J Biochem Tech (2018) Special Issue (2): 39-46
ISSN: 0974-2328

Data Quality Engineering in Electronic Health Records

Meisam Nazariani, Ahmad Abdollahzadeh Barforoush*

Received: 22 November 2017 / Received in revised form: 18 May 2018, Accepted: 23 May 2018, Published online: 05 September 2018

Abstract

Electronic Health Record system has been highly regarded in the medical world. Using these systems, medical institution may develop a clinical data repository containing extensive records of a large number of patients, which provides them with more efficient retrospective research. The presence of human factors in the process of electronic data recording causes some data quality challenges. Using similarity functions and master data, a data quality engineering framework is developed to solve these problems. The proposed framework is applied to a population based cancer registry program. Finally, some experimental results are presented to show effectiveness of the proposed framework.

Keywords: Data Quality, Electronic Health Records, Data Cleansing, Record Linkage, Entity Resolution, Cancer Registry

Introduction

The *Electronic Health Record* (EHR) acts as an infrastructure for a range of *Health Information Technology* (HIT) applications (Rumball-Smith et al., 2018). There are many reasons to use EHR (e.g. store data accurately and capture the state of a patient across time). EHR eliminates the need to track down a patient's previous paper medical records and ensures that data is accurate and legible. However, the final goal of EHR is to provide better health care by improving all aspects of patient care including safety, effectiveness, patient-centeredness, communication, education, timeliness, efficiency and equity.

Since the first concepts for EHR in the 1990s, the content, structure and technology of such records were frequently changed and adapted. The basic idea to support and enhance health care remained the same over time. To reach these goals, it is crucial that EHRs themselves adhere to rigid quality requirements (Hoerbst and Ammenwerth, 2010).

One of the main quality requirements of EHR is data quality. In (Weiskopf and Weng, 2013), a review of the clinical research literature discussing data quality assessment methodology for EHR data was performed. Using an iterative process, the aspects of data quality being measured were abstracted and categorized, as well as the employed assessment methods.

EHR systems provide opportunities for case finding and improvement of completeness in *Population Based Cancer Registry* (Leinonen et al., 2017). In Iran, the national EHR system, a.k.a. SEPAS, has been established to store data about diagnosis, and treatment of patients (Nazariani and A. Barforoush, 2016a). However, SEPAS is a large database but in this research, many data quality problems have been discovered (e.g. *duplicate*, *out of range* and *inconsistent* records). Therefore, it is necessary to refine the data and prepare it for applications in the cancer registry.

In this paper, using similarity functions and ICD-10 (Association, 2018, p. 10) master data, a data quality engineering framework is developed (Nazariani and A. Barforoush, 2016b) for cleansing SEPAS and application of clinical data in *Population Based Cancer Registry*. Finally, some experimental results are presented to show effectiveness of the proposed framework.

The rest of this paper is organized as follows: problem statement is presented in Section 2. In Section 3, the proposed solution is introduced.

Meisam Nazariani, Ahmad Abdollahzadeh Barforoush*

Department of Computer Engineering and Information Technology, Amirkabir University of Technology, Tehran, Iran.

***Email:** ahmad@aut.ac.ir

Experiments are presented in Section 4. Finally, the paper is concluded in Section 5.

Statement of Problem

Considering the importance of the *Demographic Cancer Records Program* (Sateren et al., 2002) as one of the most important developmental research infrastructures, *National Cancer Registry Program* starts with the following objectives:

1. Providing accurate statistics on incidence (Ferlay et al., 2018), outbreak and mortality from cancer in line with the *National Cancer Control Program*.
2. Determining the *Age Standardized* (Robson et al., 2007) Incidence of cancers by age, sex, tumor characteristics including tumor location and morphology in each of the partner provinces and throughout the country.
3. Determining the incidence of cancer in each of the partner provinces and throughout the country.
4. Determining the death rate from cancer and the survival rate of cancers in each of the partner provinces and throughout the country.

One of the most common activities of a *Demographic Cancer Registration Program* is the discovery and identification of cancer patients. In the conventional method, trained people come to medical records of patients in medical records of hospitals or pathologic laboratories. They collect and record information items required for *Demographic Cancer Registration Program*. Finally, they present cancer incidences at a time interval (usually one year) and in a specific population.

The new strategy is to collect data from HIS and LIS software available in hospitals and laboratories, which reduces personnel costs, significantly. However, most of the data in these applications does not necessarily follow certain standards and rules (Nazariani and A. Barforoush, 2015). Sample records to show incompleteness and inconsistency between codes and descriptions are presented in Table 1:

Table 1- Data quality problems of disease code and description

ID	Code	Description
1	1094	stone
2	-	Stone;bladder
3	U95008	Stone;kidney
4	U95009	Stone;renal
5	U95011	Stone;ureter
6	U95012	Stone;urinary tract
7	1018	stp
8	660	stpd
9	F95002	Strabismus
10	H50.9	Strabismus, unspecifie
11	h50.9	STRABISMUS, UNSPECIFIED
12	C91.0	Asadollah
13	479	Lung Cancer
14	-	Strabismus, unspecified(DVD)
15	H50.9	Strabismus, unspecified(IOOA)
16	A25.1	STREPTOBACILLOSIS

Here, purpose is to introduce a data quality engineering framework to collect data in a correct way and cleans data. As a result, patterns of data quality errors related to cancer patients can be identified and categorized for cancer incidence identification.

The Proposed Solution

In this section, proposed method for analyzing about 10 million primary records of EHR/SEPAS to detect and repair erroneous records is presented. In the first step, EHR/SEPAS database is explored and a list of ICD-10 codes assigned for patients and the diagnosis description is developed. To detect and repair data quality problems, a reliable source of master data is needed. So, ICD-10 (Association, 2018) is used as master data. The final step is to link EHR/SEPAS and ICD-10 to detect and repair data quality problems and discover cancer incidences. Even an experienced data scientist cannot tell which algorithm will perform the best before trying different algorithms. So, different record

linkage algorithms and similarity metrics are tried (Elmagarmid et al., 2007) as follows:

1. Character-Based (Kovacevic and Devedzic, 2009):
   - Jaro Distance (Rajabzadeh et al., 2012)
   - N-Grams (Brown et al., 1992)
   - Smith-Waterman (O. Sandes and Melo, 2013)
   - Edit Distance (Su et al., 2008)

2. Token-Based (Wang et al., 2011)
   - Atomic String (Monge and Elkan, 1996)
   - WHIRL (Cohen, 1998)
   - Q-Grams with tf.idf (Gravano et al., 2003)

3. Phonetic (Mielke, 2012)
   - Soundex (Holmes and McCabe, 2002)
   - NYSIIS (Taft, 1970)

4. Numeric (Bertossi et al., 2008)

After performing preprocessing operations (deleting duplicate records, removing spaces, etc.), about 75,000 non-repetitive records of the system including the disease code and disease description, were delivered to the proposed framework. The code and description of the disease should be in accordance with the ICD-10 standard format, but due to data quality problems, there are many inconsistencies (Table 1).

In order to facilitate the process of identifying and solving the problem, according to the first letter of the disease code, diagnostic data is classified into 28 classes presented in Table 2:

Table 2- Classification of disease codes and descriptions

ID	Class	Count	Percentage
1	A	1,211	1.60 %
2	B	921	1.21 %
3	C	2,321	3.06 %
4	D	2,332	3.07 %
5	E	1,328	1.75 %
6	F	1,171	1.54 %
7	G	1,410	1.86 %
8	H	1,145	1.51 %
9	I	2,902	3.83 %
10	J	1,368	1.80 %
11	K	2,709	3.57 %
12	L	955	1.26 %
13	M	6,258	8.25 %
14	N	2,255	2.97 %
15	O	2,304	3.04 %
16	P	986	1.30 %
17	Q	1,585	2.09 %
18	R	2,544	3.35 %
19	S	6,996	9.22 %
20	T	3,978	5.24 %
21	U	91	0.12 %
22	V	988	1.30 %

23	W	1,172	1.55 %
24	X	1,165	1.54 %
25	Y	777	1.02 %
26	Z	1,990	2.62 %
27	Number	22,778	30.03 %
28	Symbol	204	0.27 %

Note: According to the studies, 5 of the 28 classes above (classes: *C*, *D*, *M*, *Z*, and *Numbers*) have cancer cases, and focus is on these classes:

- Class C: Neoplasms ("ICD-10 Chapter II," 2018)
- Class D: Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism ("ICD-10 Chapter III," 2018)
- Class M: Diseases of the musculoskeletal sys. and connective tissue ("ICD-10 Chapter XIII," 2018)
- Class Z: Factors influencing health status and health services ("ICD-10 Chapter XXI," 2018)
- Class Numbers: Erroneous coding format

Among these five classes, the most important class is *Class C* (Neoplasms). The information of this class is checked using the *N-Gram* similarity algorithm with ICD-10 information and its results are presented in Figure 1:

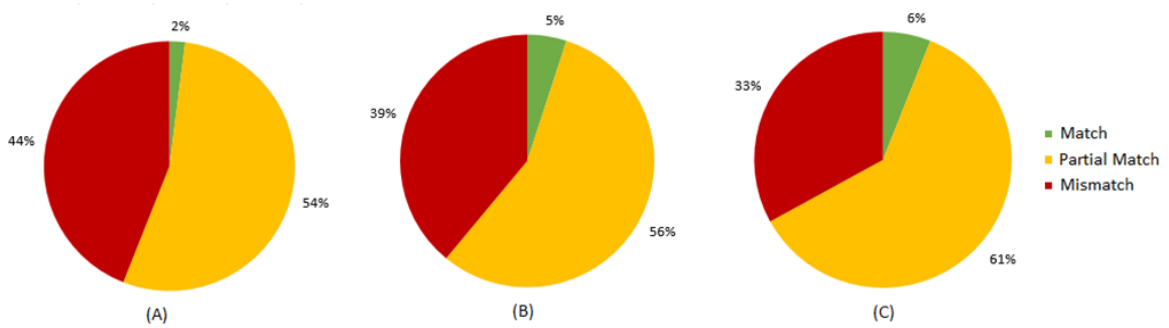

Figure 1: Percentage of source data compliance with ICD-10 (A) Initial mode (B) Ignoring case-sensitive (C) Removing spaces

Figure 1.A displays the initial status of this class (Class C) with only 2% of its codes and descriptions matching ICD-10. Figure 1.B displays the status of this class information after ignoring case-sensitive differences. In this case, 5% of codes and descriptions match ICD-10. Figure 1.C displays the status of the class information after ignoring case-sensitive differences and removing the spaces within codes and descriptions. In this case, 6% of codes and descriptions match ICD-10.

It is found that the status of other classes is similar to the class C. Consequently only a small percentage of registered diagnostic records match ICD-10 standards and most of the records have erroneous code and/or description.

To solve this problem, a data quality engineering framework is designed to match the information recorded in EHR/SEPAS with the standard ICD-10, using customized algorithms.

In Table 3, differences between *Source* (EHR/SEPAS) and *Master* (ICD-10) strings are presented using similarity functions.

The first column (ID) is a simple sequence counter. The second column (Code) is the code of the disease. The third Column (Source Description) is the description of the disease recorded in EHR/SEPAS. The fourth column (Master Description) is the description of the disease recorded in ICD-10. The fifth column (DIFF) is the difference of the third and the fourth column (using a similarity function). Other similarity metrics and algorithms (Elmagarmid et al., 2007) are also tried. Different thresholds and parameters are also used to get the optimal results.

Finally, based on characteristics of the source dataset and types of data quality problems, character-based similarity metrics are used and the comparison process is repeated using customized *Levenshtein* algorithm. In Section 4, the results are compared to the baseline method to estimate *Precision*, *Recall* and *F-measure*.

Table 3- Calculating the differences between source and master strings using similarity functions

ID	Code	Source Description	Master Description	DIFF
1	C02	Otherunspecifiedpartsoftongue	Other Unspecified Parts of Tongue	5
2	C02.9	Tongue, NOS	Tongue, NOS	1
3	C02.9	Tongue, Unspecified	Tongue, NOS	15
4	C07	Malignantneoplasmofparotidgland	PAROTID GLAND	20
5	C11	Nasopharynx	NASOPHARYNX	0
6	C11.9	Nasopharynx, Unspecified	Nasopharyngeal wall	22
7	C13	Hypopharynx	HYPOPHARYNX	0
8	C13.9	Hypopharynx, Nos	Hypopharynx, NOS	11
9	C15	Esophagus	ESOPHAGUS	0
10	C15.9	Esophagus, Unspecified	Esophagus, NOS	18
11	C16	Stomach	STOMACH	0
12	C16.0	Cardia	Gastroesophageal junction	25
13	C16.0	Cardia, Nos	Gastroesophageal junction	25
14	C16.3	Pyloricantrum	Gastric antrum	14
15	C16.9	Stomach, Unspecified	Stomach, NOS	16
16	C17	Smallintestine	SMALL INTESTINE	1
17	C18	Colon	COLON	0
18	C18.0	Cecum	Cecum	4
19	C18.1	Appendix	Appendix	7
20	C18.7	Sigmoidcolon	Sigmoid colon	12

Experimental Results

In this section, results of the proposed solution are presented. According to the classification presented in the previous sections (Section 3), it is shown that the four classes: *C*, *D*, *M,* and *Z*, are the most important identified classes. Remaining classes are merged into one class, referred to as "*Others"*.

ICD-10 is used as a master dataset and similarity function is used to find related records in *EHR-SEPAS*. Previously, expert users performed this process manually, which was very expensive, time consuming and erroneous. Results of this manual process are used as a baseline.

In order to compare results of the proposed method (Table 5) with the baseline method (Table 4), a set of data including 47,912 labeled (cancerous or non-cancerous) records of *EHR-SEPAS* is used. The detailed results of the studies are presented in Table 4 and Table 5 indicating that the proposed method outperforms baseline method in terms of *Precision*, *Recall* and *F-Measure*.

Table 4- Baseline method: *Precision*, *Recall*, and *F-Measure*

Classes	**Precision**	**Recall**	**F-Measure**
C	100	100	100
D	99.48	99.07	99.28
M	100	41.89	59.05
Z	97.83	31.03	47.12
Others	74.07	22.47	34.48
Total	**99.27**	**81.91**	**89.76**

Table 5- Proposed method: *Precision*, *Recall*, and *F-Measure*

Classes	Precision	Recall	F-Measure
C	100	100	100
D	100	99.90	99.95
M	100	99.78	99.8
Z	100	100	100
Others	79.28	100	88.44
Total	**99.01**	**99.93**	**99.47**

From Figure 2, it can be concluded that our proposed method can find more cases compared to the baseline method (*False Negative* is very low) and increase *Recall* and *F-Measure* significantly.

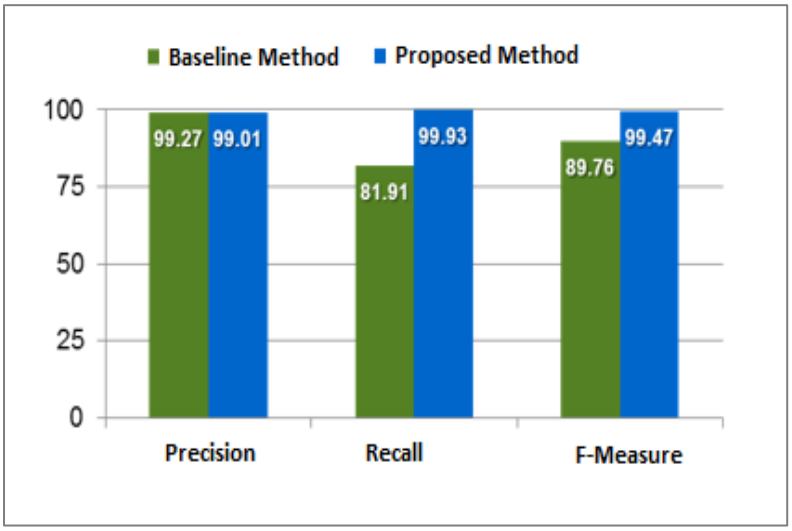

Figure 2. Comparison of Precision, Recall and F-Measure of the proposed method and the baseline method

Conclusion

In this paper, a data quality engineering framework is presented for case finding from *EHR-SEPAS* to improve completeness of the *Population Based Cancer Registry*. Presence of human factors in the process of electronic data recording causes some data quality challenges. Using similarity functions (customized Levenshtein algorithm) and master data (ICD-10), a framework is developed to solve these problems. The proposed framework is applied to a *Population Based Cancer Registry* program.

The proposed framework is validated using experimental evaluations and it is compared with a baseline method in terms of *Precision*, *Recall* and *F-Measure*. Finally, some experimental results are presented to show effectiveness of the proposed framework.

As a future work, the proposed framework can be extended to use a crowdsourcing (Ke et al., 2018; Nazariani and A. Barforoush, 2017; Wang et al., 2012) approach (instead of using expensive expert users) for data quality engineering, and finding cancerous cases which cannot be detected using algorithms.

In (Ebraheem et al., 2018; Mudgal et al., 2018) the advantages and limitations of deep learning (DP) models when applied to a diverse range of entity matching tasks, specifically entity matching over structured, textual, and dirty data is examined. Another direction for future studies is applying deep learning (DL) module to entity matching function for cancerous case finding, that is efficient and accurate.

References

Association, A.M., 2018. ICD-10-CM 2019 the Complete Official Codebook, 1 edition. ed. Amer Medical Assn, S.l.

Bertossi, L., Bravo, L., Franconi, E., Lopatenko, A., 2008. The complexity and approximation of fixing numerical attributes in databases under integrity constraints. Inf. Syst., Selected Papers from the Tenth International Symposium on Database Programming Languages DBPL 2005 33, 407–434.

Brown, P.F., deSouza, P.V., Mercer, R.L., Pietra, V.J.D., Lai, J.C., 1992. Class-based N-gram Models of Natural Language. Comput Linguist 18, 467–479.

Cohen, W.W., 1998. Integration of Heterogeneous Databases Without Common Domains Using Queries Based on Textual Similarity, in: Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data, SIGMOD '98. ACM, New York, NY, USA, pp. 201–212.

O. Sandes, E.F., Melo, A.C.M.A., 2013. Retrieving Smith-Waterman Alignments with Optimizations for Megabase Biological Sequences Using GPU. IEEE Trans. Parallel Distrib. Syst. 24, 1009–1021.

Ebraheem, M., Thirumuruganathan, S., Joty, S., Ouzzani, M., Tang, N., 2018. Distributed Representations of Tuples for Entity Resolution. Proc VLDB Endow 11, 1454–1467.

Elmagarmid, A.K., Ipeirotis, P.G., Verykios, V.S., 2007. Duplicate Record Detection: A Survey. IEEE Trans. Knowl. Data Eng. 19, 1–16.

Ferlay, J., Colombet, M., Soerjomataram, I., Dyba, T., Randi, G., Bettio, M., Gavin, A., Visser, O., Bray, F., 2018. Cancer incidence and mortality patterns in Europe: Estimates for 40 countries and 25 major cancers in 2018. Eur. J. Cancer.

Gravano, L., Ipeirotis, P.G., Koudas, N., Srivastava, D., 2003. Text Joins in an RDBMS for Web Data Integration, in: Proceedings of the 12th International Conference on World Wide Web, WWW '03. ACM, New York, NY, USA, pp. 90–101.

Hoerbst, A., Ammenwerth, E., 2010. Electronic Health Records. Methods Inf. Med. 49, 320–336.

Holmes, D., McCabe, M.C., 2002. Improving precision and recall for Soundex retrieval, in: Proceedings. International Conference on Information Technology: Coding and Computing. Presented at the International Conference on Information Technology: Coding and Computing. ITCC 2002, IEEE Comput. Soc, Las Vegas, NV, USA, pp. 22–26.

ICD-10 Chapter II: Neoplasms, 2018. . Wikipedia.

ICD-10 Chapter III: Diseases of the blood and blood-forming organs, and certain disorders involving the immune mechanism, 2018. . Wikipedia.

ICD-10 Chapter XIII: Diseases of the musculoskeletal system and connective tissue, 2018. Wikipedia.

ICD-10 Chapter XXI: Factors influencing health status and contact with health services, 2018. Wikipedia.

Ke, X., Teo, M., Khan, A., Yalavarthi, V.K., 2018. A Demonstration of PERC: Probabilistic Entity Resolution with Crowd Errors. Proc VLDB Endow 11, 1922–1925.

Kovacevic, A., Devedzic, V., 2009. Duplicate Journal Title Detection in References. Handb. Res. Digit. Libr. Des. Dev. Impact 235–242.

Leinonen, M.K., Miettinen, J., Heikkinen, S., Pitkäniemi, J., Malila, N., 2017. Quality measures of the population-based Finnish Cancer Registry indicate sound data quality for solid malignant tumours. Eur. J. Cancer 77, 31–39.

Mielke, J., 2012. A phonetically based metric of sound similarity. Lingua, Phonological Similarity 122, 145–163.

Monge, A., Elkan, C., 1996. The field matching problem: Algo-rithms and applications, in: In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. pp. 267–270.

Mudgal, S., Li, H., Rekatsinas, T., Doan, A., Park, Y., Krishnan, G., Deep, R., Arcaute, E., Raghavendra, V., 2018. Deep Learning for Entity Matching: A Design Space Explora-tion, in: Proceedings of the 2018 International Confer-ence on Management of Data, SIGMOD '18. ACM, New York, NY, USA, pp. 19–34.

Nazariani, M., A. Barforoush, A., 2017. A Crowdsourcing Ap-proach to Data Quality Engieering, in: Proceeding of 22th CSICC. Tehran, Iran.

Nazariani, M., et al., 2016a. Using Electronic Health Record System for Case Finding and Improving Completeness of Population-Based Cancer Registry, in: The 38th Annual IARC Conference. Marrakech, Moroco.

Nazariani, M., A. Barforoush, A., 2016b. Introducing the Requirements of Data Quality Engineering Framework, in: The 8th National and 2nd International Electronic Commerce and Economy Conference (ECEC 2016). Tehran, Iran.

Nazariani, M., A. Barforoush, A., 2015. Constraint-Based Data Cleaning: A Survey. Amirkabir Univ. Technol.

Rajabzadeh, M., Tabibian, S., Akbari, A., Nasersharif, B., 2012. Improved dynamic match phone lattice search using Viterbi scores and Jaro Winkler distance for keyword spotting system, in: The 16th CSI International Symposium on Artificial Intelligence and Signal Processing (AISP 2012). Presented at the 2012 16th CSI International Symposium on Artificial Intelligence and Signal Processing (AISP), IEEE, Shiraz, Fars, Iran, pp. 423–427.

Robson, B., Purdie, G., Cram, F., Simmonds, S., 2007. Age standardisation – an indigenous standard? Emerg. Themes Epidemiol. 4, 3.

Rumball-Smith, J., Shekelle, P., Damberg, C.L., 2018. Electronic health record "super-users" and "under-users" in ambulatory care practices. Am. J. Manag. Care 24, 26–31.

Sateren, W.B., Trimble, E.L., Abrams, J., Brawley, O., Breen, N., Ford, L., McCabe, M., Kaplan, R., Smith, M., Unger-leider, R., Christian, M.C., 2002. How sociodemo-graphics, presence of oncology specialists, and hospital cancer programs affect accrual to cancer treatment trials. J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol. 20, 2109–2117.

Su, Z., Ahn, B.-R., Eom, K.-Y., Kang, M.-K., Kim, J.-P., Kim, M.-K., 2008. Plagiarism Detection Using the Levenshtein Distance and Smith-Waterman Algorithm, in: 3rd International Conference on Innovative Computing Information and Control, 2008. ICICIC '08. Presented at the 3rd International Conference on Innovative Computing Information and Control, 2008. ICICIC '08, pp. 569–569.

Taft, R.L., 1970. Name search techniques. Bureau of Systems Development, New York State Identification and Intelligence System.

Wang, J., Kraska, T., Franklin, M.J., Feng, J., 2012. CrowdER: Crowdsourcing Entity Resolution. Proc VLDB Endow 5, 1483–1494.

Wang, J., Li, G., Fe, J., 2011. Fast-join: An efficient method for fuzzy token matching based string similarity join, in: 2011 IEEE 27th International Conference on Data Engineering. Presented at the 2011 IEEE International Conference on Data Engineering (ICDE 2011), IEEE, Hannover, Germany, pp. 458–469.

Weiskopf, N.G., Weng, C., 2013. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J. Am. Med. Inform. Assoc. 20, 144–151