

Generation of virtual patients for *in silico* cardiomyopathies drug development

Vasileios C. Pezoulas
Unit of Medical Technology and
Intelligent Information Systems,
University of Ioannina,
GR 45100 Ioannina, Greece
bpezoulas@gmail.com

Nikos S. Tachos
Unit of Medical Technology and
Intelligent Information Systems,
University of Ioannina,
GR 45100 Ioannina, Greece
ntachos@gmail.com

Dimitrios I. Fotiadis
Unit of Medical Technology and
Intelligent Information Systems,
University of Ioannina,
GR 45100 Ioannina, Greece
fotiadis@cc.uoi.gr

Abstract— The revolution in modelling and simulation methodologies accompanied with the recent events in the high-performance computing (HPC) helped the development of *in silico* clinical platforms which integrate advanced and individualized simulation models to support drug development. These platforms incorporate patient specific models to create and generate virtual patients (VPs). A parametric methodology for resampling and generating VPs is the multivariate normal distribution which in the current work is optimized through an iterative pipeline by the Kolmogorov-Smirnov goodness-of-fit test. The proposed VP generator is integrated in the multi-repository VP model of SILICOFCM which is a multi-modular, innovative *in silico* clinical trials solution for the design and functional optimization of the whole heart performance and monitoring effectiveness of pharmacological treatment for familial cardiomyopathies, with aim to reduce the animal studies and the human clinical trials.

Keywords— *Virtual Patients, Cardiomyopathy, In Silico Clinical Trial, Multivariate Normal Distribution*

I. INTRODUCTION

Advances in information and communication technology includes all science pillars, including the life sciences. Lately, the revolution in modelling and simulation methodologies accompanied with the recent events in the high-performance computing (HPC) industry have provided significant new understandings in the biomedicine science and technology. Exploiting the latter helps to reduce, refine and partial substitute the animal and human experimentation in the drug development pipeline. Based on this fact, European and United States Regulatory bodies allow the use of simulation and modelling in various phases during the development of new drug and therapy [1]. In this context, the clinical trial simulation (CTS) is a useful and important tool to support decision making in the drug development roadmap and the clinical trial failure [2], [3]. Specifically, the term “*in silico* clinical trials” refers to “The use of individualized computer simulation in the development or regulatory evaluation of a medicinal product, medical device, or medical intervention” [4]. Specifically, the latter includes the development of patient specific models to create and generate virtual patients (VPs) which form the virtual cohorts. These are exploited by the advanced simulation models to test the safety and efficacy of a new drug.

The generation of VPs includes two categories of methodologies the non-parametric and the parametric ones [5]. The first one is a straightforward method of resampling, where a group of virtual patients is produced by randomly selecting patient arrays along with their features from a real clinical dataset[6], [7]. Knowledge based inclusion criteria are then applied to form the refined VP data pool. The second

methodology, the parametric methods, are used to resample and generate new combinations of matrices from an existing clinical feature set. The latter uses a unique multivariate distribution (MVND) for the whole dataset. Through this methodology the correlation between the features is taking into account which leads to a realistic generation of VPs compared to the real ones. The MVND methodology was proposed by Tannenbaum et al. [8], where both continuous and categorical features are treated as continuous variables and it was tested in a dataset of 467 patients including seven continuous features (age, weight, body mass index, diastolic and systolic blood pressure, total cholesterol, fasting blood glucose) and three categorical (sex, smoking status, and diagnosis) achieving good results.

In the current work, we present a VP generator from a real clinical dataset based on the parametric MVND methodology optimized by an iterative process through the Kolmogorov-Smirnov goodness-of-fit test. The developed VP generator is integrated into the multi-repository virtual population model of SILICOFCM [9] which is an advanced cloud based *in silico* platform offering simulation capabilities and advanced tools for testing and development of drugs targeting the familial cardiomyopathies (FCM).

II. MATERIALS AND METHODS

A. Workflow

The overall workflow of the proposed approach for virtual population generation is depicted in Fig. 1. The retrospective data are first uploaded into a secure database which is part of the SILICOFCM cloud-based platform through a template .csv file. A fully automated data quality control pipeline is then applied on the uploaded data to deal with outliers, incompatible fields, unknown data types and missing values. The features are categorized into “good” or “bad” and the latter are ignored from further analysis to yield the curated dataset. The curation process is repeated until the data quality process fulfills standard quality criteria which are defined by the clinical experts in the field. Then, the virtual population pipeline is applied on a specific subset of features from the curated data, stored into the SILICOFCM repositories, to generate the virtual data. The subset of features is defined by the expert end-user and the multivariate normal distribution approach based on goodness-of-fit optimization is applied on the selected subset of features to yield high quality virtual patient data for *in silico* clinical trials. The output plausible patients are assessed by the expert end-user through the provided visualization and reporting tools and if accepted the VP is stored into the repositories accompanied with rich metadata. The generated virtual population is utilized by the

provided simulation and analysis tools included in the SILICOFCM cloud *in silico* platform.

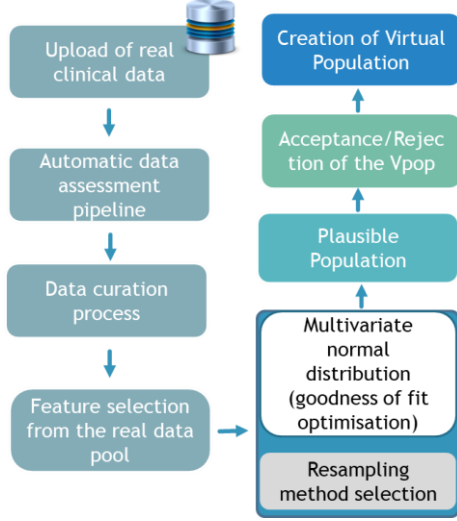


Figure 1. The proposed workflow for the virtual population generation.

C. Data quality control pipeline

A data quality control pipeline [10], [11] depicted in Fig. 2 was applied to enhance the quality of the retrospective data by detecting outliers (i.e., values that deviate from the standard population distribution), incompatible fields (e.g., inconsistent formats), unknown data types (e.g., mixed integer and string data types) and any other problematic fields that are present within the data. More specifically, the data are first annotated according to their value range and type (e.g., integer, string, float) and then univariate methods, such as, the interquartile range [12], [13] and the average/most frequent replacement [14] are used to detect outliers and impute missing values, respectively. The features are categorized into good and bad according to their overall quality state. The bad features are then automatically removed from the data. Furthermore, quality reports are provided to the clinical experts summarizing useful metadata on a feature basis (e.g., value range, data type, warnings).

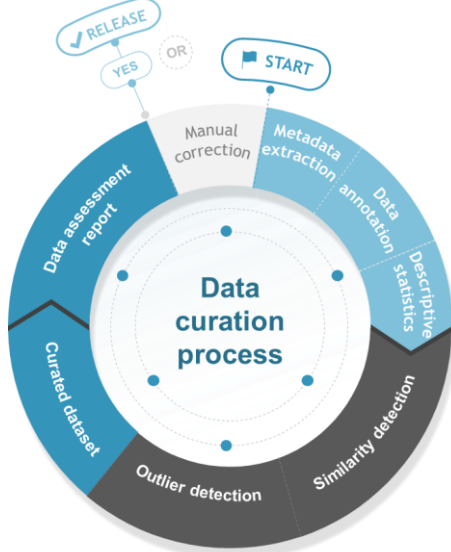


Figure 2. The proposed workflow for the data quality assessment.

C. Multivariate normal distribution

The multivariate normal distribution (MVND) strategy [8], [15] was adopted as a standard approach towards the generation of high-quality virtual patient data for *in silico* trials. For a univariate feature, say x , the normal distribution is given as:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma^2} e^{-[(x-\mu)/\sigma^2]^2/2}, \quad (1)$$

where μ is the mean and σ^2 is the variance. The exponential term in (1) can be written as:

$$\frac{x - \mu}{\sigma^2} = (x - \mu)^T (\sigma^2)^{-1} (x - \mu), \quad (2)$$

to enable its generalization in the p -th dimension as:

$$(x - \mu)^T (\Sigma)^{-1} (x - \mu), \quad (3)$$

which is in fact the Mahalanobis distance. Assuming a set of p -features, say $\mathbf{x} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_p\}$, the MVND is defined as:

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} e^{-(\mathbf{x}-\mu)\Sigma^{-1}(\mathbf{x}-\mu)/2}, \quad (4)$$

where p is the dimension, μ is the mean vector, Σ is the covariance matrix, and Σ^{-1} is the pseudoinverse of the covariance matrix which is estimated through singular value decomposition (SVD). The goal of the MVND in virtual population generation is to construct a multi-dimensional normal distribution given the mean vector and the covariance matrix which serves as the virtual distribution.

D. Kolmogorov-Smirnov goodness-of-fit

In an attempt to control for the “randomness” of the virtual data and at the same time maximize the level of “agreement” between the virtual distribution and the original one, the Kolmogorov-Smirnov (KS) goodness of fit (gof) test [16] was used to test for the null hypothesis that the distribution of the generated virtual data and the distribution of the original data come from the same distribution. More specifically, the gof test statistics, say D , is given by:

$$D = \max(|F_o(x) - F_v(x)|), \quad (5)$$

where $F_o(x)$ and $F_v(x)$ are the empirical distribution functions (EDFs) of the original and virtual populations, respectively. In fact, the gof measures whether $F_o(x)$ and $F_v(x)$ are similar by calculating the largest distance, D , between the two EDFs. If D is larger than a critical value then the null hypothesis is rejected at the given confidence level. As a matter of fact, a large gof value between $F_o(x)$ and $F_v(x)$ denotes distributions with large vertical distance and thus the null hypothesis is rejected. On the other hand, small gof values denote small vertical distances between $F_o(x)$ and $F_v(x)$ and thus similar distributions. In order to do so, the MVND process was re-applied until the overall gof of the distribution was less than a pre-defined threshold, say t . The proposed approach was compared with random executions of the MVND.

III. RESULTS

A. Data collection

Anonymized clinical data were obtained from 364 patients under the SILICOFM project which included 69 features in total [9].

B. Data curation

An instance of the curated clinical dataset is depicted in Fig. 3. According to Fig. 3, features having less than 50% missing values are depicted in blue color, features with no missing values are depicted in green color and features with more than 50% missing values are depicted in red color and are characterized as “bad” feature. The missing values are depicted in black color using the “NaN” flag and the outliers are depicted in orange. All “bad” features along with the outliers were automatically removed from further analysis and the missing values were replaced according to the mean/most frequent approach yielding the final dataset to be used for the virtual population generation.

ST_segment_abnormal	Negative_T_wave	LA	LAVs	MumaxPG	MVmeanPG
1	1	45	97	3.5	1.1
0	1	37	63	2	0.6
0	0	42	91	3.1	1.4
0	1	46	54	6	2.5
0	1	36	55	4.2	2.1
0	0	34	99	6.1	1.9
0	0	30	NaN	2.8	NaN
1	0	25	NaN	3.3	1.6
0	0	27	NaN	1.5	NaN
0	0	35	NaN	5.6	1.9
0	0	37	NaN	2.8	1.3
0	0	40	51	3.8	1.8
0	1	33	64	13.1	4.7
0	0	42	65	5.4	1.5
0	0	37	NaN	NaN	NaN
0	1	39	46	4.1	2.4
0	0	32	88	7.5	2.5
0	1	30	45	4.2	2.7
0	0	34	55	1.9	0.7

Figure 3. An instance of the curated dataset.

C. Virtual population generation

The clinical experts examined the resulting curated dataset and selected a subset of 10 features for generating 300 virtual patients. The subset of features includes the “BMI” (Body Mass Index), “age”, “sex”, “syncope”, “NYHA class”, “systolic”, “diastolic”, “heart murmur”, “LVIDs (Left Ventricle Internal Diameter in systole phase)”, and “LVIDd (Left Ventricle Internal Diameter in diastole phase)”. The mean vector and the covariance matrix of the original population were then estimated and provided as input into the MVND formula.

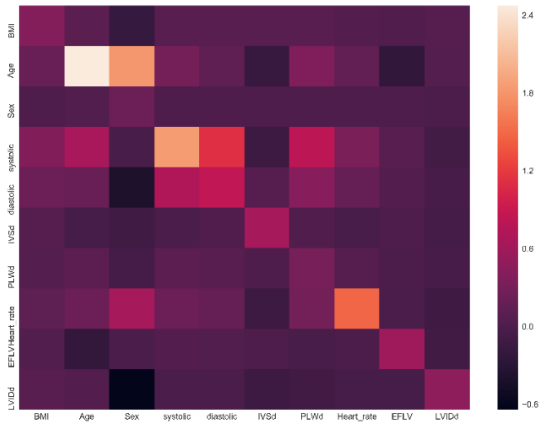


Figure 4. The covariance matrix of the real population.

The level of agreement between the multi-dimensional real distribution and the virtual one was controlled by the gof value. The 10x10 covariance matrix of the real dataset is

depicted in Fig. 4, where the non-diagonal cell (i, j) corresponds to the covariance between features i and j and the diagonal elements correspond to the variance of each feature. The pairs of features with high covariance are depicted in orange color whereas the pairs of features with dark color correspond to features which are independent.

The results of the virtual population generation are depicted in Fig. 5. In all cases, the gof values were less than (or equal to) 0.2 yielding virtual distributions similar to the real ones. The number of executions needed for the virtual population generation was approximately 5000 requiring a short amount of time (~5 sec) considering the number of virtual patients.



Figure 5. The histogram distribution between the features of the real (blue) and the virtual (red) populations.

D. Controlling the “randomness” of the virtual population

The proposed method is based on a recursive execution of the MVND formula using the gof measure to control for the randomness of the virtual distributions, where the number of iterations required to control it in the previous case was equal to 5000 runs. To demonstrate the value of the proposed approach towards the generation of high quality virtual data, the proposed method was evaluated against 10 random executions of the MVND without the gof factor as a criteria. According to Table I, the average gof values across 10 random runs are larger than 0.2 for the systolic, diastolic, and LVIDs (Left Ventricle Internal Diameter in systole phase) features, and increased for the BMI and LVIDd (Left

Ventricle Internal Diameter in dystole phase). Furthermore, in these cases, the deviation of the mean values is larger than in the original population.

Table I. Comparison results between the proposed method and 10 random executions of the MVND.					
Features	Mean/median			Goodness-of-fit	
	Real	MVND with gof	random runs	MVND with gof	random runs
BMI	27.35	27.22	27.36	0.131	0.16
Age	55.68	54.51	56.43	0.078	0.082
Sex	0	0	0	0.092	0.078
Syncope	0	0	0	0.017	0.017
NYHA class	1	1	1	0.109	0.119
Systolic blood pressure*	126.24	124.36	125.56	0.178	0.255
Diastolic blood pressure*	76.97	75.63	76.22	0.182	0.2
Heart murmur	0	0	0	0.044	0.06
LVIDs*	28.26	28.11	28.93	0.2	0.271
LVIDd	47.39	47.19	47.43	0.147	0.193
* features with significant differences between the two cases.					

IV. CONCLUSIONS

In this work, we presented a computational workflow for the generation of high-quality virtual clinical data for *in silico* clinical trials. Specifically, the proposed pipeline is integrated in the virtual patient model of SILICOFCM cloud based *in silico* platform. The latter offers advanced tools and computational solvers for drug development targeting the FCM disease.

A data quality control pipeline was applied on the anonymized clinical data from 364 patients yielding curated data with enhanced accuracy and completeness. A virtual population generation pipeline was then applied on the curated dataset to generate plausible virtual clinical data for 300 patients based on a recursive MVND approach which controls for the randomness of the generated distributions based on the Kolmogorov-Smirnov (KS) goodness-of-fit (gof) measure. Our results demonstrate the robustness and accuracy of the proposed method towards the generation of virtual clinical data for *in silico* clinical trials with high level of agreement between the densities and the distributions of the virtual and the real clinical datasets. The gof values of the proposed method were less than 0.2 compared to the average gof values obtained across 10 random executions (> 0.2 in some cases).

Although the number of iterations needed to control for the randomness of the generated virtual distribution was large enough (~5000), the execution time was small considering the number of virtual patients and the number of iterations needed to assess for the “randomness” factor. The lack of significant computational complexity followed by an increase in the quality of the virtual data is an advantage of the proposed methodology, especially in the case of large-scale *in silico* clinical trials where the number of virtual patients to be generated is significantly larger. As a future work additional methods for virtual population generation, such as the Bayesian networks [17], [18] and the modified genetic function [19], along with neural network-based strategies [20] will be employed for comparison.

V. ACKNOWLEDGEMENT

Research supported by the SILICOFCM project that has received funding from the European Union’s Horizon 2020 research and innovation program under grant 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.

REFERENCES

- [1] “in silico Clinical Trials: How Computer Simulation will Transform the Biomedical Industry,” Avicenna Coordination Support Action.
- [2] V. Zazzu, B. Regierer, A. Kühn, R. Sudbrak, and H. Lehrach, “IT Future of Medicine: from molecular analysis to clinical diagnosis and improved treatment,” *N Biotechnol*, vol. 30, no. 4, pp. 362–365, May 2013.
- [3] W. Alkema, T. Rullmann, and A. van Elsas, “Target validation in silico: does the virtual patient cure the pharma pipeline?,” *Expert Opin. Ther. Targets*, vol. 10, no. 5, pp. 635–638, Oct. 2006.
- [4] M. Viceconti, A. Henney, and E. Morley-Fletcher, “In silico clinical trials: how computer simulation will transform the biomedical industry,” *International Journal of Clinical Trials*, vol. 3, no. 2, pp. 37–46, May 2016.
- [5] J. Pilz, D. Rasch, V. B. Melas, and K. Moder, *Statistics and Simulation: IWS 8, Vienna, Austria, September 2015*. Springer, 2018.
- [6] S. Chabaud, P. Girard, P. Nony, J.-P. Boissel, and HERapeutic MOdeling and Simulation Group, “Clinical trial simulation using therapeutic effect modeling: application to ivabradine efficacy in patients with angina pectoris,” *J Pharmacokinet Pharmacodyn*, vol. 29, no. 4, pp. 339–363, Aug. 2002.
- [7] H. J. M. Lemmens, D. R. Wada, C. Munera, A. Eltahtawy, and D. R. Stanski, “Enriched analgesic efficacy studies: an assessment by clinical trial simulation,” *Contemp Clin Trials*, vol. 27, no. 2, pp. 165–173, Apr. 2006.
- [8] 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,” *J Pharmacokinet Pharmacodyn*, vol. 33, no. 6, pp. 773–794, Dec. 2006.
- [9] “SILICOFCM.” [Online]. Available: <https://silicofcm.eu/>.
- [10] V. C. Pezoulas *et al.*, “Medical data quality assessment: On the development of an automated framework for medical data curation,” *Comput. Biol. Med.*, vol. 107, pp. 270–283, Apr. 2019.
- [11] V. C. Pezoulas *et al.*, “AB0166 ENHANCING THE QUALITY OF CLINICAL DATA THROUGH DATA CURATION IN PRIMARY SJÖGREN’S SYNDROME,” *Annals of the Rheumatic Diseases*, vol. 78, no. Suppl 2, pp. 1541–1542, Jun. 2019.
- [12] S. S. Tripathy, R. K. Saxena, and P. K. Gupta, “Comparison of Statistical Methods for Outlier Detection in Proficiency Testing Data on Analysis of Lead in Aqueous Solution,” *American Journal of Theoretical and Applied Statistics*, vol. 2, no. 6, p. 233, Jan. 2014.
- [13] P. J. Rousseeuw and M. Hubert, “Robust statistics for outlier detection,” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 1, no. 1, pp. 73–79, 2011.
- [14] S. P. Mandel J, “A Comparison of Six Methods for Missing Data Imputation,” *J Biom Biostat*, vol. 06, no. 01, 2015.
- [15] D. Teutonico *et al.*, “Generating Virtual Patients by Multivariate and Discrete Re-Sampling Techniques,” *Pharm Res*, vol. 32, no. 10, pp. 3228–3237, 2015.
- [16] R. B. D’Agostino and M. A. Stephens, Eds., *Goodness-of-fit Techniques*. New York, NY, USA: Marcel Dekker, Inc., 1986.
- [17] V. Leclerc, M. Ducher, and N. Bleyzac, “Bayesian Networks: A New Approach to Predict Therapeutic Range Achievement of Initial Cyclosporine Blood Concentration After Pediatric Hematopoietic Stem Cell Transplantation,” *Drugs R D*, vol. 18, no. 1, pp. 67–75, Mar. 2018.
- [18] T. Haddad *et al.*, “Incorporation of stochastic engineering models as prior information in Bayesian medical device trials,” *Journal of Biopharmaceutical Statistics*, vol. 27, no. 6, pp. 1089–1103, Nov. 2017.
- [19] T. R. Rieger *et al.*, “Improving the generation and selection of virtual populations in quantitative systems pharmacology models,” *Prog. Biophys. Mol. Biol.*, vol. 139, pp. 15–22, 2018.
- [20] M. R. Šikonja, “Dataset comparison workflows,” *IJDS*, vol. 3, no. 2, p. 126, 2018.