistic regression, the loss function $\mathcal{L}$ took the form:

$$\mathcal{L}(\beta) = \sum_{i=1}^{n} -y_i \log\left(\sigma_\beta(x_i)\right) - (1 - y_i) \log\left(1 - \sigma_\beta(x_i)\right).$$

Finally, we incorporated regularisation terms λ to prevent overfitting, which facilitated capturing the underlying distribution of the data without the proposed model to become overly specific to the training data. This approach helped mitigate any potential biases.

$$\mathcal{L}(\beta) = \sum_{i=1}^{n} y_i \log\left(\sigma_\beta(x_i)\right) + (1 - y_i) \log\left(1 - \sigma_\beta(x_i)\right) + \frac{1}{\lambda} ||\beta||_2^2. \qquad \textit{(Equation 4)}$$

SVMs

Next, we examined Support Vector Machines. We slightly redefined our response variables from binary $\{0,1\}$ to binary $\{-1,1\}$. For instance, suppose y_i^M represents the binary response for a patient developing a mental health condition; then y_i^M is defined as:

$$y_i^M = \begin{cases} 1 & \text{if patient } i \text{ developed a mental health condition} \\ -1 & \text{if patient } i \text{ did not develop any mental health condition.} \end{cases}$$

For SVMs, the prediction function takes the form:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 May 2023 doi:10.20944/preprints202305.1337.v1

$$f_\beta(x_i) = \text{sign}(\beta^T x_i - b)$$

Where $\beta \in \mathbb{R}^p$ and $b \in \mathbb{R}$ are some weights. We considered the hinge loss function, defined as:

$$\ell_{\text{hinge}}(\beta, b) \coloneqq \max_{\beta,b} \left(0, 1 - y_i(\beta^T x_i - b)\right)$$

The function ℓ_{hinge} is 0 when $y_i(\beta^T x_i - b) \geq 1$, which occurs when $f_\beta(x_i) = y_i$ or in other words, when we have made a correct prediction. Conversely, when $f_\beta(x_i) \neq y_i$, we would incur some penalty. Therefore, for SVMs, the loss function, $\mathcal{L}$ takes the form:

$$\mathcal{L}(\beta, b) = \frac{1}{\lambda}||\beta||^2 + \frac{1}{n}\sum_{i=1}^{n} \max_{\beta,b} \left(0, 1 - y_i(\beta^T x_i - b)\right) \qquad \textit{(Equation 5)}$$

where λ is a parameter controlling the impact the of regularisation term. Similar to logistic regression, this term manages a trade-off between capturing the distribution of the entire population and overfitting to the training data.

Random Forest

The final model we fitted is the random forest predictor. These random forests classify data points through an ensemble of decision trees. The decision trees operate by separating the predictor space by a series of linear boundaries. As before, we let $y = (y_1, \dots, y_n), y \in \{0,1\}^n$ be our set of response variables with corresponding feature vectors $x = (x_1, \dots, x_n)$ where each $x_i \in \mathbb{R}^p$. To construct our random forest, we followed the procedure:

For $b = 1, \dots, B$:

1. Sample, with replacement, $x^b \in \mathbb{R}^{m \times p}$ and $y^b \in \{0,1\}^m$ from x and y respectively.
2. Fit k decision trees, $f_1^b, \dots, f_k^b$ to dataset (x^b, y^b)

When making predictions on unseen data, the model took the majority vote across all trees.

For all experiments, we split the real data in half, yielding one training set of real data and one test set of real data. The entire synthetic data is used as training data. All models utilise the test set of real data, thus enabling us to compare the performance between models trained on real data and models trained on synthetic data.

All the models contain hyperparameters that impact the performance of the model on unseen data. For each model, we performed hyperparameter optimisation by using a grid search, measuring the accuracy through cross-validation to find the optimal selection of the hyperparameter. k-fold cross-validation divides the data into k subsets, and by training the model on $k - 1$ subsets and testing on the remaining set, we obtained an estimate of how the model will perform on unseen data. This process was repeated, holding out a different subset for testing each time, and an average performance is calculated. We performed a cross-validation grid search using the training data and select the value of the hyperparameter that yields the best average accuracy, and then retrain the model on the complete training set.

Manchester Data

We divided the real-world data into a training set of 50 samples and a test set of 49 samples. The synthetic data remains as a training set of 1000 samples.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 May 2023 doi:10.20944/preprints202305.1337.v1

Logistic Regression

We used scikit-learn to fit logistic regression models of the form in equation (4). We performed a grid search using 5-fold cross-validation (CV) to investigate the optimal value of λ. The CV accuracies for each response variable can be found in Figure 7. For each response variable, the λ that yielded the highest CV accuracy was chosen, refitted to the entire training set, and tested on the test set. The accuracies on the test set and summarised in Table 7.

	Real		**Synthetic**	
	$\boldsymbol{\lambda}$	**Accuracy**	$\boldsymbol{\lambda}$	**Accuracy**
IBS	10^{-5}	59.18%	1	95.45%
Mental Health	10^{-2}	77.55%	10^{-2}	60.39%
Comorbidities (Other)	1	48.98%	10^{3}	61.04%
Combined	10^{-5}	71.43%	10^{2}	71.43%
Average		64.29%		69.90%

Table 7. Logistic Regression Model Comparison Across Real and Synthetic Data.

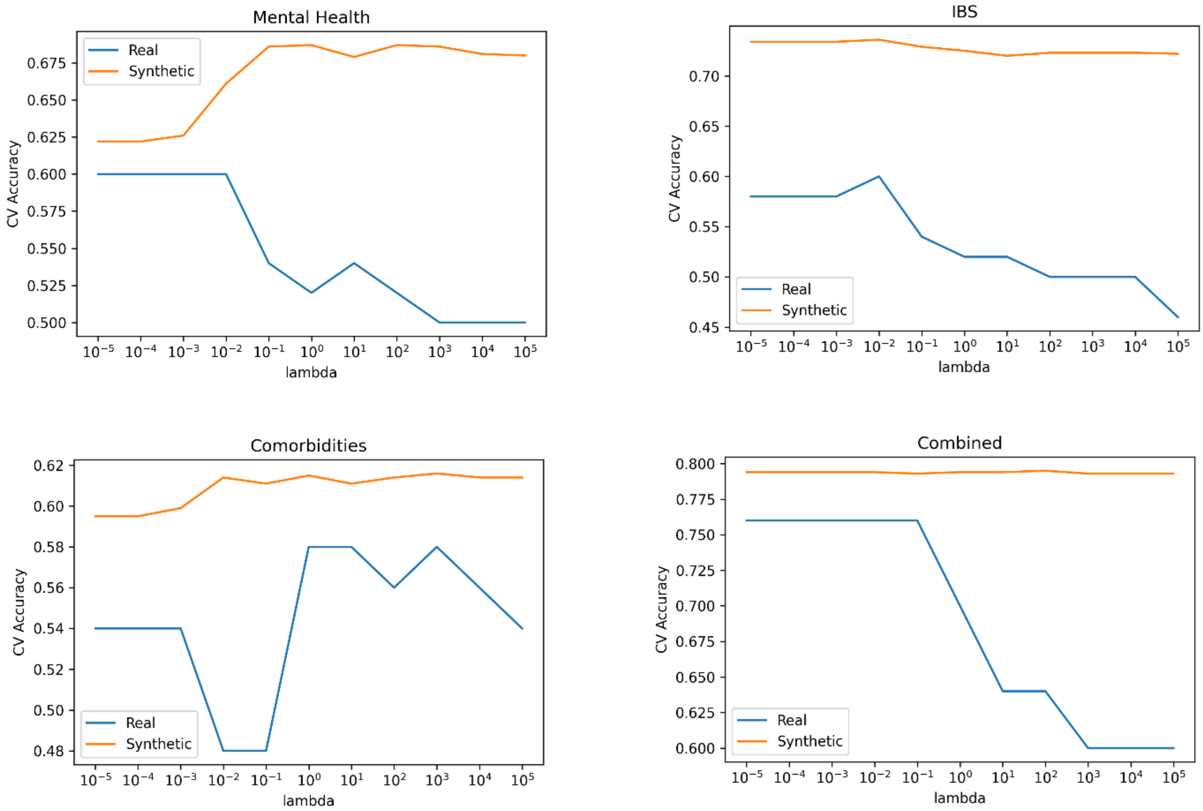

Figure 7. Logistic Regression Cross-Validation.

We can see that for all response variables, the models performed slightly better when trained on synthetic data.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 May 2023 doi:10.20944/preprints202305.1337.v1

SVM

We used Scikit-learn's svm.SVC to train and test SVMs of the form in equation (5) on our data., Scikit-learn is a popular and well-tested choice for SVMs that has shown high performance on a variety of types of datasets. We performed a grid search using 5-fold cross-validation to find the optimal value of λ; these are summarised in Figure 8. For each response variable, the λ that yielded the highest CV accuracy was chosen, refitted to the entire training set, and tested on the test set. The accuracies on the test set are summarised in Table 8. We can see that the models trained on synthetic data performed the same as the models trained on real data for all response variables except "Comorbidities (other)," where the model trained on synthetic data performed better, giving the models trained on synthetic data a better performance on average.

	Real		**Synthetic**	
	$\boldsymbol{\lambda}$	**Accuracy**	$\boldsymbol{\lambda}$	**Accuracy**
IBS	10^4	59.18%	10^5	59.18%
Mental Health	10^{-5}	79.59%	10^{-5}	79.59%
Comorbidities (other)	10^{-5}	59.18%	10^5	67.35%
Combined	10^{-5}	71.43%	10^{-5}	71.43%
Average		67.35%		69.39%

Table 8. SVM comparison with synthetic data.

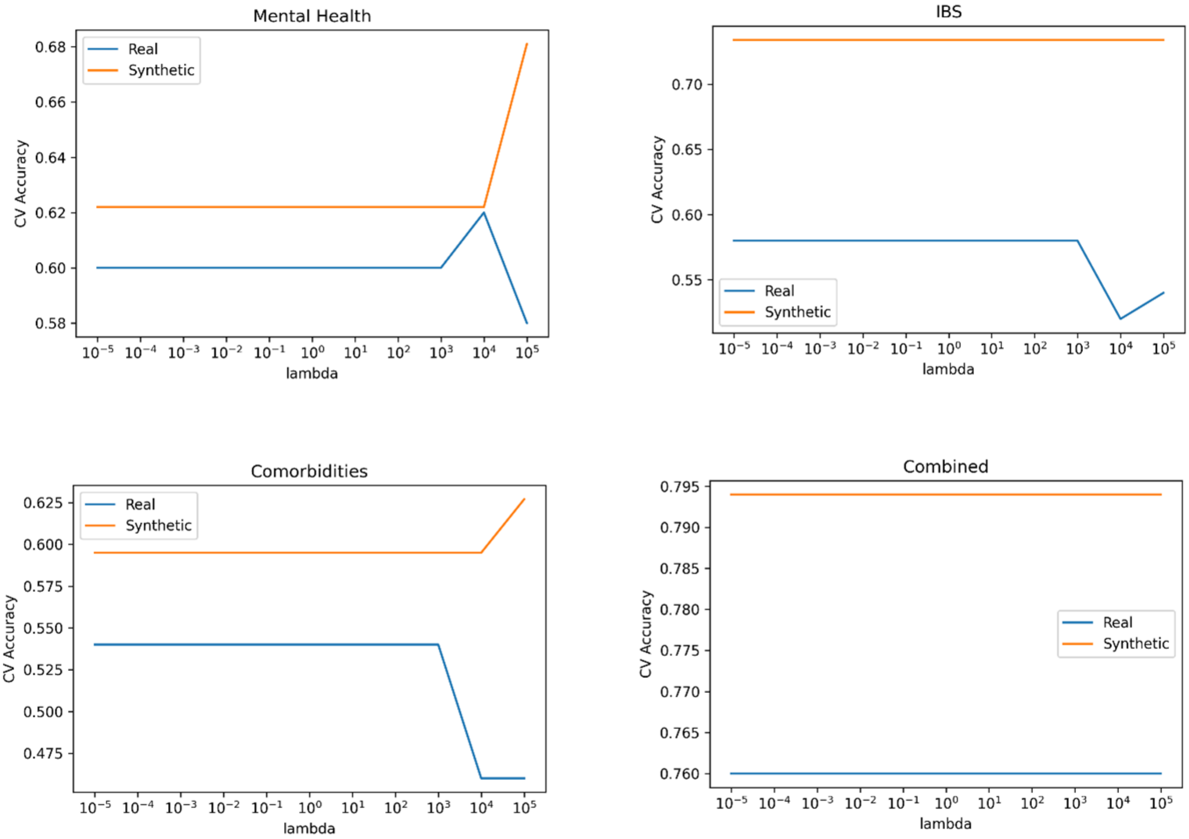

Figure 8. SVM Cross-Validation.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 May 2023 doi:10.20944/preprints202305.1337.v1

Random Forest

Finally, we fitted random forest models to the data. For each response variable, we used 1,5,10, ... ,500 trees, and performed a grid search using 5-fold cross-validation to find the best number of trees (Figure 9). The CV accuracies are summarised in Table 9. For each response variable, the number of trees that gave the highest CV accuracy was chosen, refitted to the whole training set, and tested on the test set. The accuracies on the test set and summarised in Table 9.

	Random Forest accuracy			
	Real		**Synthetic**	
	No. Trees	**Accuracy**	**No. Trees**	**Accuracy**
IBS	55	55.10%	175	59.18%
Mental Health	210	61.22%	455	87.76%
Comorbidities (other)	1	51.02%	20	57.14%
Combined	5	67.35%	480	71.43%
Average		58.67%		68.88%

Table 9. Random Forest Comparison with Synthetic Data.

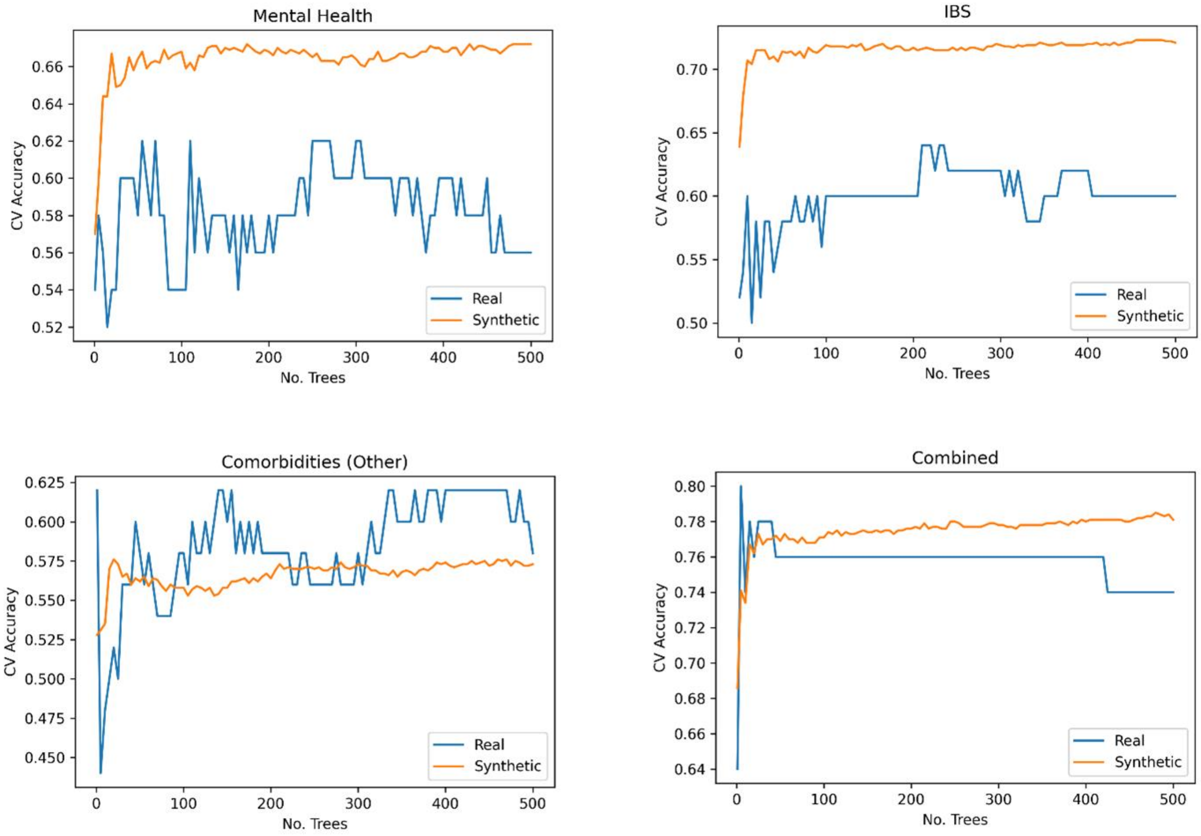

Figure 9. SVM Comparison with Synthetic Data.

Upon examining the average accuracies of all our models in Table 10, we can draw some conclusions about the performance of the models trained on synthetic data compared to those trained on real data. It is evident that on average, models trained on synthetic data performed better than those trained on real data across all response variables. This result suggests that synthetic data can effectively replace real data in training machine learning models without sacrificing performance. In some cases, it may even lead to improved performance, as demonstrated by the results.

Solver comparison

In conclusion, the use of synthetic data proves to be a promising approach to training machine learning models when real data is limited or unavailable. The models trained on synthetic data in this study were not only able to perform at a comparable level to those trained on real data, but they often outperformed them. This finding supports the adoption of synthetic data generation methods as a viable alternative or complement to real data in machine learning applications.

Data	Logistic Regression	SVM	Random Forest
Real	64.29%	67.35%	58.67%
Synthetic	69.90%	69.39%	68.88%

Table 10. Sensitivity Analysis on Manchester Data.

Sensitivity Analysis

To assess our model's sensitivity, we introduced random noise to the data and measured the impact on model accuracy. We randomly selected 1% of points in each dataset and replaced their values. Table 11 summarises the accuracy of the new models and the relative percentage change in accuracy.

Data	Logistic Regression		SVM		Random Forest	
	Accuracy	Change	Accuracy	Change	Accuracy	Change
Real	66.33%	+2.04%	67.35%	+0%	60.71%	+2.04%
Synthetic	70.40%	+0.5%	70.41%	+1.02%	67.35%	−1.53%

Table 11. Comparison of all Models.

Table 11 reveals that the accuracy of the model was impacted in some instances. The logistic regression model trained on real data was affected by more than 2% while the accuracy of its synthetically trained counterpart was only changed by 0.5%. Neither dataset shows a consistency to how the models were affected.

Liverpool Results

We divided the real-world data into a training set of 154 samples and a test set, also containing 154 samples. The synthetic data was retained as a training set comprising 1000 samples.

Logistic Regression

We employed Scikit Learn to fit logistic regression models, as described in Equation (45),and performed a grid search using 5-fold cross-validation to investigate the optimal value of λ. The CV accuracies for each response variable can be found in figure 10. For each response variable, the λ that gave the highest CV accuracy is chosen, refitted to the whole training set and tested on the test set. The accuracies on the test set are summarised in table 12.

	Real		Synthetic	
	λ	Accuracy	λ	Accuracy
Adenomyosis	10^{-5}	95.45%	10^{-5}	95.45%
Menorrhagia	10^{-5}	56.49%	10^{3}	60.39%
Combined	10^{-5}	55.19%	10^{-1}	61.04%
Average		69.04%		72.29%

Table 12. Logisitc Regression Model Comparison (Liverpool).

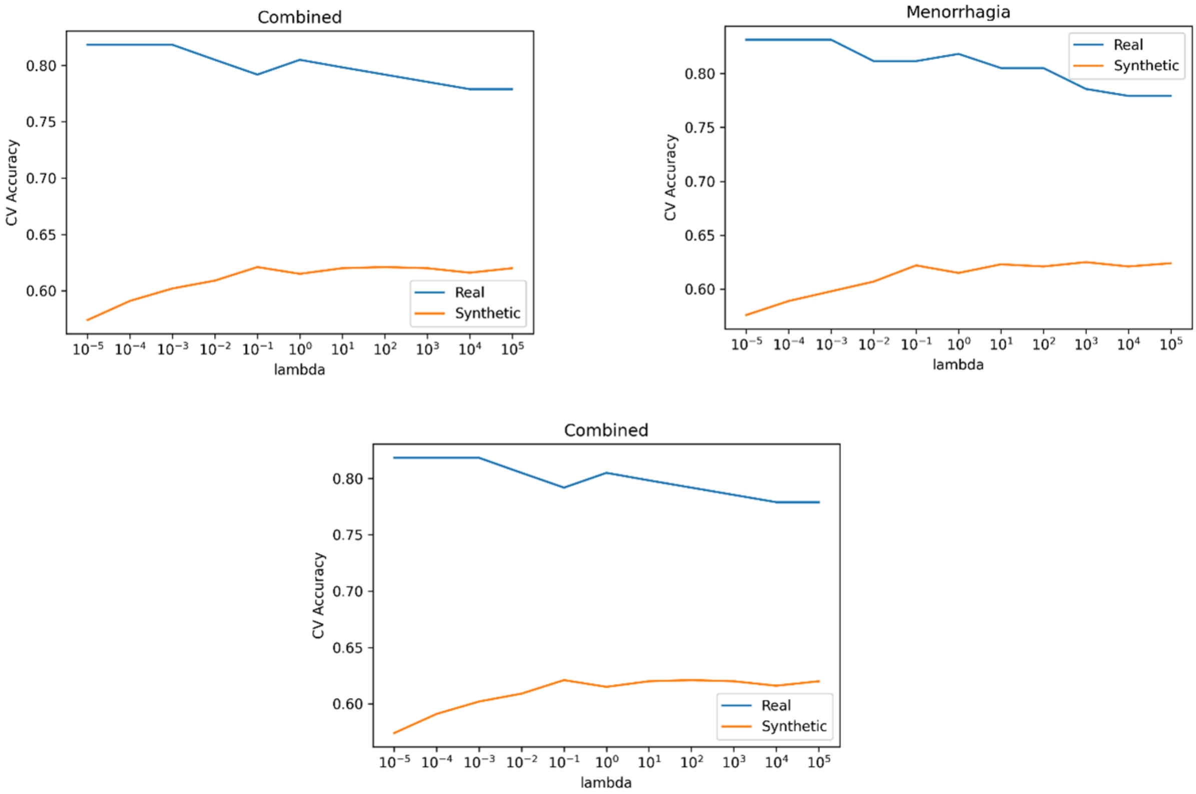

Figure 10. Logistic Regression Cross-Validation.

The models in Table 12, trained on synthetic data, exhibited similar performance in predicting adenomyosis and demonstrated improved performance for the remaining response variables

SVM

We trained and tested SMVs, as defined in Equation (5), using Scikit Learn's svm.SVC on our data. We employed a 5-fold cross-validation grid to find the optimal value of λ, summarised in figure 11. For each response variable, we selected the λ yielding the highest CV accuracy, refitted it to the entire training set, and evaluated it on the test set. Test set accuracies are provided in table 13.

	Real		Synthetic	
	λ	Accuracy	λ	Accuracy
Adenomyosis	10^{4}	94.16%	10^{-5}	95.45%
Menorrhagia	10^{-5}	56.49%	10^{4}	57.79%
Combined	10^{-5}	55.19%	10^{4}	57.14%
Average		68.61%		70.13%

Table 13. SVM Model Comparison.

The models trained on synthetic data demonstrated improved performance for all response variables (Table 13).

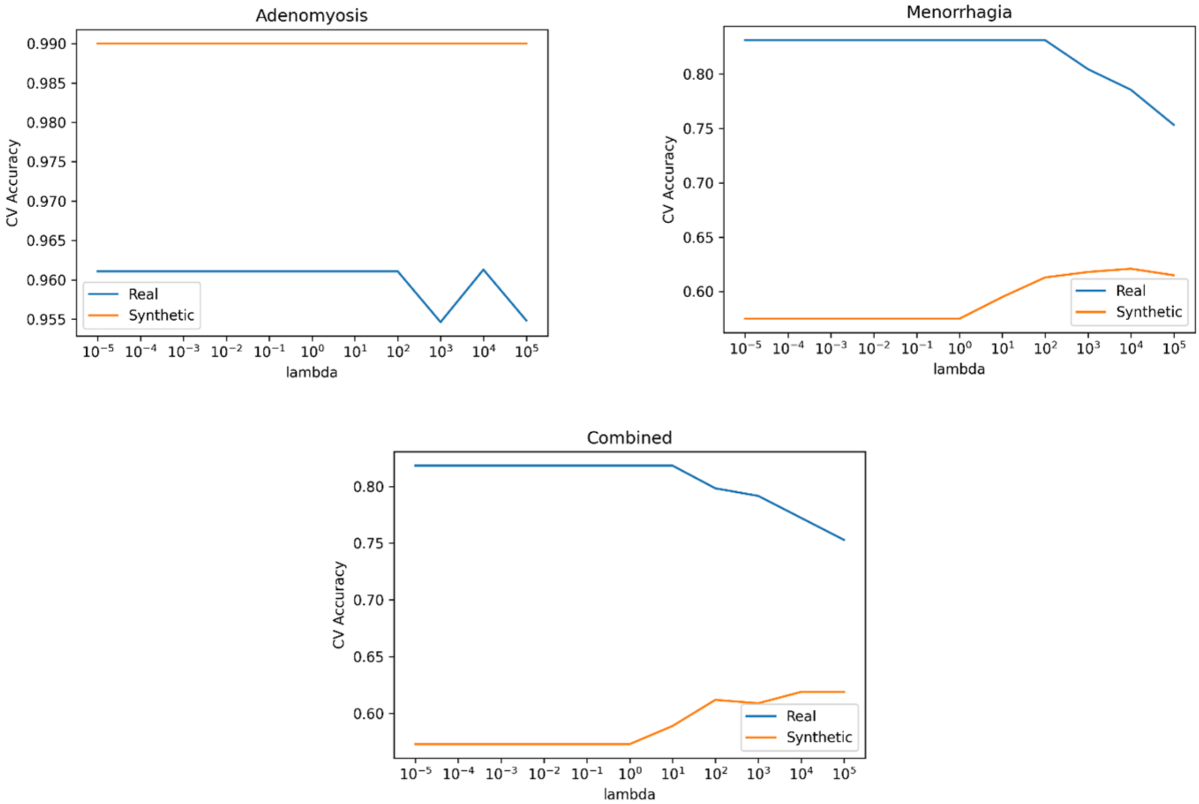

Figure 11. SVM Cross-Validation.

Random Forest

We fitted random forest and performed a grid search using 5-fold cross-validation to test 1,5,10, … , 500 trees, which displayed a higher degree of CV accuracy (figure 12). For each response variable, we chose the number of trees with the highest CV accuracy, , refitted them to the entire training set, and evaluated them on the test set. Test set accuracies are summarised in Table 14.

	Random Forest accuracy			
	Real		Synthetic	
	No. Trees	Accuracy	No. Trees	Accuracy
Adenomyosis	**5**	95.45%	**5**	95.45%
Menorrhagia	5	55.84%	495	59.09%
Combined	140	52.60%	35	55.19%
Average		73.76%		74.62%

Table 14. Random Forest Model Comparison.

The model trained on synthetic data performed comparably to the model trained on real data for adenomyosis and better for the remaining response variables. Consequently, the models trained on synthetic data performed better on avarage.

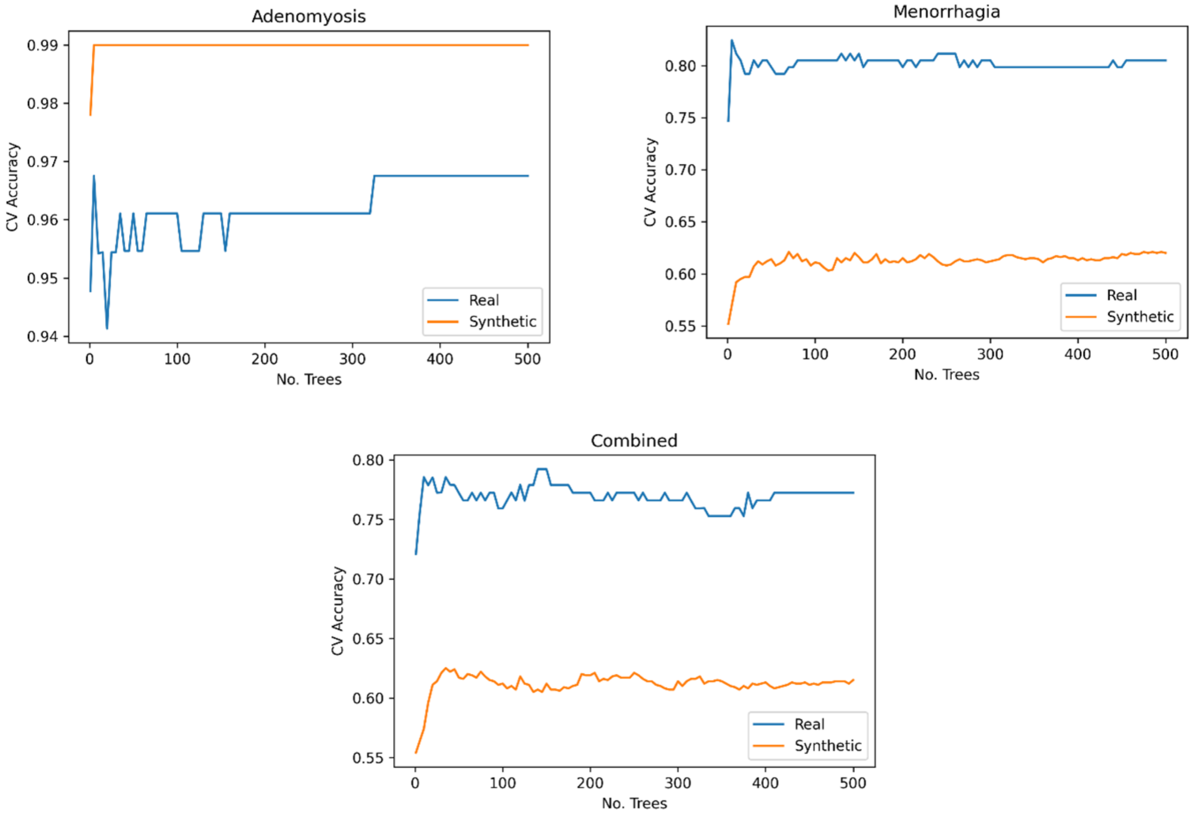

Figure 12. Random Forest Cross-Validation.

Solver Comparison

To summarise, the average accuracies of all models are presented in Table 15. The best over-all performing model is the random forest, trained on the synthetic data. We also observed that, on average, the models trained on synthetic data outperformed those trained on real data. These results support the use of synthetic data in place of the real data.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 May 2023 doi:10.20944/preprints202305.1337.v1

Data	Logistic regression	SVM	Random Forest
Real	69.04%	68.61%	73.76%
Synthetic	72.29%	70.13%	74.62%

Table 15. Comparison of all Models.

Sensitivity Analysis

To test the sensitivity of our models we added random noise to the data and measured its impact on model accuracy.. By sampling from a unform distribution, we randomly selected 1% of points in each dataset to introduce noise. The values at these points were replaced by random samples from a uniform distribution over the feature's possible values. Table 16 displays the accuracy of the new models and their relative percentage change in accuracy.

Data	Logistic Regression		SVM		Random Forest	
	Accuracy	Change	Accuracy	Change	Accuracy	Change
Real	69.05%	+0.01%	69.05%	+0.44%	69.48%	−4.28%
Synthetic	73.38%	+1.09%	70.56%	+0.43%	67.97%	−6.65%

Table 16. Sensitivity Analysis on Liverpool Data.

From Table 16, we can observe that the performance of the logistic regression and SVM models experienced minimal change. However, the random forest models demonstrated a significant drop in performance, indicating that they are sensitive to perturbations in the data. This suggests that for the random forest models, it is crucial for the synthetic data's distribution to closely resemble the real data, as the models are sensitive to small variations.

Comparison Across Datasets

	Logistic regression		SVM		Random Forest	
Data	Manchester	Liverpool	Manchester	Liverpool	Manchester	Liverpool
Real	64.29%	69.04%	67.35%	68.61%	58.67%	73.76%
Synthetic	69.90%	72.29%	69.39%	70.13%	68.88%	74.62%

Table 17. Comparison of all Models.

Table 17 compares the model accuracies across both datasets. We observed that the models trained on the Liverpool dataset consistently out-perform those trained on the Manchester dataset, for both real and synthetic data.

The two datasets documented different attributes of individuals and contained varying numbers of features and observations. The Liverpool dataset had a larger number of both features and observations, and our method performed well in both datasets. These results support the idea that our method can be applied to a diverse range of datasets. The experiments have also demonstrated the effectiveness our method is with both continuous and categorical data. From the distribution analysis of the Liverpool synthetic data, we observed that our method's performance was weakest on two continuous features.

Throughout the experiments, we showed that synthetic data performed similarly or better than those trained on real data. Since all models were tested on real data, this evidence supports the argument that synthetic data can be used as a replacement for real data.

Discussion

Multimorbidity is a growing concern within the global population, particularly for those with chronic conditions like endometriosis, where treatment options are limited. Predicting multimorbidity is challenging among endometriosis patients due to late diagnoses. Therefore, employing machine learning methods to use key features to predict the possibility of multimorbidity is valuable for healthcare services, patients and clinicians. Our findings suggest that the method could be replicated for other complex women's health conditions such as polycystic ovary syndrome, gestational diabetes or fibroids.

Our findings indicate that the real-world dataset contained one variable as a significant indicator for developing multimorbidity and highlighted the usefulness of synthetic data for future research, especially in cases with higher rates of missing data. Synthetic data can also provide more detailed information regarding the relationships between these variables, as they could be considered significant indicators. These indicators can be used to differentiate between samples with symptoms and those with disease sequalae that would influence the clinical decision-making process, particularly for patients requiring excision surgery. With a larger sample size and better representation of the overall population, synthetic data has the potential to provide more detailed information about the significance of each feature.

Previous research used methods such as pairwise comparisons to assess diseases in pairs and combined results where appropriate with similar diseases. This technique may have a higher error rate, as complex chronic diseases do not follow a *one-size fits-all* approach. Whilst the pairwise class of techniques could demonstrate co-occurrence of frequencies and predicted frequencies dissimilar, they can still show a correlation, as indicated by Hidalgo and colleagues' disease network that represented nodes and edges [6]. This is akin to a network meta-analysis approach. A limitation with this approach in disease prediction could be the lack of temporal data in the resulting network nodes, necessitating an additional analysis such as a correlation evaluation [6]. This also means that data with missing data points may be entirely deleted, impacting the final analysis and any subsequent conclusions. Correlation analyses would enable researchers and clinicians to understand the spread of the diseases based on the links shown within the network that can be modelled over time [6]. Jensen and colleagues demonstrated a similar temporal network approach, showing that a pairwise method can be combined with a correlation analysis over time [7]. Giannoula and colleagues used this approach to reveal disease clusters using a time warping along with a pairwise method to mine multimorbidity patterns and phenotyping with extensive data points [8]. In comparison, our combined approach of machine learning on a synchronised dataset can provide better multimorbidity prediction.

Another class of models used to predict multimorbidity is probabilistic methods, which focus on the relationships among diseases rather than a pairwise approach. Strauss and colleagues employed this method to model a small real-world dataset from the UK evaluating multimorbidity cluster trajectories. Individual patients were grouped in clusters based on the number of chronic conditions detected within their healthcare record over a specific period. These clusters were divided into four main categories, including the presence or absence of chronic problems in the number of comorbidities. However, this approach did not consider patients with undiagnosed symptoms aligned with chronic conditions, which is a common observation in real-world data.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 May 2023 doi:10.20944/preprints202305.1337.v1

The distribution of the synthetic data captures the true distribution of the real-world data but can have an arbitrary larger sample size, indicating that synthetic data has the potential to provide valuable insight for healthcare services To address the increasing and complex healthcare demands of a growing population, effective clinical service design is crucial for healthcare sustainability., Moreover, our results show that synthetic data accurately represents the real data and so can be used in place of the real data in cases where the real data contains sensitive or private information that cannot be shared. The accuracy measures of our models support the hypothesis that the use of synthetic data does not affect the performance of the prediction models used in this analysis.

Limitations

The model performance will need to be tested on more complex and larger datasets to ensure that a digital clinical trial can be conducted to optimise the model performance.

Conclusion

Our study created an exploratory machine learning model that can predict multimorbidity among endometriosis women using real-world and synthetic data. Before experimenting with the models developed using the real-world dataset, a quality assessment test was conducted by comparing the synthetic and real-world datasets. Distribution and similarity plots suggested that the synthetic data did indeed follow the same distribution as the real-world data. Therefore, synthetic data generation shows great promise, especially for conducting high- quality clinical epidemiology and clinical trials that could devise better precision treatments for endometriosis and, possibly prevent multimorbidity.

Conflicts of interest: PP has received a research grant from Novo Nordisk, Janssen Cilag, and other, educational from the Queen Mary University of London, other from John Wiley & Sons, outside the submitted work.All other authors report no conflict of interest. The views expressed are those of the authors and not necessarily those of the NHS, the National Institute for Health Research, the Department of Health and Social Care or the Academic institutions.

Availability of data and material: The authors will consider sharing the dataset gathered upon receipt of reasonable requests.

Code availability: The authors will consider sharing the dataset gathered upon receipt of reasonable requests.

Author contributions: FEINMAN is part of the ELEMI program developed and conceptualised by GD. GD and PP conceptualised and developed work package 1 of the FEINMAN project. GD devised the use of synthetic data to better asses' chronic diseases. GD devised the hypothesis for using synthetic data modelled on clinical symptoms to develop optimal prediction models. GD, AZ and PP furthered the study protocol. GD developed the method and furthered this with PP, AZ, DB, JQS, HC, DKP and AS. GD, DB, PP and AZ designed and executed the analysis plan. All authors critically appraised, commented and agreed on the final manuscript. All authors approved the final manuscript.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 May 2023 doi:10.20944/preprints202305.1337.v1

References

1. Delanerolle G, Ramakrishnan R, Hapangama D, Zeng Y, Shetty A, Elneil S, Chong S, Hirsch M, Oyewole M, Phiri P, Elliot K, Kothari T, Rogers B, Sandle N, Haque N, Pluchino N, Silem M, O'Hara R, Hull ML, Majumder K, Shi JQ, Raymont V. A systematic review and meta-analysis of the Endometriosis and Mental-Health Sequelae; The ELEMI Project. Womens Health (Lond). 2021 Jan-Dec;17:17455065211019717. doi: 10.1177/17455065211019717
2. Alimohammadian M, Majidi A, Yaseri M, Ahmadi B, Islami F, Derakhshan M, Delavari A, Amani M, Feyz-Sani A, Poustchi H, Pourshams A. Multimorbidity as an important issue among women: results of a gender difference investigation in a large population-based cross-sectional study in West Asia. BMJ open. 2017 May 1;7(5):e013548.
3. Tripp-Reimer T, Williams JK, Gardner SE, Rakel B, Herr K, McCarthy AM, Hand LL, Gilbertson-White S, Cherwin C. An integrated model of multimorbidity and symptom science. Nursing outlook. 2020 Jul 1;68(4):430-9.. doi:10.1016/j.outlook.2020.03.003
4. Oni T, McGrath N, BeLue R, Roderick P, Colagiuri S, May CR, Levitt NS. Chronic diseases and multi-morbidity-a conceptual modification to the WHO ICCC model for countries in health transition. BMC public health. 2014 Dec;14(1):1-7. https://doi.org/10.1186/1471-2458-14-575
5. Delanerolle GK, Shetty S, Raymont V. A perspective: use of machine learning models to predict the risk of multimorbidity. LOJ Medical Sciences. 2021 Sep 14;5(5).
6. Hassaine A, Salimi-Khorshidi G, Canoy D, Rahimi K. Untangling the complexity of multimorbidity with machine learning. Mechanisms of ageing and development. 2020 Sep 1;190:111325.
7. Jensen AB, Moseley PL, Oprea TI, Ellesøe SG, Eriksson R, Schmock H, Jensen PB, Jensen LJ, Brunak S. Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients. Nature communications. 2014 Jun 24;5(1):4022.
8. Giannoula A, Gutierrez-Sacristán A, Bravo Á, Sanz F, Furlong LI. Identifying temporal patterns in patient disease trajectories using dynamic time warping: A population-based study. Scientific reports. 2018 Mar 9;8(1):1-4.