The Importance of Health Informatics in Public Health during a Pandemic
J. Mantas et al. (Eds.)

doi:10.3233/SHTI200514

K-Means Clustering in Monitoring Facility Reporting of HIV Indicator Data: Case of Kenya

Milka B. GESICHO [a,d, 1], Ankica BABIC [a,b] and Martin C. WERE [c,d]
[a] *Department of Information Science and Media Studies, University of Bergen, Norway*
[b] *Department of Biomedical Engineering, Linköping University, Sweden*
[c] *Vanderbilt University Medical Center, US*
[d] *Institute of Biomedical Informatics, Moi University, Kenya*

Abstract. Health management information systems (HMISs) in low- and middle-income countries have been used to collect large amounts of data after years of implementation, especially in support of HIV care services. National-level aggregate reporting data derived from HMISs are essential for informed decision-making. However, the optimal statistical approaches and algorithms for deriving key insights from these data are yet to be fully and adequately utilized. This paper demonstrates use of the k-means clustering algorithm as an approach in supporting monitoring of facility reporting and data-informed decision-making, using the case example of Kenya HIV national reporting data. Results reveal four homogeneous cluster categories that can be used in assessing overall facility performance and rating of that performance.

Keywords. HIV-indicator, dhis2, k-means clustering, monitoring, data-use

1. Introduction

Implementation of Health Management Information Systems (HMISs) for purposes of improving monitoring and evaluation efforts toward eradication of HIV in low- and middle-income countries has resulted in large amounts of data. Facilities using HMISs are required to submit various reports to aggregate-level HMISs[1], such as the District Health Information Software Version 2 (DHIS2) used in many countries [2]. These aggregate data are essential for program monitoring and evaluation (M&E) and for data-informed decision making (DIDM). DIDM is essential in informing policy and advocacy, and in program design, improvement, operations and management [2]. The ultimate aim of DIDM is achievement of improved health outcomes. For the submitted reports to be of best use to monitoring and evaluation (M&E) efforts, they must be complete, accurate and submitted in a timely manner. For the case of HIV, a weakness in understanding HIV information use infrequently addressed in previous studies is how M & E teams at the national level can utilize various approaches to derive insights from HIV facility reporting data aggregated in HMISs. In this study, we demonstrate use of the k-means

[1] Corresponding Author, Milka Gesicho, Department of Information Science and Media Studies, University of Bergen, Norway; E-mail: milcagesicho@gmail.com.

clustering algorithm as an approach in supporting monitoring of facility reporting and DIDM, using the case example of Kenya HIV national reporting data.

2. Methods

A retrospective observational study was used in monitoring performance trends in HIV reporting from health facilities in Kenya. DHIS2, the national aggregate reporting system, was used in extracting facility HIV reporting completeness and timeliness data for all health facilities in all 47 counties in Kenya, for the year 2011 to 2018. Systematic procedures were used in cleaning the data prior to analysis. Facilities in this study included only those offering HIV care and treatment services. This study explored an automated approach of grouping facilities based on their reporting completeness and timeliness, as a way of determining overall facility performance in reporting. Facility reporting completeness was defined as the extent to which facilities submit the expected number of reports, and timeliness as reporting submission within the defined reporting deadline. The actual number of reports submitted by facilities are automatically calculated within DHIS2, against the expected number of reports. The k-means clustering algorithm was used in identifying homogeneous groups within the data. The average silhouette coefficient was used in measuring the quality of the selected clusters [3]. All analyses were conducted in SPSS.

3. Results

A total of 18,394 HIV care and treatment reports from a total 3,242 facilities for the period 2011-2018 were evaluated. Based on the average silhouette measures for each year (ranging from 0.58 to 0.70); the k value used was four (k=4), with the four homogeneous groups of facilities identified as: best performers, average performers, poor performers, and outlier performers. Figure 1 to Figure 4 illustrate the exact performance (report timeliness and completeness) over time by facilities in each of these clusters. Figure 1 illustrates results for facilities in the best performers cluster, where average percentage completeness and timeliness was high (80% and above) in the various years (2012 to 2018).

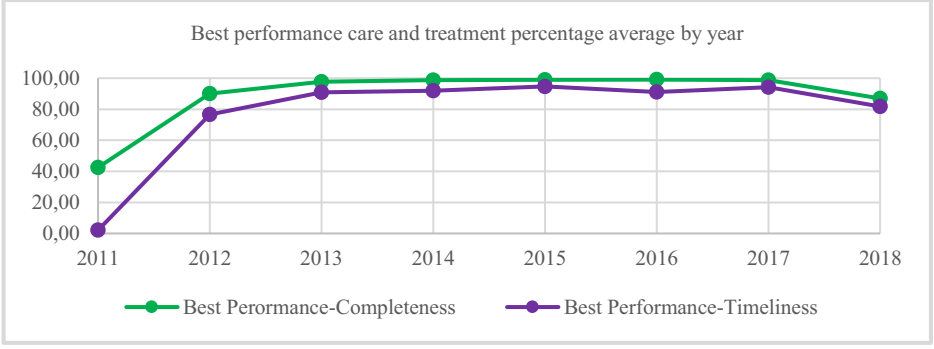

Figure 1. Care and treatment facility reporting best performance.

Figure 2 illustrates results for facilities in the average performance cluster, where percentage completeness and timeliness were lower in comparison to best performance

facilities in the various years respectively. For instance, performance in 2015 for timeliness and completeness is lower by 28.65% and 5.37% respectively compared to performance in 2015 for best performance (Figure 1.).

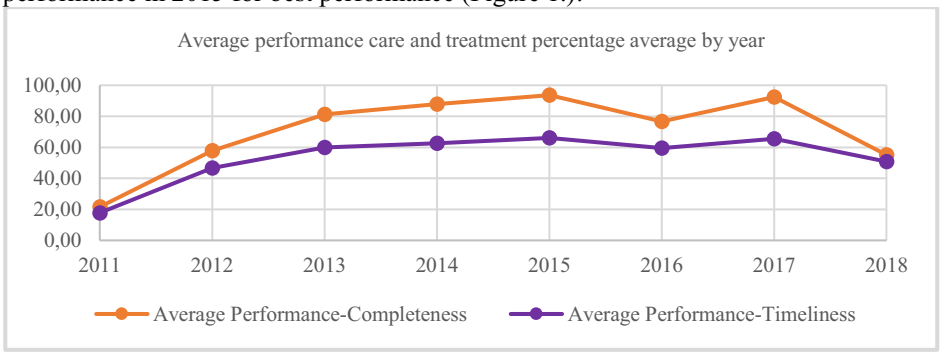

Figure 2. Care and treatment facility reporting average performance

Figure 3 illustrates results for facilities in the poor performance cluster, where percentage completeness and timeliness was low (below 50%) in the various years.

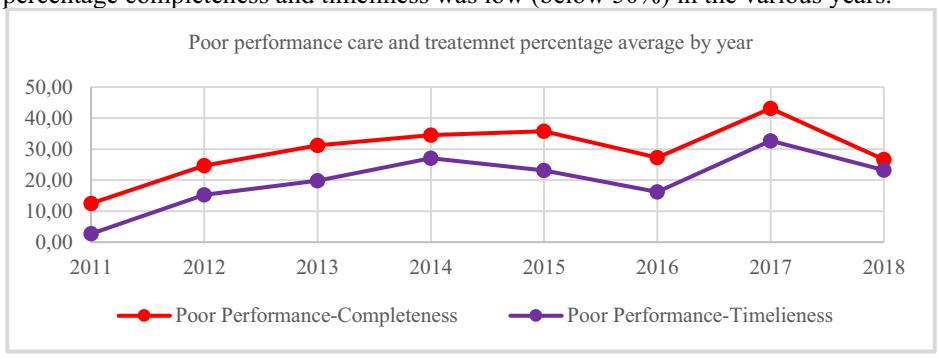

Figure 3. Care and treatment facility reporting poor performance.

Figure 4 illustrates results for facilities in the outlier performance cluster, where there was an evidently big gap between percentage completeness and timeliness in the various years. This depicts scenarios where timeliness was a problem despite good performances in completeness.

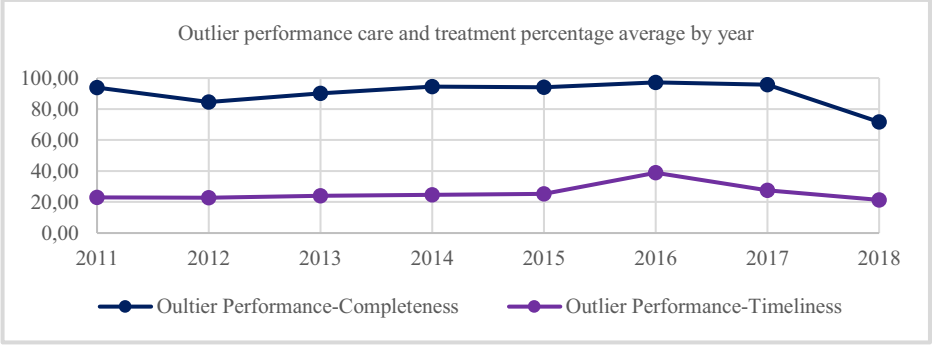

Figure 4. Care and treatment facility reporting outlier performance.

4. Discussion

In this paper we illustrate use of the k-means clustering algorithm as an approach in assessing retrospective HIV facility reporting to determine facility performance, as a way of informing reporting improvement mechanisms. These analyses provide at a glance view of various categories that emerge based on performance of facilities in meeting completeness and timeliness requirements in reporting. These results can be used by M&E teams to identify facilities whose performance is satisfactory or not, therefore providing a baseline for further evaluations and development of sustainable solutions.

Furthermore, due to the volume, velocity and veracity of health data consolidated from various sources, representing various facets of data in a way that makes sense is a challenge. In this study, we used line graphs, which are simple visualizations that can be used to represent data in a way that promotes development of insights at a glance. Figure 1 portrays an ideal situation of good facility reporting. If success is to be achieved in terms of meeting reporting requirements, then the ultimate goal for these evaluations should be to enable all facilities to attain and maintain similar results as illustrated in the best performing category. A common attribute among average, poor and outlier performance is the discrepancy between completeness and timeliness as represented by the gaps observed between them in the respective performance categories. There is therefore need for investigating issues that bring about delays more so in the outlier performance group, which has the largest gap in the completeness and timeliness measures. As the next step, we will further disaggregate the results by facility characteristics and geographic region, and also look at additional reporting domains.

5. Conclusions

The k-means clustering algorithm is essential in automatically finding homogenous groups within aggregate reporting data. This serves as a good baseline for monitoring the progression of health facility reporting performance by management and M&E teams that use large amounts of data collected from integrated data sources.

Ethical approval and Acknowledgements

Ethical approval was obtained from the IREC in Moi University-No.0003362. This work was supported in part by the NORHED program (Norad: Project QZA-0484). The content is solely the responsibility of the authors.

References

[1] Manya A, Nielsen P, Reporting practices and data quality in health information systems in developing countries: an exploratory case study in Kenya, J. Health Inform. Dev. Ctries 10 (2016), 114–126.
[2] Karuri J, Waiganjo P, Orwa D, Manya A, DHIS2: The Tool to Improve Health Data Demand and Use in Kenya, J. Health Inform. Dev. Ctrie 8 (2014), 38–60.
[3] Kaufman L, Rousseeuw PJ, Finding Groups in Data, An Introduction to Cluster Analysis. John Wiley & Sons, Inc, New Jersey, 1990.