Informatics for Health: Connected Citizen-Led Wellness and Population Health
R. Randell et al. (Eds.)

doi:10.3233/978-1-61499-753-5-539

A Standardized and Data Quality Assessed Maternal-Child Care Integrated Data Repository for Research and Monitoring of Best Practices: A Pilot Project in Spain

Carlos SÁEZ[a,b,1], David MONER[a,b], Ricardo GARCÍA-DE-LEÓN-CHOCANO[c], Verónica MUÑOZ-SOLER[b,c], Ricardo GARCÍA-DE-LEÓN-GONZÁLEZ[c], José Alberto MALDONADO[a,b], Diego BOSCÁ[a,b], Salvador TORTAJADA[a,e], Montserrat ROBLES[a], Juan M GARCÍA-GÓMEZ[a], Manuel ALCARAZ[c], Pablo SERRANO[d], José L BERNAL[d], Jesús RODRÍGUEZ[d], Gerardo BUSTOS[d], and Miguel ESPARZA[b]
[a] *Instituto Universitario de Tecnologías de la Información y Comunicaciones. Universitat Politècnica de València. Camino de Vera s/n. 46022 Valencia, España*
[b] *VeraTech for Health S.L., Valencia, España*
[c] *Hospital Virgen del Castillo, Yecla, España*
[d] *Hospital 12 de Octubre, Madrid, España*
[e] *Insituto de investigación sanitaria La Fe, Hospital Universitari i Politècnic La Fe, Valencia, España*

Abstract. We present the results of a pilot project of the Spanish Ministry of Health, Social Services and Equality, envisaged to the development of a national integrated data repository of maternal-child care information. Based on health information standards and data quality assessment procedures, the developed repository is aimed to a reliable data reuse for (1) population research and (2) the monitoring of healthcare best practices. Data standardization was provided by means of two main ISO 13606 archetypes (composed of 43 sub-archetypes), the first dedicated to the delivery and birth information and the second about the infant feeding information from delivery up to two years. Data quality was assessed by means of a dedicated procedure on seven dimensions including completeness, consistency, uniqueness, multi-source variability, temporal variability, correctness and predictive value. A set of 127 best practice indicators was defined according to international recommendations and mapped to the archetypes, allowing their calculus using XQuery programs. As a result, a standardized and data quality assessed integrated data respository was generated, including 7857 records from two Spanish hospitals: Hospital Virgen del Castillo, Yecla, and Hospital 12 de Octubre, Madrid. This pilot project establishes the basis for a reliable maternal-child care data reuse and standardized monitoring of best practices based on the developed information and data quality standards.

Keywords. Integrated data repositories, Data quality, Normalization, ISO 13606, Archetypes, Best practices, Quality indicators, Data reuse

[1] Corresponding autor: Carlos Sáez (carsaesi@ibime.upv.es), Instituto ITACA, Universitat Politècnica de València, Camino de Vera s/n, 46022, Valencia, Spain / VeraTech for Health SL, Avenida del Puerto 237-1, 46011, Valencia, Spain

1. Introduction

Integrated Data Repositories (IDRs) are becoming an essential resource enabling the biomedical data reuse on larger amounts and sources of data [1]. Several initiatives have been carried out on IDRs to provide access to biomedical research data, either as federated query tools [2,3] or as centralized repositories [4]. In most solutions, the adoption of a common data format was key. However, to our knowledge, the use of specific health information standards was limited [5]. Besides, it is agreed that the reliability of data reuse depends greatly on its Data Quality (DQ) [6]. Certainly, DQ assessment is considered a key component to any IDR [7], where some successful examples can be found in the recent literature [4,8].

We present the results of a pilot project of the Spanish Ministry of Health, Social Services and Equality (2015/07PN0010), envisaged to the development of a National IDR of maternal-child care information. The project had two main motivations. First, the evaluation of maternal and child health strategies [9,10] with the ultimate aim of disseminating best practices (BPs). Second, to provide a repository for population-based research, with a special focus on breastfeeding, as one of the main determinants for maternal and child health [10]. An IDR was developed as solution, which, based on health information standards, ensured a common interface for monitoring BPs of different hospitals and regions, and having its DQ assessed ensures a reliable data reuse.

2. Materials and Methods

The pilot was divided in three main workpackages: (1) definition of clinical information models and standardization of data, (2) DQ assessment and (3) definition of a proposal of BPs indicators. Clinical and data support was provided by the two participating hospitals: Virgen del Castillo Hospital, Yecla (VCH) and 12 de Octubre Hospital, Madrid (12OH). Figure 1 shows the architecture of the proposed solution.

Regarding to standardization, we used ISO 13606 archetypes [11] to provide the IDR with an information model about the data structure (how data is organized) and the data constrains to be fulfilled (which values are valid). To create the required archetypes, a multidisciplinary group of professionals was arranged. A proper information modelling is crucial to ensure that the relevant clinical information will be available for the particular data reuse purposes. Hence, the archetype creation methodology included the identification of the main clinical data structures, the selection and aggrupation of relevant data items for the required clinical domains, the search of reusable archetypes, the creation of new archetypes or adaptation of existing ones, and their validation by clinical experts. Finally, archetypes were mapped to the maternal-child care data extracted from the original data sources, which were transformed into the ISO 13606 archetype-compliant documents. Archetype edition and data transformation were performed using LinkEHR Studio [12]. The IDR was implemented in ExistDB, and queries were defined using native XQuery language.

The DQ assessment was carried out with a dedicated procedure based on seven dimensions [13]. Data completeness (non-missing data, weighting obligatory and optative elements), consistency (conformance to schema rules) and uniqueness (non-replicated identifiers) were calculated according to the archetype conformance requirements and based on our previous studies [14]. The multi-source and temporal variability of data (degree of data concordance among different sources and over time)

were assessed based on our probabilistic DQ methods [15-17]. Data correctness accounted for the number of possibly anomalous records (multivariate outliers). Finally, the predictive value dimension measured the baseline dataset potential to predict the breastfeeding continuity at one month as the AUC of a Naïve Bayes classifier with a 10-fold cross-validation estimation, as a measure of the data usefulness for this task.

Lastly, the BPs workpackage included two tasks. First, the formal definition of a set of BPs indicators based on an extensive review of literature. And second, the operative definition of the indicators and mapping of their variables to the archetype information for their automatic calculation.

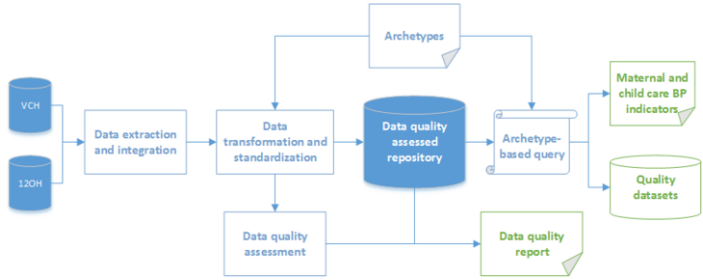

Figure 1. Architecture of the IDR solution for a standardized and reliable maternal-child care data reuse

3. Results

Two main archetypes were created as the basis for (1) a report of perinatal health information (Figure 2), including information about family history, gestation, delivery, birth, and maternity, and (2) a report of the infant feeding up to two years, including information about breastfeeding and the dates of introduction of different types of food. The two main archetypes were composed of 43 sub-archetypes. The two archetypes can be accessed at http://mm.linkehr.com/ (currently available in Spanish only).

To populate the IDR, the archetypes were mapped to the hospitals data sources and data were transformed to ISO 13606 instances. The VCH provided data from 3781 records for the newborn report and 2133 for the infant feeding report. The 12OH provided 1949 records for the newborn report, but only from their neonatal database.

DQ was assessed on the VCH and 12OH data separately for the perinatal and infant feeding datasets, evaluating the original EHR data (PRE) and after its standardization in the IDR (POST). A DQ report was generated, including the following main findings. Uniqueness: 100%, no replicated patient identifiers found. Completeness: in POST, completeness decreased due to stricter information requirements by the archetypes. Particularly, despite the higher completeness of the PRE 12OH neonatal dataset (77%), this filled a minor part of the more detailed perinatal archetype in POST (8%), whilst the average completeness of the VCH remained more stable (56% to 52%), filling in a higher degree the information required by the archetype. Consistency: in this pilot, due to the large amount of variables, data types and range checks were not included, thus, consistency results accounted for the un-conformance to data obligatoriness, and high measurements were obtained in general. Temporal stability: the method warned about a minor number of wrongly dated records in 12OH, showing up as an anomalous temporal subgroup; the VCH perinatal data showed three temporal subgroups (showing non-concordant data among their periods) related to two changes in the original EHR system; the VCH infant

feeding data was stable over time. Multi-source stability: a low stability between the VCH and 12OH perinatal data was found (0.08 out of 1), as expected due to their different populations (maternity vs neonatal units) and completeness derived from this reason. Correctness: an average of 1% of outlier records were found in all the datasets. Predictive value: an AUC of 0.60 was obtained using all the variables of VCH perinatal dataset. Six records did not pass the standardization process due to strong quality faults.

Finally, a set of 127 BPs indicators, grouped in six categories involving different clinical processes (from gestation to primary-care follow-up), was defined according to national and international recommendations (Euro-Peristat, WHO, UNICEF). All the variables in the operative definitions of indicators were mapped to the archetypes. A BP monitoring system was developed using XQuery programs on the standardized data.

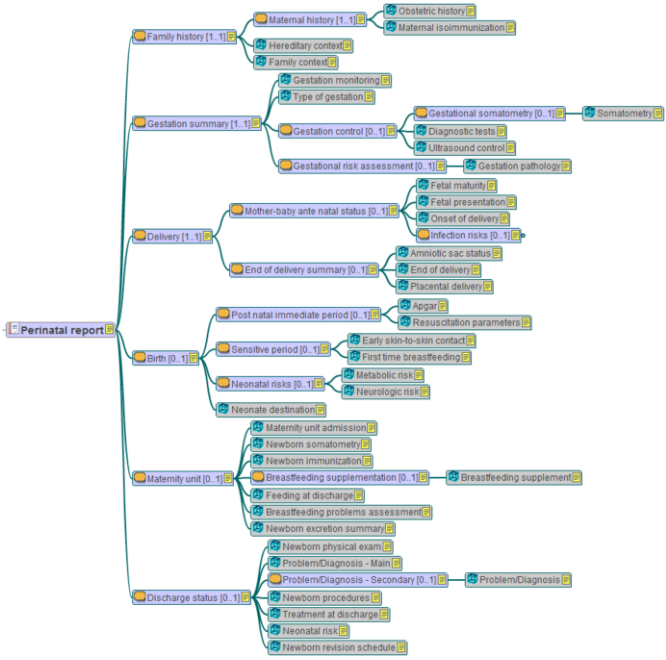

Figure 2. Contents of the perinatal report archetype

4. Discussion

The defined archetypes should be considered as an initial version, candidate to be revised by further professionals, towards a harmonized detailed clinical model. Further versions should also incorporate terminology bindings (e.g., to SNOMED-CT) to provide the semantics of the information structures and data values, not covered in this pilot. Mapping the local vocabularies of the two hospitals to the controlled vocabulary used in the archetypes was an intense task. Using standard terminologies in both the archetypes and the original data sources can solve this problem in the future.

The DQ assessment provided novel insights about the effect on DQ of data standardization using health information standards. Stricter information requirements can improve DQ in terms of completeness and consistency, and improve the usability of data given their contextualization. However, the change of the variable space given by the standardization must be considered in comparing PRE and POST measurements:

archetypes defined a larger set of information compared to the existing data. Consistency assessment can be improved including data type and range checks, what can be supported by the use of terminologies. The temporal variability results can be of utmost utility to support the mapping of data with variable representations over time.

The BPs monitoring based on standardized data would allow further enrolled hospitals getting instant BPs monitoring, comparison, and DQ assessment, just by providing equivalent standardized data. Finally, other repository technologies are to be explored to support advanced query needs and to improve the efficiency of the IDR.

5. Conclusion

This pilot project established the basis for a national IDR for a reliable maternal-child data reuse and standardized monitoring of BPs. The discussed lessons learned can facilitate the scaling-up of the project in national and international actions. The developed approach can be replicated in additional healthcare domains.

Acknowledgements

Work co-supported by grants: RTC-2014-1530-1, AEST/2016/023 and DI-14-06564.

References

[1] L. Toubiana and M. Cuggia, Big Data and Smart Health Strategies: Findings from the Health Information Systems Perspective. *IMIA Yearbook.* 2014;9(1):125–7.
[2] A.J. McMurry et al. SHRINE: Enabling Nationally Scalable Multi-Site Disease Studies. Carter KW, editor. *PLoS ONE*. 2013 7;8(3):e55811.
[3] G.M. Weber et al. Direct2Experts: a pilot national network to demonstrate interoperability among research-networking platforms. *J Am Med Inform Assoc*. 2011;18:157–60.
[4] S.N. Murphy et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *J Am Med Inform Assoc*. 2010;17(2):124–30.
[5] K.L. Walker et al. Using the CER Hub to ensure data quality in a multi-institution smoking cessation study. *J Am Med Inform Assoc*. 2014;21(6):1129–35.
[6] N.G. Weiskopf and C. Weng, Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research, *J Am Med Inform Assoc,* 2013;20(1):144–51.
[7] S.L. MacKenzie et al, Practices and perspectives on building integrated data repositories: results from a 2010 CTSA survey, *J Am Med Inform Assoc* 2012;19(1):119–24.
[8] M.G. Kahn et al, A Pragmatic Framework for Single-site and Multisite Data Quality Assessment in Electronic Health Record-based Clinical Research: *Med Care*. 2012;50:S21–9.
[9] Estrategia de atención al parto normal y nacimiento en el Sistema Nacional de Salud. Observatorio de Salud de la Mujer y del Sistema Nacional de Salud. Ministerio de Sanidad y Consumo, 2008.
[10] Evidence for the Ten Steps to Successful Breastfeeding. Geneva: WHO; 1998.
[11] ISO 13606:2008 - Health informatics - Electronic health record communication. 2008.
[12] J.A. Maldonado et al, LinkEHR-Ed: a multi-reference model archetype editor based on formal semantics, *Int J Med Inf*, 2009:78;8:559–70.
[13] C. Sáez et al. Organizing data quality assessment of shifting biomedical data. *Stud Health Technol Inform*, 180:721–725, 2012.
[14] R. García-de-León-Chocano et al. Construction of quality-assured infant feeding process of care data repositories: Construction of the perinatal repository (Part 2). *Comput Biol Med*, 2016:71(1):214–22.
[15] C. Sáez et al. Applying probabilistic temporal and multi-site data quality control methods to a public health mortality registry in Spain: A systematic approach to quality control of repositories. *J Am Med Inform Assoc,* 2016;23:1085-95.
[16] C. Sáez et al. Stability metrics for multi-source biomedical data based on simplicial projections from probability distribution distances. *Stat Methods Med Res*. 2017:26(1):312–336.
[17] C. Sáez et al. Probabilistic change detection and visualization methods for the assessment of temporal stability in biomedical data quality. *Data Min Knowl Discov*. 2015;29(4):950–75.