Data in Brief 47 (2023) 108921

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

A synthetic dataset of liver disorder patients

Giovanna Nicora [a,b,*], Tommaso Mario Buonocore [a], Enea Parimbelli [a,c,*]

[a] Department of Electrical, Computer and Biomedical Engineering, University of Pavia, Italy
[b] enGenome Srl, Italy
[c] Telfer School of Management, University of Ottawa, Ottawa, ON, Canada

ARTICLE INFO

Article history:
Received 20 December 2022
Revised 10 January 2023
Accepted 16 January 2023
Available online 20 January 2023

Dataset link: A synthetic dataset of Liver Disorder patients (Original data)

Keywords:
Synthetic patients
Machine learning
Bayesian network
Dataset shift
Causal model

ABSTRACT

The data in this article include 10,000 synthetic patients with liver disorders, characterized by 70 different variables, including clinical features, and patient outcomes, such as hospital admission or surgery. Patient data are generated, simulating as close as possible real patient data, using a publicly available Bayesian network describing a casual model for liver disorders. By varying the network parameters, we also generated an additional set of 500 patients with characteristics that deviated from the initial patient population. We provide an overview of the synthetic data generation process and the associated scripts for generating the cohorts. This dataset can be useful for the machine learning models training and validation, especially under the effect of dataset shift between training and testing sets.

DOI of original article: 10.1016/j.artmed.2022.102471

* Corresponding authors at: Department of Electrical, Computer and Biomedical Engineering, University of Pavia, Italy.
E-mail addresses: giovanna.nicora@unipv.it (G. Nicora), enea.parimbelli@unipv.it (E. Parimbelli).
Social media: @GiovannaNicora (G. Nicora), @detsutut (T.M. Buonocore), @NeneParimbelli (E. Parimbelli)

https://doi.org/10.1016/j.dib.2023.108921

Specifications Table

Subject	Applied Machine Learning
Specific subject area	Synthetic data can support the implementation of trustworthy, privacy-preserving machine learning approaches, especially in high-stakes application, such as medicine.
Type of data	Comma-separated values files containing tables where each row represent a synthetic patient and each column represent a variable (i.e. feature)
How the data were acquired	We simulated 10,000 patients data by sampling from the Bayesian network describing liver disorder patients proposed by [1] and implemented within the R bnlearn package [2]. For each synthetic patient, we report 70 different clinical characteristics, such as age, sex, platelet count, comorbidities, and other features relevant for liver disorder diagnosis that were modeled by [1] to determine a casual model. Additionally, by modifying the prior conditional probabilities and by sampling from the corresponding network, we simulated an additional set of 500 patients as Out-of-distribution (O.O.D.) samples. The R script implemented to simulate all the data is made available through this article.
Data format	Raw
Description of data collection	The data were generated based on the casual probabilistic model that describe how different clinical characteristics interplay in liver disorder.
Data source location	Generative causal model: Bayesian network
Data accessibility	Repository name: Zenodo, Github Data identification number: 10.5281/zenodo.6726768 Direct URL to data: https://zenodo.org/record/6726768#.Y2PWDi1aZQJ, https://github.com/GiovannaNicora/hepar_simulation
Related research article	E. Parimbelli, T. M. Buonocore, G. Nicora, W. Michalowski, S. Wilk, R. Bellazzi, Why did AI get this one wrong? — Tree-based explanations of machine learning model predictions, Artificial Intelligence in Medicine, Volume 135, 2023,102471, ISSN 0933-3657, https://doi.org/10.1016/j.artmed.2022.102471

Value of the Data

- The data generate within this study represent realistic patients' profiles that can be exploited as features set to train and test machine learning models, which are increasingly applied to medical-related problems [3,4].
- Synthetic data have the potential to support developers in training and testing ML classifiers, using a large set of bona-fide data to investigate the behavior of the model, thus helping model inspection and debugging.
- By simulating data from a casual model, it is possible to simulate realistic patient profiles that can be used for training. In turn, modifying features prior probabilities in the causal model allows the simulation of what-if scenarios for which real-world data might not be available or accessible. One straightforward application is the simulation patients that significantly deviate from the training population, i.e. simulating dataset shift. This out-of-distribution samples can be exploited to test the robustness, reliability and explainability of ML classifiers [5].

1. Objective

Our aim is to generate a dataset of synthetic patients with liver disorders, that can be used to develop machine learning models to predict patient outcomes such as the need for hospitalization, based on clinical features.

2. Data Description

This article describes data generated using a Bayesian network documented in [1], in order to build a casual model for liver disorders. An interactive visualization of the network can be found in the Colab file (https://github.com/GiovannaNicora/hepar_simulation). For the dataset described in this article we provide two different files:

- HEPAR_simulated_patients.csv: this comma-separated file contains synthetic data for 10,000 liver disorder patients (*population dataset*). Each row represents one synthetic patient. On each column, a different clinical characteristics (*features*), in this case representing a node of the Bayesian network in [1], is reported, for a total of 70 columns. The majority of the features are binary, indicating the presence/absence of a comorbidity (for instance, *diabetes*) or intervention (for instance *surgery* or *hospital*). Other features, such as age or platelets, are reported as ordinal features. For instance, in case of age, different categories such as age0_30, age31_50, etc. are reported.
- HEPAR_simulated_patients_ood.csv: comma-separated file containing 500 synthetic patients (*O.O.D. dataset*) by changing the prior probability of sex, age and hospitalization, as reported in the https://github.com/GiovannaNicora/hepar_simulation/blob/main/hepar_simulate_patients.R) file.

Table 1. lists the different variables (reported in the columns of the aforementioned files) that can be used to build machine learning models.

Table 1
List of the columns (features) corresponding to the Bayesian network nodes. The features are grouped into 3 different categories (clinical features, comorbidities and interventions), and their values types (binary, mostly "absent" or "present"), categorical and nominal (numerical values that have been discretized by [1] in order to build the Bayesian network.

	Clinical features	Comorbidities	Interventions
Binary (mostly absent/present)	Sex, fatigue, itching, upper pain, fat, pain ruq, pressure ruq, skin, AMA (antimitochondrial autoantibodies), le_cells, joints, pain, edema, bleeding, flatulence, ascites, hepatomegaly, density, nausea, spleen, consciousness, spider angioma, jaundice, EDGE (endoscopic ultrasound directed transgastric ERCP), irregular_liver, hbc_anti, hcv_anti, palm, hbeag	Alcoholism, hepatotoxic, Thepatitis, gallstones, PBC (Primary biliary cholangitis), fibrosis, diabetes, obesity, steatosis, Hyperbilirubinemia, RHepatitis, encelopathy, hepatalgia, anorexia, carcinoma	Hospital, surgery, choledocholithotomy, injections, transfusion,
Categorical		ChHepatitis (persistent/absent/active), Cirrhosis (absent, compensate, decompensate)	
Nominal	Age, triglycerides, birilubin, phosphatase, proteins, platelet, inr, urea, ESR (erythrocyte sedimentation rate), alt (alanine transaminase), ast (Aspartate aminotransferase), amylase, ggtp (Gamma-glutamyltranspeptidase), cholesterol, hbsag (Hepatitis B surface antigen), hbsag_anti, albumin		

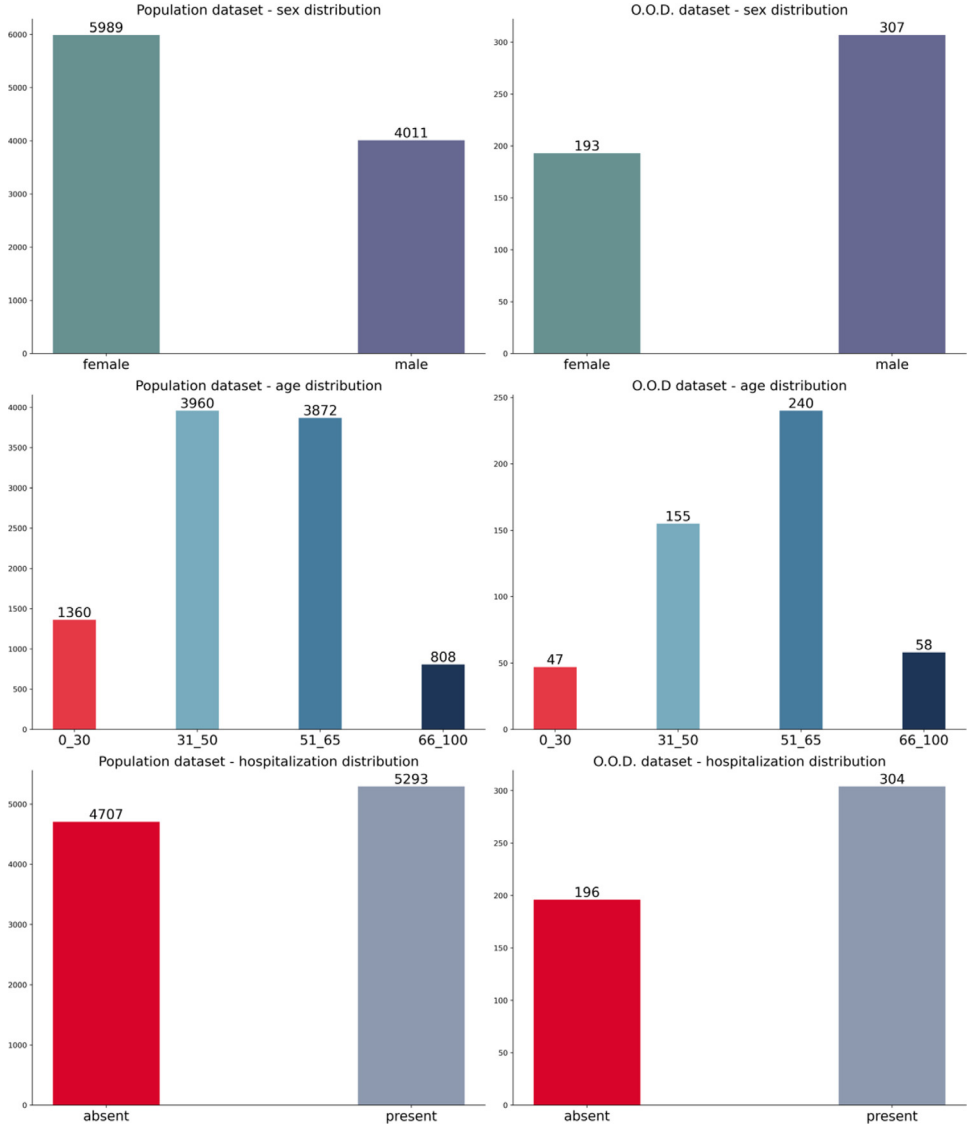

Fig. 1. Number of synthetic patients in the 2 different datasets, stratified for sex, age and hospitalization.

Fig. 1. shows the number of synthetic patients stratified for sex, age and hospitalization, both in the *Population dataset* and in the *O.O.D. dataset*.

3. Experimental Design, Materials and Methods

To acquire these data, we used the bnlearn package in R. We downloaded the Hepar2.rda file from the bnlearn repository (https://www.bnlearn.com/bnrepository/discrete-large.html#hepar2). Once loaded, the structure contains the list of nodes and arcs of the Bayesian network developed by [1], as well as all the prior probabilities. To generate the synthetic data,

we use the function rbn(), whose input are the loaded network structure and the number of samples to generate. In this case, we set the number of samples as 10,000 to generate an initial set of patients ("*population dataset*") (detailed code at https://github.com/GiovannaNicora/hepar_simulation/blob/main/hepar_simulate_patients.R). Subsequently, with a modified network with different prior probability distribution, we generated a second set of synthetic patients "out-of-distribution" ("*O.O.D dataset*") with respect to the "*population dataset*". To generate the "*O.O.D. dataset*", we changed the prior probabilities of sex, age and hospitalization (see the script https://github.com/GiovannaNicora/hepar_simulation/blob/main/hepar_simulate_patients.R). In particular, in our experiment we changed the prior probability distribution of sex, from $p(female) = 0.6)$ to $p(female) = 0.4$ (and subsequently, $p(male) = 0.6)$). The "hospitalization" distribution was changed as well: the initial $p(hospitalization = absent) = 0.65$ was changed to 0.35. Finally, the age probability distribution was set as $p(age\ between\ 0\ and\ 30) = 0.1$, $p(age\ between\ 31\ and\ 50) = 0.3$, $p(age\ between\ 51\ and\ 65) = 0.5$ and $p(age\ between\ higher\ than\ 65) = 0.1$. We then saved the resulting data into csv files. By changing parameters, such as number of patients or prior probabilities, users can repeat the process in order to simulate their own custom cohort of liver disorder patients.

The obtained datasets were used to train and test machine learning models as we described in [6]. We selected "hospitalization" as desired outcome, i.e. the class to predict. Briefly, we select the 10,000 patients as our "true population" and we sampled 50% of these patients. Of these, 53% were hospitalized at least once. We then selected 70% of these 50,000 patients for training and parameter tuning, while the remaining was kept as "inner distributed" test set. The *O.O.D. dataset* of 500 patients was used as additional O.O.D test set. In [6] we show how the performance of different classifiers vary between identically distributed (i.i.d.) dataset and O.O.D. sets, and we show that explanations built by our XAI method are impacted under the effects of dataset shift, i.e. more unstable when evaluated on unreliable O.O.D. samples.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: GN is a full employee of enGenome srl.

Data Availability

A synthetic dataset of Liver Disorder patients (Original data) (Zenodo).

CRediT Author Statement

Giovanna Nicora: Conceptualization, Methodology, Software; **Tommaso Mario Buonocore:** Data curation, Methodology, Writing – review & editing; **Enea Parimbelli:** Investigation, Writing – review & editing.

References

[1] A. Oniśko, M.J. Druzdzel, H. Wasyluk, A probabilistic causal model for diagnosis of liver disorders, in: Proceedings of the Seventh International Symposium on Intelligent Information Systems (IIS–98), Malbork, Poland, June 15–19, 1998, pp. 379–387

[2] M. Scutari, Learning Bayesian networks with the bnlearn R package, J. Stat. Softw. 35 (2010) 1–22, doi:10.18637/jss.v035.i03.

[3] G. Briganti, O. Le Moine, Artificial intelligence in medicine: today and tomorrow, Front. Med. (Lausanne) 7 (27) (2020) Accessed: Oct. 28, 2022. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fmed.2020.00027 , doi:10.3389/fmed.2020.00027. PMID: 32118012; PMCID: PMC7012990.

[4] N. Peek, C. Combi, R. Marin, R. Bellazzi, Thirty years of artificial intelligence in medicine (AIME) conferences: a review of research themes, Artif. Intell. Med. 65 (1) (2015) 61–73, doi:10.1016/j.artmed.2015.07.003.
[5] G. Nicora, M. Rios, A. Abu-Hanna, R. Bellazzi, Evaluating pointwise reliability of machine learning prediction, J. Biomed. Inform. (2022) 103996, doi:10.1016/j.jbi.2022.103996.
[6] E. Parimbelli, T.M. Buonocore, G. Nicora, W. Michalowski, S. Wilk, R. Bellazzi, Why did AI get this one wrong? — tree-based explanations of machine learning model predictions, Artif. Intell. Med. 135 (2023) 102471 ISSN 0933-3657, doi:10.1016/j.artmed.2022.102471.