Advances in Informatics, Management and Technology in Healthcare
J. Mantas et al. (Eds.)

doi:10.3233/SHTI220664

Data Cleaning Process for mHealth Log Data to Inform Health Worker Performance

Simon Muhindi SAVAI[a,1], Md Kamrul HASAN[b], Jemimah KAMANO[c], Lawrence MISOI[d], Peter WAKHOLI[e] and Martin C WERE[f]
[a]*Institute of Biomedical Informatics, Moi University, Eldoret, Kenya*
[b]*Department of Computer Science, Vanderbilt University, TN, USA*
[c]*School of Medicine, Moi University, Eldoret, Kenya*
[d]*Moi Teaching and Referral Hospital, Eldoret, Kenya*
[e]*School of Computing and Informatics Technology, Makerere University, Kampala*
[f]*Department of Biomedical Informatics, VUMC, Nashville, TN, USA*

Abstract. Log data, captured during use of mobile health (mHealth) applications by health providers, can play an important role in informing nature of user engagement with the application. The log data can also be employed in understanding health provider work patterns and performance. However, given that these logs are raw data, they require robust cleaning and curation if accurate conclusions are to be derived from analyzing them. This paper describes a systematic data cleaning process for mHealth-derived logs based on *Broeck's* framework, which involves iterative screening, diagnosis, and treatment of the log data. For this study, log data from the demonstrative *mUzima* mHealth application are used. The employed data cleaning process uncovered data inconsistencies, duplicate logs, missing data within logs that required imputation, among other issues. After the data cleaning process, only 39,229 log records out of the initial 91,432 usage logs (42.9%) could be included in the final dataset suitable for analyses of health provider work patterns. This work highlights the significance of having a systematic data cleaning approach for log data to derive useful information on health provider work patterns and performance.

Keywords. mHealth, developing countries, paradata, data cleaning

1. Introduction

Mobile health (mHealth) applications are increasingly used in low- and middle-income countries to support care. Beyond improving care, these applications can help to strengthen the health systems. Many mHealth solutions incorporate functionality to collect paradata, defined as "process data documenting users' access, participation, and navigation through an mHealth application [1]. When collected securely and used ethically, these paradata can be used to generate metrics that inform user engagement with the mHealth application, as well as for better understanding work performance and patterns for health workers [2]. In the age of COVID-19, where direct supervision for providers can be challenging, approaches that augment supportive supervision, such as innovative use of logs, are of particular relevance.

[1] Corresponding Author, Simon Muhindi SAVAI, Institute of Biomedical Informatics, Moi University, Eldoret, Kenya; E-mail: mssavai@gmail.com.

mHealth-derived logs are typically raw data that require significant data cleaning and curation to avoid inaccurate and unreliable conclusions that can lead to erroneous decision making [3]. Unfortunately, few studies provide details on systematic procedures and processes used for mHealth paradata cleaning based on best practices, despite data cleaning being an important and often time-consuming part of any paradata-based initiative. The aim of this study is to report on the method and results of a systematic and replicable data cleaning approach employed on mHealth-derived log-based paradata that was intended for use in health worker performance evaluation.

2. Methods

2.1. Study Setting and Participant

This study involved data cleaning of log data derived from a widely deployed mHealth application, *mUzima* [4]. *mUzima* is a robust open-source Android application used by health providers to primarily review and capture clinical data as part of patient care. The application has multiple other features, such as offline capabilities and clinical decision support. *mUzima* can capture usage log data as health providers use the application. The log data are transmitted securely to a MongoDB database. For this study, approval was obtained from the Institutional Review and Ethics Committee at Moi University School of Medicine in Eldoret, Kenya. After consent, 23 health care providers serving 89 care facilities and 24 community-based groups were equipped with the *mUzima* application, and usage logs captured over a 3-month period between 2nd December 2019 and 2nd March 2020. The goal was to use the captured mHealth log data to evaluate work patterns by the study participants including: (a) number of patients seen per day; (b) number of days worked by providers during the study period; (c) work hours per day; and (d) length of patient encounters. The above analyses could only be conducted after robust data cleaning and curation of the collected logs.

2.2. Data Cleaning Framework

Cleaning or transformation of raw data is essential to obtain quality data to support analysis and data mining. Several frameworks exist for assessing data quality in health information systems. For this study, a conceptual data-cleaning framework proposed by *Van den Broeck* et. al. (*Broek's* framework), was employed, with the goal of retaining only data fit for subsequent analyses [5]. This framework provides deliberate, systematic, and customizable data cleaning guidelines, and has three iterative data cleaning phases, namely: (1) data screening, (2) data diagnosis, and (3) data editing. The screening process involves identification of lacking or excess data, outliers and inconsistencies and strange patterns. Diagnosis involves determination of errors or missing data and any true extremes and true normal. Editing involves correction or deleting of any identified errors [5].

2.3. Data Cleaning Procedure

Cleaning of the data involved a step-by-step approach that incorporated the *Broeck's* framework. These detailed steps included: *Step 1* - outlining the analyses or evaluation questions, which for our purposes was provider performance metrics; *Step 2* - describing data and study variables; *Step 3* - creating the initial database by extracting data from relevant databases to generate the final data set; *Step 4* - applying *Broeck's* framework for data cleaning - during this step, the three data cleaning phases (screening, diagnosis, and treatment) are iteratively applied on data set (**Figure 1**); *Step 5* – analyzing the resulting data to provide a summary of the data quality issues discovered, the eliminated data after the treatment exercise, and the retained final data set.

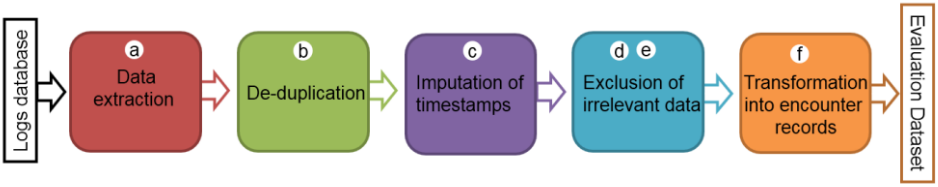

Figure 1. Creation of evaluation dataset.

Specific steps involved in applying the *Broeck's* framework for data cleaning included: (a) extraction of the usage logs from the initial database and filtering the data to retrieve records with fields of interest for the study, while excluding logs related to test users. During this extraction, timestamps, user identifiers, patient identifiers, form identifiers and device identifiers were reformatted and standardized for easier data handling; (b) removal of duplicate records to retain unique entries only; (c) imputation of correct timestamps for the records that had incorrect device timestamps - done using the GPS timestamps or the server timestamps; (d) exclusion of records with timestamps falling outside the study period; (e) exclusion of logs that were recorded from a device that lacked meaningful data; and (f) derivation of encounter characteristics by grouping records in the cleaned dataset by encounters. All scripts for data cleaning were developed using *python* programming language which was used to extract data from the MongoDB database, to clean and transform the data, to generate datasets for the target metrics, and to export the datasets into CSV file format for analyses. Specifically, Python's Pandas library was used for manipulation of the data and for generating the various datasets, while Matplotlib library was used for visualization.

3. Results

A total of 91,432 usage logs were captured during the study period between 2nd December 2019 to 2nd March 2020. A total of 4,848 (5.3%) logs that were not relevant to evaluating work performance metrics were then removed, leaving 86,584 logs from 19 study participants. Next, 33,612 duplicate logs were identified and removed, resulting in 52,972 records for the 19 participants. Of the remaining logs, 6,536 (12.3%) had incorrect device timestamps, and these timestamps were imputed. After timestamp imputation, a total of 13,739 (25.9%) of the logs fell outside of the study period and were also removed, leaving 39,233 logs for 15 participants. An additional four logs were removed as these belonged to one provider who attempted to log in four times but failed. The remaining 39,229 log records, which was only 42.9% of the logs in the initial dataset, were determined to be the cleaned logs available for performance metrics analyses.

4. Discussion

In this study, we describe a systematic approach for cleaning data from mHealth logs, providing evidence of this important step when leveraging paradata for accurate decision-making. The use of *Broeck*'s framework in data cleaning process allows for iterative data cleaning, helping uncover data issues that can easily missed. The rule-based data cleaning approach used identified numerous issues, including integrity constraints, duplication, inconsistencies, and missing values. Of relevance to mHealth applications in LMICs that often operate offline is the issue of incorrect timestamps due to users changing the time setting on their devices. In addition, timestamps from various devices are often captured in inconsistent formats. Approaches to improve mHealth applications to ensure correct times on devices, to standardize capture of timestamps that includes the time zone and to effectively impute times on logs, are of particular importance. While this study is limited by the fact that it was conducted in one setting using a single mHealth application, it serves as a demonstrative example relevant to other instances where paradata cleaning is needed. Moving forward, we plan to automate the data cleaning process for faster and scalable use of mHealth log data for data analytics and data mining.

5. Conclusions

A systematic data cleaning approach can uncover multiple data quality issues in mHealth log data and should be an integral part of any analyses using mHealth paradata.

Acknowledgement

This work was supported through USAID (Grant #: 7200AA18CA00019) and the NORHED program (Norad: Project QZA-0484). The contents are solely the responsibility of the authors and do not necessarily represent the official views of USAID, Norad, or the United States Government. The authors thank the participating healthcare providers and clinics, as well as the PIC$_4$C project in Western Kenya.

References

[1] Bonett S, Connochie D, Golinkoff JM, Horvath KJ, Bauermeister JA. Paradata Analysis of an eHealth HIV testing intervention for young men who have sex with men. AIDS Education and Prevention. 2018;30(5):434-47.

[2] Hightow-Weidman LB, Bauermeister JA. Engagement in mHealth behavioral interventions for HIV prevention and care: making sense of the metrics. Mhealth. 2020;6:7. Epub 2020/03/20. doi: 10.21037/mhealth.2019.10.01. PubMed PMID: 32190618; PubMed Central PMCID: PMCPMC7063263.

[3] Weiskopf NG, Bakken S, Hripcsak G, Weng C. A Data Quality Assessment Guideline for Electronic Health Record Data Reuse. EGEMS (Wash DC). 2017;5(1):14. Epub 2018/06/09. doi: 10.5334/egems.218. PubMed PMID: 29881734; PubMed Central PMCID: PMCPMC5983018.

[4] Were MC, Savai S, Mokaya B, Mbugua S, Ribeka N, Cholli P, et al. mUzima Mobile Electronic Health Record (EHR) System: Development and Implementation at Scale. Journal of medical Internet research. 2021;23(12):e26381. Epub 2021/12/15. doi: 10.2196/26381. PubMed PMID: 34904952; PubMed Central PMCID: PMCPMC8715359.

[5] Van den Broeck J, Argeseanu Cunningham S, Eeckels R, Herbst K (2005) Data Cleaning: Detecting, Diagnosing, and Editing Data Abnormalities. PLOS Medicine 2(10): e267. doi: 10.1371/journal.pmed.0020267. PubMed PMID: 16138788; PubMed Central PMCID: PMC1198040