International Journal of Population Data Science

Journal Website: www.ijpds.org

Unified health database creation: 125 million brazilian cohort from information systems of hospital, outpatient, births, notifications and mortalities

Pereira, Ramon[1*], Dias, Leonardo[2], Ávila, Juliano[3], Santos, Núbia[3], Gurgel, Eli Iola[4], Cherchiglia, Mariangela Leal[4], AcÚrcio, Francisco[5], Reis, Afonso[6], Junior, Wagner Meira[1], and Guerra Junior, Augusto Afonso[5]

[1]DCC/UFMG
[2]CCATES/UFMG
[3]DAAED
[4]MEDICINA/UFMG
[5]FARMACIA/UFMG
[6]DEMAS

Objectives

Our objectives were unify and deduplicate databases' of patients registration information coming from Information Systems of SUS in Brazil: Hospital, Outpatient, Births, Notifications and Mortalities, between the years 2008-2015, to get an individualize data and plot patients' lines of care during the period, enabling pharmacoeconomic and epidemiological studies that parameterize effectiveness and efficiency of public policies and embedded technologies.

Methods

Semantic analysis of data was performed to describe and understand different meanings of different fields existing in the studied bases. In addition, there were four main procedures, executed with database operations tools and PLSQL programming language: cleaning and standardization of databases(document's numbers was checked in the brazilian national people's database, with a string approximator algorithm to decide if the document's number belonged or no the register); registration information extraction, deterministic and probabilistic deduplication thereof. The procedures were first performed on each database separately and after the unification of the records, was held again a deterministic deduplication. Except the probabilistic deduplication which was performed only on the final deterministic deduplicated's database.

Performed procedures allowed a decision-making to chose fields used in data model for the unified database creation. Nine database's representative fields related to patients were selected: patient's name; patient mother's name; sex; birth date; state; city; zip code; cpf and cns(brazilian documents).

Results

Initially, the unified registration database resulted in 705.599.785 records, after deterministic deduplication there was a reduction culminating in 198.400.762 records. This reduction is explained because these databases are not fully integrated. Moreover, there is not always agreement between systems' semantics and in some cases changes occur in the data format over the period within the same system. After probabilistic deduplication, the number of unique records decreased to 124.545.186 which is explained by non-linked pairs by deterministic process. This result is guaranteed with a estimate error of at most 3.3% of false positive and at most 12.3% of false negative pairs.

Conclusion

The results show that data deduplication is necessary and should be carried out thoroughly. Where the databases had limited patients' registration information, the technique enabled to capture, in more complete basis, additional information. Futhermore, it allowed to identify and assist in the understanding of positive and negative aspects within systems and trace clinical condition of patients, enabling pharmacoeconomic and epidemiological studies that define effectiveness and efficiency of public policies and embedded technologies. As future work, is important ensure the univocity of records and link this database with past period.

*Corresponding Author:
Email Address: ramonbhb@ufmg.br (R. Pereira)

http://dx.doi.org/10.23889/ijpds.v1i1.225