DATA NOTE

A synthetic dataset for the exploration of survival and classification models: prediction of heart attack or stroke within a 10-year follow-up period [version 1; peer review: awaiting peer review]

Dan Burns[1,2], Kathryn Richardson[3], Corine Driessens[2]

[1]School of Electronics and Computer Science, University of Southampton, Southampton, SO17 1BJ, UK
[2]NIHR Applied Research Collaboration Wessex, University of Southampton, Chilworth, Southampton, SO16 7NP, UK
[3]Norwich Medical School, University of East Anglia, Norwich, England, NR4 7TJ, UK

v1 **First published:** 01 Nov 2024, **4**:67
https://doi.org/10.3310/nihropenres.13651.1
Latest published: 01 Nov 2024, **4**:67
https://doi.org/10.3310/nihropenres.13651.1

Open Peer Review

Approval Status *AWAITING PEER REVIEW*

Any reports and responses or comments on the article can be found at the end of the article.

Abstract

Machine learning methodologies are becoming increasingly popular in healthcare research. This shift to integrated data science approaches necessitates professional development of the existing healthcare data analyst workforce. To enhance this smooth transition, educational resources need to be developed. Real healthcare datasets, vital for healthcare data analysis and training purposes, have many barriers, including financial, ethical, and patient confidentiality concerns. Synthetic datasets that mimic real-world complexities offer simple solutions. The presented synthetic dataset mirrors the routinely collected primary care data on heart attacks and strokes among the adult population. Training experiences using this synthetic dataset are elevated as the data incorporate many of the practical challenges encountered in routinely collected primary care systems, such as missing data, informative censoring, interactions, variable irrelevance, and noise.

By openly sharing this synthetic dataset, our goal was to contribute a transformative asset for professional training in health and social care data analysis. The dataset covers demographics, lifestyle variables, comorbidities, systolic blood pressure, hypertension treatment, family history of cardiovascular diseases, respiratory function, and experience of heart attack and/or stroke. Methods for simulating each variable are detailed to ensure a realistic representation of the patient data. This initiative aims to bridge the gap in sophisticated healthcare datasets for training, fostering professional development in the healthcare and social care research workforce.

Plain Language Summary

In healthcare research, computer programs that can learn patterns are becoming increasingly common. These programs are known as machine learning programs. This means that healthcare data analysts must learn about this approach. Analysts must test these methods in real healthcare data. However, it is difficult to access these data. However, these datasets are costly and have limited use. Privacy and ethical concerns also play an important role.

There is a need for teaching resources to help analysts learn machine learning. A good solution is to use mock datasets that resemble real world data. This work describes a mock dataset created to resemble real adult patient data. The purpose of the dataset was to predict the risk of a heart attack or stroke. The dataset includes problems that a data analyst encounters and needs to solve. This includes missing data and complex links between the data. It also includes data that are not useful for predicting heart attack or stroke. Finally, some data included recording errors. These problems make the training experience practical and helpful.

We aimed to support the growth of machine-learning skills using this mock dataset. We also aimed to provide better support for health and social care research. This study fills this gap in freely available quality healthcare datasets for training.

Keywords

synthetic dataset, heart attack, stroke, machine learning, statistics, training, education, data

Corresponding author: Dan Burns (d.burns@soton.ac.uk)

Author roles: Burns D: Conceptualization, Data Curation, Formal Analysis, Methodology, Resources, Validation, Writing – Original Draft Preparation, Writing – Review & Editing; **Richardson K**: Conceptualization, Formal Analysis, Writing – Original Draft Preparation, Writing – Review & Editing; **Driessens C**: Conceptualization, Formal Analysis, Funding Acquisition, Project Administration, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This project is funded by the National Institute for Health and Care Research (NIHR) [ARC Wessex (Reference Number NIHR2018870)] and the National Centre for Research Methods. The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care. Dr. Dan Burns and Dr. Corine Driessens were the lead applicants for the funding applications.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Burns D, Richardson K and Driessens C. **A synthetic dataset for the exploration of survival and classification models: prediction of heart attack or stroke within a 10-year follow-up period [version 1; peer review: awaiting peer review]** NIHR Open Research 2024, **4**:67 https://doi.org/10.3310/nihropenres.13651.1

First published: 01 Nov 2024, **4**:67 https://doi.org/10.3310/nihropenres.13651.1

Introduction

One of the primary objectives of healthcare research is to classify individuals into health status groups to predict the likelihood of developing certain health conditions based on their risk characteristics. Traditionally, this classification is performed using statistical methodologies, with only a few potential risk factors considered simultaneously[1]. Alongside increased computing capacity to analyse large datasets and increased awareness of the knowledge potential embedded in routinely collected healthcare data, the approach to data analysis and decision-making has evolved from a theoretical to a more data-driven decision-making process[2].

Machine learning (ML) methodologies have become more popular for classifying healthcare-related projects, and integrated data science approaches should be considered to enhance the ability to extract meaningful insights from the data[3,4]. The strengths of ML methodologies include the accurate generation of general-purpose learning algorithms to identify patterns of variables related to specified health conditions within large datasets that contain vast collections of different variables[5,6]. The strengths of statistical methods include the ability to delve deeper and improve understanding of the underlying relationships between the identified risk factors and health conditions in question[6]. Therefore, in this data-driven era, it is important for modern-day healthcare and social care data analysts to familiarize themselves with a more integrated data science approach[7].

Organisations such as the National Institute for Health and Care Research (NIHR), whose aim is to improve health and social care through research, must facilitate a smooth transition to this integrated data science approach by developing professional development programmes that highlight the connection between statistical concepts and ML algorithms and similarities in terms of model validation, uncertainty estimation, and feature selection. To reduce the learning curve and enhance the integration of these novel approaches, accessible user-friendly educational resources must be developed[8].

The 2023 topic of the Routine Data section of the NIHR Statistics Group[9] annual workshop is the analysis of classification problems using machine learning and statistical methodologies in routine data. While organising this meeting, we observed a lack of freely available, sophisticated, and comprehensive healthcare datasets with sufficiently large sample sizes for training purposes. Access to healthcare datasets is an essential element of health data science training is access to healthcare datasets[10]. To protect patient confidentiality, healthcare data are not available for training[11]. Synthetic datasets can alleviate privacy and ethical concerns associated with accessing real healthcare data for training purposes[12]. A bespoke synthetic healthcare dataset was created for the annual meeting of the 2023 NIHR Statistics Group Routine Data section. Through meticulous simulation techniques[13,14], the synthetic dataset presented here mirrors the complexities and nuances of real-world primary care heart attack or stroke data from adult patients (age 18+ years) within a 10-year follow-up period.

The QRISK tool, used in the United Kingdom (UK) to assess the risk of cardiovascular disease in adults, was taken as inspiration for the creation of the synthetic dataset[15]. Therefore, the synthetic dataset includes factors such as age, sex, smoking status, blood pressure, and medical history. The synthetic dataset has been developed to represent the heterogeneity in heart attack and stroke data commonly encountered in datasets collected from primary care databases, such as the Clinical Practice Research Datalink (CPRD)[16], and incorporates the naturally occurring noise of real-world data. To provide an authentic immersive learning experience, the dataset incorporates incomplete information to enhance trainees' missing data analysis and informative censoring skills. In addition, the synthetic dataset contained variables that did not significantly contribute to the analyses, thereby developing trainees' variable selection and model optimization skills. By making this synthetic dataset available for open access, we endeavored to produce a transformative asset for professional training in the field of health and social care data analysis that focuses on one of the leading causes of global mortality[17].

Methods

Our philosophy for dataset development is to produce a dataset that sufficiently reflects the realism challenges encountered in practice: missing data, informative censoring, interactions, variable irrelevance, and noise. The purpose of the dataset is to enable training in the handling of these realisms, and knowledge of the exact mechanisms generating these effects in the dataset can help facilitate training in this manner. The aim is not to produce a dataset that represents real life: although the data should have a degree of realism, we do not strive to optimize the accuracy.

To maintain some realism in the dataset, we generated covariate values based on the simulated age and sex of each patient based on relationships derived from published studies. This naturally introduces correlations between all variables, while remaining relatively simple to generate. We simulated the data for 100,000 patients using this methodology.

A summary of the fields in the dataset is provided in Table 1. A single row per patient was identified using patient_id. The data included age and sex as demographic factors; lifestyle variables such as body mass index (BMI) and smoking status; and comorbidities including hypertension, family history of cardiovascular disease, atrial fibrillation, chronic kidney disease, rheumatoid arthritis, diabetes, and chronic obstructive pulmonary disorder (COPD). It also includes simulated measurements of the systolic blood pressure (SBP) and forced expiratory volume 1 (FEV-1). Finally, the recording of whether a heart attack or stroke occurred was included, with the associated time to event or censoring. We discuss how these variables were simulated in subsequent sections.

Demographics

To simulate age, we utilised 2020-based interim national population projections from the Office of National Statistics[18]. These data are provided in age groups of 5 years from age 0 to

Table 1. Metadata regarding fields within the dataset.

Field	Description
patient_id	The identifier for the patient.
gender	The sex of the patient. "M" = "Male", "F" = "Female"
age	The age of the patient in years
body_mass_index	The body mass index of the individual, if recorded
smoker	Binary variable indicating whether the person has a record of being a smoker
systolic_blood_pressure	The systolic blood pressure taken at the consultation in units of mmHg, if recorded
hypertension_treated	Binary variable indicating whether the person is currently on hypertension treatments
family_history_of_cardiovascular_disease	Binary variable indicating whether the person has a record of family history of cardiovascular disease
atrial_fibrillation	Binary variable indicating whether person has atrial fibrillation
chronic_kidney_disease	Binary variable indicating presence of chronic kidney disease
rheumatoid_arthritis	Binary variable indicating presence of rheumatoid arthritis
diabetes	Binary variable indicating presence of diabetes
chronic_obstructive_pulmonary_disorder	Binary variable indicating whether the person has chronic obstructive pulmonary disorder
forced_expiratory_volume_1	The forced expiratory volume, the volume of air that an individual can exhale during a forced breath in 1 second, as a percentage of their predicted FEV1
time_to_event_or_censoring	The time to event, or time to censoring, in years
heart_attack_or_stroke_occurred	Binary variable indicating whether a heart attack or stroke occurred

100 years and are represented in terms of 1000s within each group. For this dataset, we considered only individuals aged 18 – 79 years. We assume that within each 5-year age group, the ages are uniformly distributed, that is, each age is equally likely. With this simplification, we contracted the 15 – 19 age group to an 18–19 age group by replacing the corresponding count with a reduced count by a factor of 2/5, that is, two years out of the five years within that group. To sample an individual's age, we sampled the age group based on the counts post-modifications. We then sampled an individual's age from a uniform distribution within that age group.

To simulate sex, we used a Bernoulli distribution with p = 0.5 to indicate whether a person is male or female.

Body Mass Index

To simulate BMI, we used the 2021 Health Survey for England dataset on obesity[19]. This dataset provides aggregated statistics regarding the percentage of individuals in three categories: non-overweight or obese (BMI < 25 kg/m^2), overweight (BMI = 25 kg/m^2), and obese (BMI = 30 kg/m^2). The data were also stratified by sex and the following age groups: 16–24, 25–34, 35–44, 45–54, 55–64, 65–74, and 75+ years.

To sample BMI, we assumed that within each age group and sex combination, BMI follows a normal distribution. To fit a normal distribution to the percentage data, we represented each percentage through the cumulative distribution function (CDF) of a normal distribution for that particular group. Based on this assumption, CDF must be equal to the percentage value at each threshold. Let F represent the CDF for a particular group; then, $F(25)$ represents the percentage of individuals in the "not overweight or obese" group, and $F(30)$ represents those in the former group, combined with those in the overweight group. Finally, to include information on obese patients, we set this to $F(50)$ where 50 kg/m^2 is an arbitrarily high upper bound. We then fit the mean and standard deviation of the normal distribution by minimising the least squares between the CDF values and observed percentages:

$$L = (F(25) - p_{25})^2 + (F(30) - p_{30}) + (F(50) - 1)^2$$

where p_{25}, p_{30} are the percentages observed for the age group and sex combination.

To sample BMI, we used the previously sampled age and sex to match up to the corresponding mean and standard deviation calculated from the above procedure and sample from the corresponding normal distribution.

Systolic Blood Pressure

To simulate SBP, we harness the results from Balijepalli *et al.*[20], who calculated the 5th, 25th, 50th, 75th, and 95th percentiles of the SBP distribution as a function of age group. We fitted the percentiles for each age group to a normal distribution in a similar fashion to the case for BMI: minimising the least squares of the CDF values matched with the observed proportions. To sample SBP, we therefore find the age group of the individual and then sample from the corresponding normal distribution. We refer to this SBP as true SBP.

SBP as a measurement variable suffers from various sources of noise in routine datasets in practice, such as measurement errors and missingness. We modeled the measurement stochasticity arising from physiological variation and measurement device error by creating a measured version of the true SBP. To achieve this, we sampled from a normal distribution centered on the true SBP. If the true SBP is S_t, then the measured SBP S_m is distributed as

$$S_m \sim \mathcal{N}(S_t, \sigma_m^2)$$

We used σ_m = 10 mmHg to reflect the median standard deviation observed by Li *et al.*[21], an ambulatory blood pressure monitoring study based in China. We cover this missingness in a later section.

Hypertension treatment

Hypertension is known to be undertreated. It is estimated that only approximately 40% of hypertensive patients are treated worldwide, and the primary drivers for this low percentage are a lack of diagnosis and patients refusing treatment[22]. Furthermore, therapeutic inertia also has an impact[23] where cases are explicitly missed; this depends explicitly on the true SBP of the individual. From our simulated true SBPs, approximately 33% of the patients had a true SBP above 135 mmHg, leading to approximately 13% of the patients in the dataset being treated for hypertension.

To simulate the effect of an individual being missed as a hypertensive patient, we used a logistic curve to define the probability of treatment. Let $p_{treated}$ be the probability that a patient is treated. Then, we define

$$p_{treated} = \frac{p_{diag}}{1 + \exp\left(\frac{(S_t - 137.5)}{10}\right)}$$

where we set p_{diag} to ensure that the overall prevalence of hypertension treatment was approximately 13%. This leads to $p_{diag} = 0.2$ for this case.

Family history of cardiovascular disease

To simulate whether an individual has a family history, we used the prevalence measured by Chacko *et al.*[24] and used a Bernoulli distribution with p = 0.24 to represent this variable. There are no correlations with the other variables.

Comorbidities

To simulate comorbidities, we utilised joint distributions between each comorbidity, age, and sex. Each comorbidity is sampled using a Bernoulli distribution, where the probability is governed by the individual's age group and sex. For chronic kidney disease, we used the study by Kampmann *et al.*, which provides prevalence estimates as a function of age and sex in Southern Denmark[25]. This study reports the prevalence in three age groups: 18–39, 40–69, and 70+ years for both men and women. For atrial fibrillation, we used Go *et al.*[26], which assessed prevalence as a function of age and sex in the United States in 2001. The age groups are given as 18–54, 55–59 and 5-year age groups, ending with an 85+ age group. For rheumatoid arthritis, we used Symmons *et al.*[27], which stratified the prevalence by age and sex for adults in the UK, with age groups of 16–44, 45–64, 65–74, and 75+ years. For diabetes, we used the Health Survey for England data[19], using prevalence estimates for combined diagnosed and undiagnosed diabetes by age groups 16–44, 45–64 and 65+ years. For COPD, we used Ntritsos *et al.*[28], which provides prevalence estimates by age and sex for the global population, with age groups 15–39, 40–69, and 70+. We adjusted the lower age group to 18–39 for simulation and assumed that the prevalence holds for this subgroup.

FEV1

FEV1 differs significantly between patients with COPD and those who do not have the condition. To simulate FEV1 in patients with COPD, we used the study by Le *et al.*[29], which assessed the prognostic ability of GOLD classifications in Norway. The GOLD categories correspond to FEV1 ranges: ≥ 80% for GOLD 1, 50 – 80% for GOLD 2, 30 – 50% for GOLD 3, and <30% for GOLD 4. This study reports the prevalence of GOLD grades 2, 3, and 4 in patients with COPD. We arbitrarily assumed that the prevalence of GOLD 1 is 50% in COPD patients. Therefore, according to the study, the prevalence of GOLD 2, 3, and 4 was 28%, 15%, and 7%, respectively. To sample FEV1 using these data, we first sampled the individual's GOLD grouping according to the above prevalence. We then sample uniformly from the following ranges: 80 – 82 for GOLD 1, 50 – 80 for GOLD 2, 30 – 50 for GOLD 3, and 20 – 30 for GOLD 4. To measure FEV1 in healthy individuals, we used uniform samples from 90 to 100 irrespective of age and sex.

Smoking status

Smoking status prevalence has been assessed in the Health Survey for England study by age group and sex[19]. The age groups are described in the body mass index section. To simulate smoking status, we extracted the proportion of individuals who currently smoked within each age group and sex combination. We then used a Bernoulli distribution with p equal to the prevalence for that individual's age and sex.

Heart attack or stroke

To simulate a heart attack or stroke, we used a modification of the QRISK3 model[15] to consider only the variates contained in this synthetic dataset. Baseline survival is taken as a step function with 10 steps, with each step occurring each year. The value of the survival function at the steps forms a linear function, starting at probability 1 at t = 0 years, dropping to 0.977 for males and 0.989 for females at t = 10 years. The step-like nature of the survival function replicates an event measured in discrete intervals.

We simplified the QRISK3 algorithm to match our reduced dataset by removing the impact of the majority of variables not present in our dataset. This includes variables corresponding to angina, migraines, systemic lupus erythematosus, severe mental illness, atypical antipsychotic medication, steroid use, erectile dysfunction, ethnicity, and deprivation. For variables within our dataset, QRISK3 had different contributions to the type of diabetes, as well as history of smoking. We set the impact of diabetes on the risk to be equivalent to those with type 2 diabetes and the impact of smoking on the risk as equivalent to light smokers within the QRISK nomenclature. We identified that the cholesterol/HDL ratio and systolic blood pressure standard deviation both had significant impacts on the risk; therefore, we chose to include non-zero values overall. We arbitrarily chose a cholesterol/HDL ratio of 3 and a systolic blood pressure standard deviation of 10 mmHg to align with the median observed standard deviation[21]. The model was evaluated using the individual's true SBP, not their measured SBP, to further introduce noise.

To determine whether the event occurred, we used the model to calculate the likelihood of an event occurring at 10 years for each individual. We then sample the time at which the event occurred using the individual's survival curve using inverse transform sampling, that is, we draw a uniformly random variable on (0,1) and match this to the predicted CDF. If the uniformly random variable is larger than the corresponding point for t = 10 years, then the event is censored. For those who did not experience the event, we set their censoring time to 10 years.

To simulate study dropout and include informative censoring effects, we used a dropout rate based on whether the event occurred or not. If an individual has an event at *t* years, then the likelihood of dropout is $p = 0.01 * t$.

Missingness

After sampling the heart attack or stroke event, we introduced two types of missingness mechanisms among the other variables in the simulated dataset that can occur in routine datasets: Missing at Random (MAR), where the missingness is dependent on other variables in the dataset (e.g., a numerical variable is not explicitly recorded, e.g., BMI), or Missing Completely at Random (MCAR, e.g., a feature of the individual has not been made known to the healthcare system).

We modeled missingness in smoking status, family history of cardiovascular disease, SBP, and BMI variables as MCAR[30]. For smoking status and family history of cardiovascular disease, we converted 1 to 0 with a probability of 30% to obscure the true smoking status and family history. To simulate missingness for the SBP measurements, we sampled from a Bernoulli distribution with $p = 0.1$ to determine whether the variable value should be dropped. We excluded the measured SBP value based on this indicator. For BMI measurements, we removed those with $p = 0.3$.

For FEV1, missingness is modelled as MAR with missingness less likely if the patient has COPD[30]. If an individual had COPD, we removed the variable with $p = 0.05$. If they did not have COPD, we removed the variable with $p = 0.75$.

Finally, we completely excluded the true SBP column, which represents an unknown characteristic of the patient that can only be interpreted through other variables (including the measured SBP).

Patient and Public Involvement

This data note and corresponding dataset was created without patient involvement. Patients were not involved in the design or creation of this synthetic dataset: the dataset's intended purpose is aimed at healthcare data analysts who wish to develop their machine learning skills. Patients were not invited to contribute to this document.

Discussion

Sophisticated synthetic datasets are becoming an increasingly important resource in health data science. We have provided a freely available and comprehensive synthetic healthcare dataset that mirrors the natural associations within UK primary care data to assess the 10-year risk of two commonly studied health outcomes: heart attack and stroke.

We are aware of the CPRD synthetic dataset available for predicting heart attack or stroke using information for 499,344 simulated patients and 21 predictor variables[13]. Although the simulated data in the CPRD dataset are also based on the QRISK models[15], the dataset is generated differently by using a Bayesian Network model trained on real primary care data. This leads to governance issues and, hence, a fee to access and use data. Our synthetic data were built from an a priori modelling framework informed by a literature search. This allows the data to not be generated directly from real patient data, but retains many of the realisms found in such datasets. These synthetic datasets differ in their purpose: ours is for training and learning how to handle such data in practice, whereas the purpose of the CPRD synthetic dataset is to be a faithful representation of CPRD Aurum. The CPRD synthetic dataset only provides a binary outcome variable of heart attack or stroke within five years, whereas ours also provides the follow-up time to censoring or the event, thus additionally enabling classical survival analysis methods to be evaluated on the dataset and the censoring mechanisms to be examined. This was particularly useful in our training event, where we successfully used the synthetic dataset to compare the results of the classical survival analysis with those of machine learning for predicting heart attack or stroke.

Our synthetic dataset was strengthened by including sophisticated realistic relationships between predictor variables and outcomes, as informed by literature searches. Missingness patterns, measurement errors, and informative censoring allow data scientists to tailor their classification training to include more advanced methodologies focused on the inclusion of missing data handling techniques (e.g., multiple imputation and ML imputation) and discussion of classification bias due to informative censoring. The additional benefit of spurious variables in the synthetic dataset encourages data scientists to stay up-to-date with the latest variable selection and model optimization techniques. Limitations include the synthetic dataset having less fidelity than synthetic datasets built directly from real primary care data[13]; however, as mentioned, this enabled us to avoid governance issues and allow the data to be made freely available. Our dataset was limited by not including all the predictive variables within QRISK3. However, we wanted to strike a balance between simulating a reasonably sized realistic dataset for training purposes and covering many types of variables (e.g., continuous, categorical, and various degrees of missingness) without adding unnecessary complexity.

In conclusion, by making this realistic simulated dataset on a highly applicable health research topic available for open access, we endeavored to enhance professional training in the field of health and social care data analysis.

Data availability

The data were hosted by ARC Wessex and are openly available on the ARC Wessex website at https://www.arc-wx.nihr.ac.uk/data-sets.

The data is also hosted by Zenodo, available at https://doi.org/10.5281/zenodo.12567416[31].

This project contains following dataset:

1. cvd_synthetic_dataset_v0.2.csv
2. cvd_synthetic_dataset_v0.2_metadata.xlsx

Data are available under the terms of the Creative Commons Zero "No rights reserved' data waiver (CC0 1.0 Public domain dedication).

Acknowledgements

We would like to acknowledge all attendees of the 2023 NIHR Statistics Group – routine data event. Without your keen interest in professional development, this event would not have been possible. We would also like to thank the members of the organizing committee (Dr. Jessica Harris, Dr. Jianhua Wu, Dr. Jacqueline Birks, Dr. Ge Yu, and Dr. David Culliford) for their stimulating discussions that helped conceptualize the creation of the synthetic dataset.

References

1. Rothman KJ: **Epidemiology: an introduction**. Oxford university press, 2012.
Reference Source
2. Krittanawong C, Virk HUH, Bangalore S, *et al.*: **Machine Learning prediction in cardiovascular diseases: a meta-analysis.** *Sci Rep.* 2020; **10**(1): 16057.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**
3. Alsuliman T, Humaidan D, Sliman L: **Machine Learning and Artificial Intelligence in the service of medicine: necessity or potentiality?** *Curr Res Transl Med.* 2020; **68**(4): 245–251.
PubMed Abstract | **Publisher Full Text**
4. Naseem M, Akhund R, Arshad H, *et al.*: **Exploring the potential of Artificial Intelligence and Machine Learning to combat COVID-19 and existing opportunities for LMIC: a scoping review.** *J Prim Care Community Health.* 2020; **11**: 2150132720963634.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**
5. Rathore DK, Mannepalli PK: **A review of Machine Learning techniques and applications for health care.** *2021 International Conference on Advances in Technology, Management & Education (ICATME).* 2021; 4–8.
Publisher Full Text
6. Bzdok D, Altman N, Krzywinski M: **Statistics versus Machine Learning.** *Nat Methods.* 2018; **15**(4): 233–234.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**
7. Miller S, Hughes D: **The quant crunch: how the demand for data science skills is disrupting the job market**. Burning Glass Technologies, 2017.
Reference Source
8. Kolaczyk ED, Wright H, Yajima M: **Statistics practicum: placing "practice" at the center of data science education.** *Harvard Data Science Review.* 2021; **3**(1).
Publisher Full Text
9. NIHR Statistics Group: **Routine Data Section member list**. 2024.
Reference Source
10. Gonzales A, Guruswamy G, Smith SR: **Synthetic data in health care: a narrative review.** *PLOS Digit Health.* 2023; **2**(1): e0000082.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**
11. Beauchamp TL, Childress JF: **Principles of biomedical ethics**. Oxford University Press, 2019.
Reference Source
12. Ibrahim ZM, Schmitt W, Palma AM: **A comprehensive review on healthcare simulation research.** *Simul Healthc.* 2021; **16**(1): 61–71.
13. Tucker A, Wang Z, Rotalinti Y, *et al.*: **Generating high-fidelity synthetic patient data for assessing Machine Learning healthcare software.** *NPJ Digit Med.* 2020; **3**(1): 1–13, 147.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**
14. Draghi B, Wang Z, Myles P, *et al.*: **Identifying and handling data bias within primary healthcare data using synthetic data generators.** *Heliyon.* 2024; **10**(2): e24164.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**
15. Hippisley-Cox J, Coupland C, Brindle P: **Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study.** *BMJ.* 2017; **357**: j2099.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**
16. Wolf A, Dedman D, Campbell J, *et al.*: **Data resource profile: Clinical Practice Research Datalink (CPRD) Aurum.** *Int J Epidemiol.* 2019; **48**(6): 1740–1740g.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**
17. World Health Organization: **The top 10 causes of death**. 2020; Accessed 29th February 2024.
Reference Source
18. Office for National Statistics: **National population projections: 2020–based interim [Dataset].** Office for National Statistics. 2020; Accessed 29th February 2024.
Reference Source
19. NHS England: **Health Survey for England 2019 [Dataset].** NHS England. 2019; Accessed 29th February 2024.
Reference Source
20. Balijepalli C, Lösch C, Bramlage P, *et al.*: **Percentile distribution of blood pressure readings in 35683 men and women aged 18 to 99 years.** *J Hum Hypertens.* 2014; **28**(3): 193–200.
PubMed Abstract | **Publisher Full Text**
21. Li W, Yu Y, Liang D, *et al.*: **Factors associated with blood pressure variability based on ambulatory blood pressure monitoring in subjects with hypertension in China.** *Kidney Blood Press Res.* 2017; **42**(2): 267–275.
PubMed Abstract | **Publisher Full Text**

22. NCD Risk Factor Collaboration: **Worldwide trends in hypertension prevalence and progress in treatment and control from 1990 to 2019: a pooled analysis of 1201 population-representative studies with 104 million participants.** *Lancet.* 2021; **398**(10304): 957–980.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

23. Augustin A, Coutts L, Zanisi L, *et al.*: **Impact of therapeutic inertia on Long-Term Blood Pressure Control: a Monte Carlo simulation study.** *Hypertension.* 2021; **77**(4): 1350–1359.
PubMed Abstract | **Publisher Full Text**

24. Chack M, Sarma PS, Harikrishnan S, *et al.*: **Family history of Cardiovascular Disease and risk of premature coronary heart disease: a matched case-control study [version 2; peer review: 2 approved].** *Wellcome Open Res.* 2020; **5**: 70.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

25. Kampmann JD, Heaf JG, Mogensen CB, *et al.*: **Prevalence and incidence of chronic kidney disease stage 3–5 – results from KidDiCo.** *BMC Nephrol.* 2023; **24**(1): 17.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

26. Go AS, Hylek EM, Phillips KA, *et al.*: **Prevalence of diagnosed atrial fibrillation in adults: national implications for rhythm management and stroke prevention: the AnTicoagulation and Risk Factors in Atrial Fibrillation (ATRIA) Study.** *JAMA.* 2001; **285**(18): 2370–2375.
PubMed Abstract | **Publisher Full Text**

27. Symmons D, Turner G, Webb R, *et al.*: **The prevalence of rheumatoid arthritis in the United Kingdom: new estimates for a new century.** *Rheumatology (Oxford).* 2002; **41**(7): 793–800.
PubMed Abstract | **Publisher Full Text**

28. Ntritsos G, Franek J, Belbasis L, *et al.*: **Gender-specific estimates of COPD prevalence: a systematic review and meta-analysis.** *Int J Chron Obstruct Pulmon Dis.* 2018; **2018**(13): 1507–1514.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

29. Le LAK, Johannessen A, Hardie JA, *et al.*: **Prevalence and prognostic ability of the GOLD 2017 classification compared to the GOLD 2011 classification in a Norwegian COPD cohort.** *Int J Chron Obstruct Pulmon Dis.* 2019; **2019**(14): 1639–1655.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

30. Molenberghs G, Fitzmaurice G, Kenward MG, (Eds.), *et al.*: **Handbook of missing data methodology.** CRC Press, 2014.
Publisher Full Text

31. Burns D, Richardson K, Driessens C: **A synthetic dataset for the exploration of survival and classification models: prediction of heart attack or stroke within a 10–year follow-up period [Dataset].** *Zenodo*. 2024; Accessed 27th June 2024.
http://www.doi.org/10.5281/zenodo.12567416