Generation of virtual patient data for in-silico cardiomyopathies

Vasileios C. Pezoulas, Student Member, IEEE, Grigoris I. Grigoriadis, Nikolaos S. Tachos, Member IEEE, Fausto Barlocco, Iacopo Olivotto, and Dimitrios I. Fotiadis, Fellow, IEEE

Abstract—In-silico clinical platforms have been recently used as a new revolutionary path for virtual patients (VP) generation and further analysis, such as, drug development. Advanced individualized models have been developed to enhance flexibility and reliability of the virtual patient cohorts. This study focuses on the implementation and comparison of three different methodologies for generating virtual data for in-silico clinical trials. Towards this direction, three computational methods, namely: (i) the multivariate log-normal distribution (log-MVND), (ii) the supervised tree ensembles, and (iii) the unsupervised tree ensembles are deployed and evaluated against their performance towards the generation of high-quality virtual data using the goodness of fit (gof) and the dataset correlation matrix as performance evaluation measures. Our results reveal the dominance of the tree ensembles towards the generation of virtual data with similar distributions (gof values less than 0.2) and correlation patterns (average difference less than 0.03).

Keywords: virtual population generation, cardiomyopathy, in-silico clinical trials, log-MVND, tree ensembles.

I. INTRODUCTION

Virtual population (VP) development is a very popular and emerging aspect of healthcare technology, where the recent computational advances have shed light into the impact of this rapidly evolving biomedical field. According to the literature, many studies refer to the use of VP models towards the 3D visualization of human patients which could be potentially used to train clinicians. On the other hand, other applications focus on the design of geometrical models using VP models, which can be reconstructed from the virtual data to extract valuable clinical information in order to represent big varied cohorts depending on the disease under investigation.

In this study, we refer to virtual population generation as a method towards the generation of virtual patient data for clinical trials where mathematical models are used to describe and predict a patient’s progression based on an existing therapy. Such virtual patient generation models are used to enhance the population size of clinical trials for numerous applications, including effective drug development and robust decision making in clinical trials in terms of clinical trial simulations (CTS). Knowledge mining from these models leads to the reduction, refinement and partial substitution of the animal and human experimentation for drug development. The definition of a CTS though assumes the reproduction of the clinical trial design, the drug efficacy and the disease progression. A difficult issue of this field is the quality of the provided data. Towards this direction, data curation methods can be used to resolve any errors that are present within the data (e.g., outliers, data inconsistencies) and thus yield high quality data for more accurate virtual population generation.

The major baseline approaches for VP generation can be grouped in two categories, namely the parametric which are used to resample the real data towards the generation of synthetic data and the non-parametric where virtual patients are produced by randomly selecting patients from a real clinical dataset [1]. The multivariate normal distribution (MVND) method is a parametric approach which has been proposed by Tanenbaum [1] towards the generation of a target group of virtual patients according to realistic covariates. Teutonio et al. [2] evaluated and compared the MVND approach using resampling. Furthermore, some other studies that used different techniques are worth to be mentioned due to their increased performance. Regnier et al. [3] compared three different approaches to generate VPs that would match the real distributions from clinical cohorts. Moreover, Allen et al. [4], introduced a novel technique to generate VPs and demonstrated an efficient approach for VP generation without the need for weighting. Apart from the statistical methods, more straightforward methods, such as, the tree ensembles have been recently proposed by Robnik-Sikonja [5] towards the generation of robust semiartificial data, however, without any reported application for in-silico clinical trials.

In this work, we deploy three computational methods to generate virtual patient data from real clinical data. We extend a previous study [6], where the MVND method was used to generate 300 virtual patients by comparing the multivariate log-normal distribution with the tree ensembles to generate 1000 virtual patients for in-silico clinical trials targeting the drug development for familiar cardiomyopathies (FCM).

*This project has received funding from the European Union’s Horizon 2020 research and innovation programme under agreement No 777204. This paper reflects only the author’s view and the Commission is not responsible for any use that may be made of the information it contains. The data are from patients of the Cardiomyopathies Unit, Careggi University Hospital, Florence.

V.C. Pezoulas, G.I. Grigoriadis and N.S. Tachos are with the Unit of Medical Technology and Intelligent Information Systems, Dept. of Materials Science and Engineering, University of Ioannina, GR 45110 Ioannina, Greece (e-mails: bpezoulas@gmail.com, greg@grigorriadis@gmail.com, ntachos@gmail.com).

D.I. Fotiadis is with the Department of Biomedical Research, Institute of Molecular Biology and Biotechnology, FORTH, Ioannina, Greece and the Dept. of Materials Science and Engineering, Unit of Medical Technology and Intelligent Information Systems, University of Ioannina, GR 45110, Ioannina, Greece (e-mail: fotiadis@ce.aoi.gr).

F. Barlocco and I. Olivotto are with the Department of Experimental and Clinical Medicine, University of Florence and Cardiomyopathies Unit, Azienda Ospedaliera Careggi, Florence, Italy (e-mails: fausto.barlocco@unifi.it, iacopo.olivotto@gmail.com).
These methods have been integrated into the in-silico clinical trial SILICOFCM cloud-based platform [7]. The latter incorporates straightforward simulation tools for FCM drug development. Our results confirm the dominance of the tree ensembles as a prominent method for virtual population generation yielding an increased level of agreement between the real and the virtual data with goodness of fit (gof) values less than 0.2 (empirical threshold) and correlation patterns highly similar to the real patient data (2.71% difference in the average correlation values).

II. MATERIALS AND METHODS

B. Data sharing

Anonymized data were obtained from 1227 patients at two timepoints (2454 records in total) under the SILICOFCM project [7]. The dataset included 29 features (17 discrete, 12 continuous) related to demographics (e.g., age, gender), laboratory measures (e.g., diastolic pressure), and gene-related information (e.g., ACTC1, CSRP3). All the data came from the same centre, the Cardiomyopathies Unit at Careggi Hospital, Florence, and were collected by a very limited number of clinicians over more than 20 years.

B. Data quality evaluation

A data quality control pipeline presented in a previous study [8,9] was applied on the clinical data to deal with outliers, incompatible fields, duplicated features and missing values, where the Interquartile range (IQR) method was used to detect outliers and the Jaro distance [10] for lexical similarities. Features with more than 50% missing values, outliers, inconsistent fields were ignored from the analysis.

C. Methods for generating virtual data

1) Multivariate log-normal distribution

The multivariate normal distribution (MVND) [1,2] was adopted as a standard approach towards the generation of virtual patient data. Assuming a set of p features, say \( x = \{x_1, x_2, ..., x_p\} \), the MVND is defined as:

\[
f(x) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} e^{-(x-\mu)^T \Sigma^{-1} (x-\mu)/2},
\]

where \( p \) is the dimension, \( \mu \) is the mean vector, \( \Sigma \) is the covariance matrix, and \( \Sigma^{-1} \) is the pseudoinverse of the covariance matrix which is estimated through singular value decomposition (SVD). The goal of the MVND is to construct a multi-dimensional normal distribution given the mean vector, \( \mu \) and the covariance matrix, \( \Sigma \). In an attempt to strengthen the assumption of the normal distribution among the real data, we generated samples that follow a log-normal distribution, where the exponential of (1) was used to generate virtual data so that:

\[
ln(e^{f(x)}) \sim N(\mu, \Sigma),
\]

The generated values from \( e^{f(x)} \) were finally log-normally transformed and compared with the original ones.

2) Supervised tree ensembles

Another straightforward approach for virtual population generation is to train tree ensembles [11] given a set of training features along with a target feature. The virtual population generation problem can be treated as a supervised learning problem, where the tree ensembles are trained on 50% of the real data and tested on the remaining subset. The classifier can then be turned into a virtual data generator to generate semi-artificial data that follow the same distribution as the original data [11]. During the training process, the Gini impurity index [12] is used to measure the probability of a variable being classified in the wrong class as in:

\[
l(p) = 1 - \sum_{i=1}^{n} p_i^2,
\]

where \( p_i \) is the probability of a sample falling in class \( i \in \{1,2, ..., k\} \), where \( k \) is the total number of classes (\( k = 2 \) for a binary target feature). The Semiartificial package in R [11] was used to create a tree ensemble.

3) “Unsupervised” tree ensembles

A similar and rather straightforward approach for virtual population generation is to see the virtual population as a regression problem using tree ensembles [11]. The regressor can then be turned into a virtual data generator to generate semi-artificial data that follow the same distribution as the original data [11]. During the training process, the Mean Square Error (MSE) [11] is used to measure the difference between the real and the virtual distributions. For two variables (features), assume \( x_r \) and \( x_v \), in the real and virtual distributions, respectively, the MSE is given as:

\[
MSE = \frac{1}{n} \sum_{i=1}^{n} (x_{r,i} - x_{v,i})^2,
\]

The Semiartificial package in R [11] was used to construct a tree ensemble which was used as virtual generator.

D. Evaluation measures

1) Goodness of fit (gof)

The Kolmogorov-Smirnoff goodness of fit (gof) statistical test [13] is widely used in VP studies to measure the level of agreement between the real and the virtual distributions. The gof test measures whether the virtual data instances and the real data instances come from the same distribution. More specifically, the gof test statistics, say \( g \), is given as:

\[
g = \max \{|f_r(x) - f_v(x)|\},
\]

where \( f_r(x) \) and \( f_v(x) \) are the empirical distribution functions of the original and virtual data, respectively. Large gof values denote distributions with large vertical distance between them, whereas small gof values denote distributions with small vertical distances and thus similar.

2) Correlation

To further evaluate the similarity of the real and virtual data we computed the Pearson’s correlation coefficient between each pair of features in the virtual data and the real data. For two variables (features), assume \( x_r \) and \( x_v \), in the real and virtual distributions, respectively, the Pearson’s correlation coefficient [14], assume \( r \), is given as:

\[
r = \frac{E[(x_r - \mu_{x_r})(x_v - \mu_{x_v})]}{\sigma_{x_r} \sigma_{x_v}},
\]

5344
where $\mu_{x_r}, \mu_{y_r}$ are the mean values and $\sigma_{x_r}, \sigma_{y_r}$ are the standard deviations of $x_r$ and $y_r$, respectively, and $E[.]$ is the expected value. An $r$ value 1 denotes a perfect positive correlation between $x_r$ and $y_r$, whereas an $r$ value of 0 denotes no correlation.

### III. Results

#### A. Data quality assessment

The total number of missing values within the data was 10.12% (13 features had less than 50% missing values and 16 features had no missing values at all). Neither outliers nor duplicated features or inconsistent fields were detected in the anonymized data.

#### B. Virtual population generation

The clinical experts examined the resulting curated dataset and selected a subset of 10 features for generating 1000 virtual patients. The subset of features consists of the following features: “LVOTO_Rest” (Left ventricular outflow tract obstruction during resting state) (mean = 19.79, std = 21.79), “Evel” (E wave velocity) (mean = 74.8, std = 22.46), “lat_Eprime” (lateral e’ wave) (mean = 9.8, std = 3.02), “sep_Eprime” (septal e’ wave) (mean = 6.86, std = 2.14), “LA” (Left Atrium) (mean = 44.23, std = 7.39), “LVEF” (Left Ventricular Ejection Fraction) (mean = 64.25, std = 8.6), “Max_LVT” (maximum Left Ventricular Thickness) (mean = 19.24, std = 5.88), “PW” (mean = 10.68, std = 2.08), “NYHA” (New York Heart Association class) (mean = 1.65, std = 0.71) and “Age” (mean = 51.33, std = 18.19).

The performance evaluation results for each type of virtual population generation method are presented in Table 1, where the gof values where the smallest for the features “Max_LVT” and “Age” in the log-MVND, for the features “sep_Eprime”, “LA”, “PW”, and “Age” in the supervised tree ensembles and for the remaining features in the unsupervised tree ensembles.

The distribution of the gof values for each type of virtual population generation method is depicted in Figure 1 for the set of features which is presented in Table 1.

#### Table 1: Performance evaluation results for each type of virtual population generation method.

| Feature      | Performance evaluation measures | Multivariate log-normal distribution | Supervised tree ensembles |
|--------------|----------------------------------|--------------------------------------|---------------------------|
|              | mean    | std    | gof*     | LVOTO_Rest | Evel | lat_Eprime | sep_Eprime | LA | LVEF | Max_LVT | PW | NYHA | Age |
| LVOTO_Rest   | 22.93   | 16.76  | 0.28     |            | 19.85| 26.52      | 0.19       |    |      |        |    |      | 52.03|
| Evel         | 75.10   | 22.36  | 0.18     |            | 78.25| 31.71      | 0.15       |    |      |        |    |      | 10.04|
| lat_Eprime   | 9.79    | 3.04   | 0.25     |            | 9.63 | 3.64       | 0.19       |    |      |        |    |      | 1.51 |
| sep_Eprime   | 6.89    | 2.09   | 0.24     |            | 6.56 | 2.74       | 0.24       |    |      |        |    |      | 0.62 |
| LA           | 43.89   | 7.30   | 0.19     |            | 42.97| 7.26       | 0.14       |    |      |        |    |      | 0.17 |
| LVEF         | 64.08   | 8.69   | 0.18     |            | 64.28| 7.54       | 0.07       |    |      |        |    |      | 0.13 |
| Max_LVT      | 19.08   | 6.00   | 0.10     |            | 17.70| 4.73       | 0.13       |    |      |        |    |      | 0.16 |
| PW           | 10.51   | 1.99   | 0.19     |            | 10.39| 2.03       | 0.17       |    |      |        |    |      | 0.17 |
| NYHA         | 1.63    | 0.76   | 0.27     |            | 1.51 | 0.62       | 0.07       |    |      |        |    |      | 0.06 |
| Age          | 51.49   | 18.36  | 0.04     |            | 52.03| 17.59      | 0.04       |    |      |        |    |      |      |

*The smallest gof values on each method are filled with gray color.

Figure 1: Distribution plots for the real (upper panel; blue color) and the virtual data (lower panel; light blue: log-MVND, green: supervised tree ensembles, orange: unsupervised tree ensembles).

The correlation matrix of the real data is depicted in Figure 2 whereas the correlation matrix for the virtual data that were derived by the unsupervised tree ensembles (as the method that achieved the highest number of “optimal” gof values) is depicted in Figure 3, showing similar association patterns within the data. High correlation values are depicted in deep blue color whereas low correlation values are depicted in yellow. Since this is a qualitative way to view the correlation between the features in the real and the virtual populations, we have also computed the absolute value of the difference between the average correlation values from the upper (or
lower) triangular part of the matrices, for quantitative purposes. The difference in the average correlation values was 4.45% for the log-MVND, 8.28% for the supervised tree ensembles and 2.71% for the unsupervised tree ensembles.

The proposed methods could potentially provide significant insight in the field of virtual population generation to re-adjust the perspective of Clinical Trials (CTs) in other domains. As a future work, we also plan to deploy artificial neural networks (ANNs) that make use of radial basis functions (RBFs) as activation functions towards the generation of even more robust clinical data for in-silico clinical trials.

REFERENCES

[1] S. J. Tannenbaum, N. H. G. Holford, H. Lee, C. C. Peck, and D. R. Mould, “Simulation of Correlated Continuous and Categorical Variables using a Single Multivariate Distribution,” Journal of Pharmacokinetics and Pharmacodynamics, vol. 33, no. 6, pp. 773–794, Nov. 2006, doi: 10.1007/s10928-006-9033-1.

[2] D. Teuttonico et al., “Generating Virtual Patients by Multivariate and Discrete Re-Sampling Techniques,” Pharmaceutical Research, vol. 32, no. 10, pp. 3228–3237, Oct. 2015, doi: 10.1007/s11095-015-1699-x.

[3] T. R. Rieger et al., “Improving the generation and selection of virtual patients in quantitative systems pharmacology models,” Progress in Biophysics and Molecular Biology, vol. 139, pp. 15–22, Nov. 2018, doi: 10.1016/j.pbiomolbio.2018.06.002.

[4] R. Allen, T. Rieger, and C. Musante, “Efficient Generation and Selection of Virtual Populations in Quantitative Systems Pharmacology Models: Generation and Selection of Virtual Populations,” CPT: Pharmacometrics & Systems Pharmacology, vol. 5, no. 3, pp. 140–146, Mar. 2016, doi: 10.1002/pspd.12063.

[5] M. Robnik-Šikonja, “Dataset comparison workflows,” p. 20.

[6] V. Pezoulas, N. Tachos, and D. Fotiadis, “Generation of Virtual Patients for in-silico Cardiomyopathies Drug Development,” in 2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE), Athens, Greece, 2019, pp. 671–674, doi: 10.1109/BIBE.2019.00126.

[7] “Home - SILICOFCM.” [Online]. Available: https://silicofcm.eu/.

[8] V. C. Pezoulas, K. D. Kourou, F. Kalatzis, T. P. Exarchos, A. Venetsanopoulou, E. Zampeli, S. Gandolfi, F. Skopoulis, S. De Vita, A. G. Tzioufas, and D. I. Fotiadis, “Medical data quality assessment: On the development of an automated framework for medical data curation,” CBM, vol. 107, pp. 270–283, 2019.

[9] V. C. Pezoulas, K. D. Kourou, F. Kalatzis, T. P. Exarchos, A. I. Venetsanopoulou, E. Zampeli, S. Gandolfi, F. N. Skopoulis, S. De Vita, A. G. Tzioufas, and D. I. Fotiadis, “Enhancing medical data quality through data curation: a case study in primary Sjögren’s syndrome,” Clinical and experimental rheumatology, vol. 37, no. 3, pp. 90–96, 2019.

[10] K. International Conference on Advances in Databases and Data Application and F. Laux, DBKDA 2015 the Seventh International Conference on Advances in Databases, Knowledge, and Data Applications, Rome, Italy, May 24 - 29 2015. S. l.: s. n., 2015.

[11] M. Robnik-Sikonja, M. Radovíc, S. Drorövić, B. Andjelkovic Cirkovic, and N. Filipovic, Modeling ischemia with finite elements and automated machine learning. 2018.

[12] V. C. Pezoulas, T. P. Exarchos, and D. I. Fotiadis, “Medical data sharing, harmonization and analytics,” Academic Press, 2020. Paperback ISBN: 9780128165072, eBook ISBN: 9780128165591.

[13] R. B. D’Agostino and M. A. Stephens, Eds., Goodness-of-fit Techniques. New York, NY, USA: Marcel Dekker, Inc., 1986.

[14] P. Schober, C. Boer, and L. A. Schwarte, “Correlation coefficients: appropriate use and interpretation,” Anesthesia & Analgesia, vol. 126, no. 5, pp. 1763–1768, 2018.