Journal of Health Informatics in Africa

8th Health Informatics in Africa Conference (HELINA 2013)
Peer-reviewed and selected under the responsibility of the Scientific Programme Committee

Factors Associated with HIV Prognosis in Rural Uganda: Sequence-Mining the Medical Record

M Velez [a], WO Ogallo [a,]*, H Cole-Lewis [a], R Pivovarov [a], H Salmasian [a], E Atuhairwe [b], E Toko [b], N Elhadad [a], AS Kanter [a]

[a] Columbia University in the City of New York, USA
[b] Millennium Villages Project, Ruhiira, Uganda.

Background and Purpose: With the increased adoption of the electronic health record, data mining emerges as an important technique for the empirical discovery of new knowledge from health data. Itemset sequence mining is a technique that enables the discovery of temporal (ordered) associations between entities in a database. The purpose of this study is to demonstrate the use of itemset sequence mining to discover factors associated with good and poor HIV prognosis amongst a cohort of HIV patients in rural Uganda.

Methods: We obtained a de-identified and date-shifted OpenMRS-based medical dataset from the Millennium Villages Project site in Ruhiira Uganda. The dataset contained the records of 129,080 patients with 5,825,579 observations made in 335,000 encounters between 2008 and 2011. We extracted relevant demographic and clinical variables for all HIV positive patients in the dataset. The extracted cohort was stratified into good and poor prognosis groups. The good prognosis group comprised patients whose modal stages over the duration of follow-up were either WHO stage I or II. The poor prognosis group comprised patients whose modal stages were either WHO stage III or IV. We applied the Sequential Pattern Discovery Equivalence classes (SPADE) algorithm via the arulesSequences package in R (programming language) to discover frequent sequence rules in each group. Pruning strategies were used to filter out sequence rules that were insignificant. Internal validation was done using 10-fold repeated random sub-sampling and only sequence rules common to all folds were assumed to be representative findings. External evaluation was done by two independent researchers who assessed the relevance of the generated sequence rules.

Results: A total of 3,353 records pertaining to HIV positive patients were extracted from the dataset. The good and poor prognosis groups contained of 1,441 patients and475 patients respectively. 18 sequence rules in the good prognosis group and 14 sequence rules in the poor prognosis group were obtained. The external evaluation of these sequence rules was conflicted, with an inter-rater agreement by Cohen's Kappa being 32% (poor) for the good prognosis rules and 57.1% for the poor prognosis rules. The good prognosis group was associated with having a normal body mass index, being single or married, being diagnosed with urinary tract infections, and having signs and symptoms of fever and headache. The patients in this group were more likely to be newly diagnosed and less likely to be on antiretroviral therapy. The poor prognosis group was associated with being underweight, being widowed or divorced, and being diagnosed with respiratory tract infections. The patients in this group were more likely to be in the WHO stage III and reported the HIV status of their partners as either positive or unknown.

Conclusions: Our study suggests that sequence mining may indeed substantiate known associations as well as generate useful hypotheses concerning HIV prognosis. Further investigations using larger and more diverse datasets are warranted.

Keywords: HIV Prognosis, Sequence Mining, Rural Uganda

*Corresponding author: Department of Biomedical Informatics, Columbia University Medical Center
622 West 168th Street, VC5 New York, NY, 10032. Email: woo7001@dbmi.columbia.edu, Tel: +(1)-(347) (3871213)
HELINA 2013 M. Korpela et al. (Eds.)
 DOI: 10.12856/JHIA-2013-v1-i1-64