Data in Brief 36 (2021) 107147

Contents lists available at ScienceDirect

Data in Brief

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

Data Article

Processed HIV prognostic dataset for control experiments

Moses E. Ekpenyong[a,b,*], Philip I. Etebong[a], Tenderwealth C. Jackson[c], Edidiong J. Udofa[c]

[a] *Department of Computer Science, University of Uyo, P.M.B. 1017 520003, Uyo, Akwa Ibom State, Nigeria*
[b] *Centre for Research and Development, University of Uyo, P.M.B. 1017 520003, Uyo, Akwa Ibom State, Nigeria*
[c] *Department of Pharmaceutics and Pharmaceutical Technology, University of Uyo, P.M.B. 1017 520003, Uyo, Akwa Ibom State, Nigeria*

A R T I C L E I N F O

Article history:
Received 21 April 2021
Revised 6 May 2021
Accepted 10 May 2021
Available online 14 May 2021

Keywords:
Adverse drug reaction
Antiretroviral therapy
HIV control data
Treatment change episode
Treatment failure classification

A B S T R A C T

This paper provides a control dataset of processed prognostic indicators for analysing drug resistance in patients on antiretroviral therapy (ART). The dataset was locally sourced from health facilities in Akwa Ibom State of Nigeria, West Africa and contains 14 attributes with 1506 unique records filtered from 3168 individual treatment change episodes (TCEs). These attributes include sex, before and follow-up CD4 counts (BCD4, FCD4), before and follow-up viral load (BRNA, FRNA), drug type/combination (DTYPE), before and follow-up body weight (Bwt, Fwt), patient response to ART (PR), and classification targets (C1-C5). Five (5) output membership grades of a fuzzy inference system ranging from very high interaction to no interaction were constructed to model the influence of adverse drug reaction (ADR) and subsequently derive the PR attribute (a non-fuzzy variable). The PR attribute membership clusters derived from a universe of discourse table were then used to label the classification targets as follows: C1=no interaction, C2=very low interaction, C3=low interaction, C4=high interaction, and C5=very high interaction. The classification targets are useful for building classification models and for detecting patients with ADR. This data can be exploited for the development of expert sys-

DOI of original article: 10.1016/j.imu.2021.100568
* Corresponding author.
E-mail address: mosesekpenyong@uniuyo.edu.ng (M.E. Ekpenyong).

https://doi.org/10.1016/j.dib.2021.107147

tems, for useful decision support to treatment failure classification [1] and effectual drug regimen prescription.

Specifications Table

Subject	Health and Medical Sciences
Specific subject area	Adverse Drug Reaction
Type of data	Table Figure
How data were acquired	Excavation and pre-processing Instruments: hardware, software, program Make and model and of the instruments used: hardware (Intel HP Core i5 8th Gen), software (Microsoft Excel, JuzzyOnline Fuzzy Toolkit)
Data format	Raw Analysed Filtered
Parameters for data collection	Prognostic indicators of HIV were excavated and analysed.
Description of data collection	Data of HIV patients were obtained directly from HIV patients' records distributed across different health facilities.
Data source location	Institution: University of Uyo City/Town/Region: Uyo/Akwa Ibom Country: Nigeria
Data accessibility	With the article
Related research article	[1] M.E. Ekpenyong, M.E. Edoho, I.J. Udo, P.I. Etebong, N.P. Uto, T.C. Jackson, N.M. Obiakor, A transfer learning approach to drug resistance classification in mixed HIV dataset, Informatics in Medicine Unlocked. 100,568. https://doi.org/10.1016/j.imu.2021.100568

Value of the Data

- This paper presents very useful datasets for engendering research on HIV/AIDS in the Sub-Saharan African region.
- Computer scientists can use the data to develop classification models and expert systems for drug pattern analysis, adverse drug reaction and failed treatment. Clinicians/physicians and pharmacists can use the developed expert system to support meaningful decisions on drug prescription, recommendation, and administration.
- By providing access to clinical (control) HIV data, research progress can be accelerated towards individualised medicine, where on-treatment variables influencing a set of study outcomes are analysed for the purpose of predicting patient drug response with precision.
- The developed models and algorithms could be made available as open-source tools with adaptive and replicable features for diverse domains/environments.

1. Data Description

We provide a control dataset (SupplFile.xlsx) containing average prognostic indicators of HIV (sex; before CD4 count, BCD4; follow-up CD4 count, FCD4; before viral load, BRNA; follow-up viral load, FRNA; before body weight in Kg, BWt; and follow-up body weight in Kg, FWt), treatment type/drug(s) combination (DrugNo/DrugComb) and patient response to treatment (PR). The dataset is divided into two sets, the individual treatment change episodes (TCEs) and unique records. The first set, the TCEs (or raw data) lists on each row repeated instances of other

Table 1
Drugs administered to patients on ART (https://hivdb.stanford.edu) [2].

DrugNo	DrugCode	DrugName	DrugNo	DrugCode	DrugName
1	RTV	Ritonavir	13	DDI	Didanosine
2	IDV	Indinavir	14	LPV	Lopinavir
3	D4T	Stavudine	15	APV	Amprenavir
4	3TC	Lamivudine	16	NVP	Nivarapine
5	SQV	Squatonavir	17	DRV	Darunavir
6	T20	Nfoviritide	18	FTC	Emtricitabine
7	FPV	Fosamprenavir	19	ATV	Atazanavir
8	NFV	Nelfinavir	20	TPV	Tipranavir
9	AZT	Zidovudine	21	RAL	Raltenovir
10	ABC	Abacavir	22	ETR	Etravirine
11	TDF	Tenofovir	23	MVC	Maraviroc
12	EFV	Efavirenz	24	DLV	Delavirdine

Table 2
Analysis of control datasets.

	Type of Control Dataset	
Analysis	Stanford	Akwa Ibom
Male	–	704
Female	–	352
Total number of drugs administered	24	5
Minimum drug combination	1	3
Maximum drug combination	7	3
Number of Patients with most frequent drug combinations (actual drug combination)	37 (D4T+DDI+EFV)	698 (TDF+3TC+EFV)
Number of Patients with less frequent drug combinations (actual drug combination)	1 (3TC)	27 (AZT+3TC+EFV)
Patients with at most 2 TCEs	31	0
Patients with at least 3 TCEs (Total TCEs)	1490 (5780)	1506 (3168)

variables, save the individual drugs (or DrugNo–a number or numeric value used to identify each drug taken by the patient) which are listed on separate rows for each patient ID (PID). Table 1 populates the corresponding drug code (DrugCode) and drug name (DrugName) of the respective DrugNo, for each drug administered to patients on ART. The prognostic indicators are results of laboratory analysis conducted using biological fluid sample (the blood), while sex and body weight are determined by physical appearance and measurement using scale reader, respectively. A total of 3168 TCEs are documented. The TCEs were further processed to achieve individual unique records of 1506 patients. The unique records are condensed instances of the TCEs, with DrugNo converted into its DrugCode equivalent and concatenated to form a single, unique record. The PR is a non-fuzzy output value obtained from a fuzzy inference system evaluation of the prognostic indicators with 5 output membership grades indicating the level of drugs interaction as follows (very high interaction, high interaction, very low interaction, low interaction, and no interaction). The classification targets (C1-C5) are binary digits (0/1) used to indicate or label the occurrence of a particular membership grade.

Important statistics revealing more insight into the control dataset compared with the Stanford dataset are as presented in Table 2.

2. Experimental Design, Materials and Methods

Locally sourced data were collected directly from case files of patients receiving treatment at various health centres in Akwa Ibom State of Nigeria, including a Community Anti-Retroviral Therapy Programme–periodically carried out to reach rural dwellers. A total of 13 health facilities

Table 3
Input and output fuzzy sets from domain knowledge.

S/N	Membership grade (MG)	l_1	P_1	r_1	l_2	P_2	r_2
		BCD4/FCD4 (Input)					
1	Low {L}	0	225	450	50	275	500
2	Medium {M}	300	575	850	350	625	900
3	High {H}	700	1075	1450	750	1125	1500
		BRNA/FRNA (Input)					
1	Undetected {U}	0	0.60	1.20	0.30	0.90	1.50
2	Supressed {S}	1.00	2.15	3.30	1.20	2.35	3.50
3	Not Supressed {NS}	2.50	4.00	5.50	3.00	4.50	6.00
		PR (Output)					
1	No Interaction {NI}	0	27.50	55	5	32.50	60
2	Very Low Interaction {VLI}	30	47.50	65	35	52.50	70
3	Low Interaction {LI}	62	68.50	75	67	73.50	80
4	High Interaction {HI}	72	78.50	85	77	83.50	90
5	Very High Interaction {VHI}	82	88.50	95	87	93.50	100

were used as data collection points and covers patients with both resistant and non-resistant cases who registered for treatment at the various facilities from 2015 to 2018. The investigated facilities were found to accommodate up to 10,000 patients receiving treatment in the southeast region. Due to limited resources and the high cost of treatment, only 5 drug combinations in 3 consistent treatment regimens were administered to patients free of charge, through a Family Health International (FHI) HIV/AIDS intervention programme. The number of row(s) rendered depend(s) on the patient's ART regimen administered over the treatment period. Hence, if a patient was administered a combination of 3 drugs, then, three rows are rendered (see data on individual TCEs).

Collection of the control data did not involve direct contact with the patients. Instead, access to patients' medical histories and treatment was granted by the responsible authorities after satisfying the ethical consent procedure required for the purpose of filtering the relevant data. At the University level, ethical approval was granted by the University of Uyo Institutional Health Research Ethics Committee (UNIUYO–IHREC). At the hospital level, Informed consent through written permission was obtained from the responsible health authority before embarking on the data collection. To protect patient records, details that could expose the patients' personal details (e.g., name, address, occupation, etc.) were not documented. Each patient data was further validated for consistency before recording, while questionable, inconsistent, or not properly documented records were dropped. The control dataset holds only first line treatment episodes (initial 6 months) excavated from existing patients' records/files under the supervision of a medical superintendent.

From the control dataset, universe of discourse (UoD) membership ranges, were created to align with established ranges from domain experts/physicians. Table 3 shows the input and output fuzzy sets derived from the control dataset.

The column labels indicate the internal structure of the IT2FS [3,4], where: l is the left end point bounded by both UMF (l_2) and LMF (l_1), and r is the right end point, also bounded by both UMF (r_2) and LMF (r_1). The triangular peak location or mean, P, of each end point is also bounded by P_1 and P_2, representing the triangular peak locations of end points l_1 and r_1, and l_1 and r_2, respectively. Expressions deriving the IT2FL LMF and UMF can be found in [5].

To enable precise knowledge representation of PR and minimise the influence of confusing input/output boundaries, an Interval Type-2 Fuzzy Logic (IT2FL) system was developed using the JuzzyOnline Fuzzy Toolkit (http://juzzy.wagnerweb.net/) [6,7] – an open-source toolkit for design, implementation, evaluation and sharing of Type-1 and Type-2 fuzzy logic systems. Applying Microsoft Excel functions and commands, the individual TCEs were condensed to produce a second set of data called unique records (a single row of patient record), with the DrugNo replaced with DrugCode and then concatenated with the '+' symbol to form the drug combination (DrugComb). Hence, if the following drugs were administered to a patient (3TC, ABC, AZT) over the study period, then the DrugComb cell is rendered as 3TC+ABC+AZT. Microsoft Excel command was also used to label the classification targets of the unique records, based on the non-fuzzy PR values. Guided by the derived IT2FL expressions in [5], the correct target class is determined, with 1 placed in the correct target class and 0 s placed in other target classes.

Ethics Statement

The University of Uyo Institutional Health Research Ethics Committee determined that this study did not qualify as human subjects because no protected health information was collected, accessed, or distributed (UU/CHS/IHREC/014).

CRediT Author Statement

Moses Ekpenyong: Conceptualization, Methodology, Writing-Original draft, Funding acquisition, Supervision; **Philip Etebong**: Data curation, Investigation, Writing-Original draft; **Tenderwealth Jackson**: Investigation, Validation, Supervision; **Edidiong Udofa:** Data curation, writing – Reviewing & Editing.

Declaration of Competing Interest

Moses Ekpenyong was funded by a research grant on drug-drug combination in treatment-enable patients on antiretroviral therapy by the Tertiary Education Trust Fund (TETFund), Nigeria. The authors declare that they have no known competing financial interests or personal relationships which have or could be perceived to have influenced the work reported in this article.

Acknowledgments

The authors thank The Stanford University for allowing their data (the Stanford HIV drug resistance database) to be shared publicly, and the Hospital facilities for assisting in the collection of the Akwa Ibom dataset.

Supplementary Materials

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.dib.2021.107147.

References

[1] M.E. Ekpenyong, M.E. Edoho, I.J. Udo, P.I. Etebong, N.P. Uto, T.C. Jackson, N.M. Obiakor, A transfer learning approach to drug resistance classification in mixed HIV dataset, informatics in medicine unlocked. In Press. 2021.
[2] M.W. Tang, T.F. Liu, R.W. Shafer, The HIVdb system for HIV-1 genotypic resistance interpretation, Intervirology 55 (2) (2012) 98–101, doi:10.1159/000331998.
[3] J.M. Mendel, R.I. John, F. Liu, Interval type-2 fuzzy logic systems made simple, IEEE Trans. Fuzzy Syst. 14 (6) (2006) 808–821 http://dx.doi.org/, doi:10.1109/TFUZZ.2006.879986.
[4] J.M. Mendel, X. Liu, Simplified interval type-2 fuzzy logic systems, IEEE Trans. Fuzzy Syst. 21 (6) (2013) 1056–1069, doi:10.1109/TFUZZ.2013.2241771.
[5] M.E. Ekpenyong, P.I. Etebong, T.C. Jackson. Fuzzy-multidimensional deep learning for efficient prediction of patient response to antiretroviral therapy. Heliyon. 2019; 5(7): 1–14. https://doi.org/10.1016/j.heliyon.2019.e02080.

[6] C. Wagner, Juzzy – a java based toolkit for type-2 fuzzy logic, in: Proceedings of the IEEE Symposium Series on Computational Intelligence, Singapore, 2013.
[7] C. Wagner, M. Pierfitt, J. McCulloch, Juzzy online: an online toolkit for the design, implementation, execution and sharing of type-1 and type-2 fuzzy logic systems, in: 2014 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), IEEE, 2014, pp. 2321–2328, doi:10.1109/FUZZ-IEEE.2014.6891548.