International Journal of Engineering & Technology, *7 (3.34) (2018) 103-106*

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Relational Forecast Limiter Algorithm for ICD based EMRs

Nithya.M1, 2, Sheela.T3

1Research Scholar, Faculty of Computer Science & Eng., Satyabhama Institute of Science and Technology
2Assistant Professor, Computer Science & Eng., Sri Sairam Engineering College, Chennai, India,
3Professor, Department of Information & Technology, Sri Sairam Engineering College, Chennai, India
**Corresponding author E-Mail: 1nithya0817@gmail.com, 2nithya.cse@sairam.edu.in, 3hot.it@sairam.edu.in*

Abstract:

Forecasting individual privacy based on acquired knowledge across associated or related diseases is often a concern during data publishing. EMR (Electronic Medical Record) of an individual have more than one associated diseases which are potential knowledge nodes for analyst to exploit privacy. RFL (Relational Forecast Limiter) algorithm aims in reducing the forecast or prediction level by introducing generalization, suppression and noise addition techniques based on relational forecast detection and relational forecast height of diseases classified in ICD (International Classification of Diseases) table. These techniques delimit the forecasting capability to bring privacy under control. Generalization and suppression techniques are applied on sensitive attributes while noise addition is applied on quasi identifiers. Generalization is realized at lower twig and branch level, while suppression and noise addition are realized at bough level. Primary objective of the algorithm focuses on sharing minimum privacy data enabling the data analyst to extract maximum useful information. Accuracy is retained to ensure data analysis yields useful information for social causes. Experimental results on privacy and accuracy loss demonstrates algorithm efficiency.

Keywords: *Accuracy, Generalization, Noise, Privacy, Suppression*

1. Introduction

Typical EMR system houses various hospitalization stages starting from registration, lab results, and Pharmacy & radiology data and finally billing information. Registration stage collects identifier information (Name) of the patient, including their quasi identifiers (age, sex, location). Lab result adds blood & culture results, DNA pattern, and disease identification. Generally the patients are associated with more than one disease. And the diseases are inter related with each other. This inter relation is due to various reasons like common cause, consumption of specific drug, geographic location, secondary symptoms and other natural causes. Interrelated diseases have hidden knowledge buried in them. When they are picked up for a good cause, they provide useful information for the society. However if the hidden knowledge used impacts privacy of the individual, it becomes an infringement of law. Therefore it becomes mandatory to break the association between related diseases.

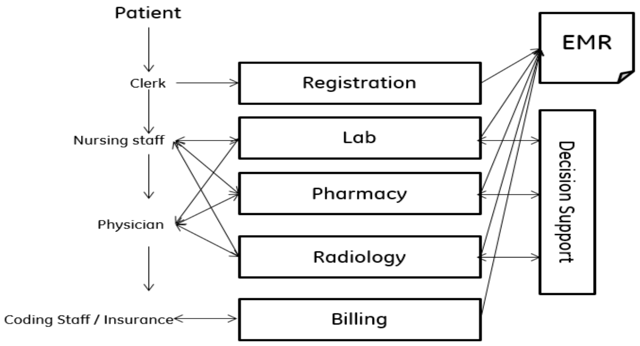

Fig .1: EMR System

Figure.1 shows standard EMR system showing interfaces and stakeholders involved in handling the EMR database. Individual EMR records are updated at various stages and collated at the end. EMR database is stored in the hospital medical system. When the records are requested for analysis, sufficient care should be exercised in cleaning, formatting and presenting. Apart from the format and intensity of data released, type and intension of receiver who does the data analysis should be considered. The data analyst can be a third party agency, research student appointed by medical council or medical data analyst expert in the hospital. Prior NDU should be signed to ensure data integrity is preserved. Next level of scrutiny looks for what level of data to be shared. Will it be individual EMR records or aggregated summary of records? Sharing individual EMR records is vulnerable for privacy attacks when compared to aggregated data. However the decision on type of data to be shared solely depends on the purpose and objective of analysis.

2. Motivation for Relational Forecast Limiter Algorithm

Past research studies dealing with EMR records have thrown light on delimiting the associated diseases when publishing EMR records. However they suffer from limitations. Proposed relational forecast limiter algorithm is developed considering the drawbacks of conventional systems. Khaled El Emam et.al (2009) proposed OLA (Optimal Lattice Anonymization) algorithm which satisfies K-anonymity and defines suppression and generalization requirements. However the algorithm only uses suppression and generalization de-identification methods. Proposed RFL algorithm uses noise addition methods in addition to suppression and generalization. Khaled El Emam (2011) discusses masking, generalization and suppression methods which can be applied on electronic medical data for de-identification. Amit Thakkar et.al

(2015) proposed correlation based anonymization algorithm using generalization and suppression for disclosure problems. However method of calculation of privacy and accuracy loss is not evident in the research work. Carlos Moque et al. (2012) proposed an interactive tool for medical dataset anonymization. This method lacks user interaction for selecting the parameters for anonymization. Wayne Newhauser et al. (2014) proposed a new method for anonymizing radiation therapy treatment plans which would retain data privacy. Major limitation of this method was it only tested radiation therapy treatment plans and not EMR records. Thus this method can't be generalized and used in EMR medical records. Acar Tamersoy et al. (2012) proposed a unique method to publish individual longitudinal data that retains privacy while offering maximum accuracy. This method offers definitive information loss due to generalization and suppression techniques used. Further this method suffers from computational challenges when dealing with large datasets. Raymond Heatherly et al. (2016) proposed anonymization on 3 medical centres using conventional k- anonymization algorithm with k as 5. Further the algorithm used is heuristic and there is no assurance for optimization. Soohyung Kim et al. (2017) compares 3 types of anonymization methods for EMR data cube. EMR data cubes are aggregated data with complex EMR attributes. Analysis shows the data accuracy widely varies for each method and the usage of the method is purely based on application environment. Adebayo Omotosho and Justice Emuoyibofarhe (2014) emphasises on bio cryptographic techniques for encrypting and anonymizing medical data records. This method requires accountability of handling crypto keys in order to guarantee privacy. However accountability cannot be measured and trusted. Melanie L. Balestra (2017) lists down methods and best practices for handling EMRs by medical personal, in order to ensure privacy information is retained. The journal also narrates about the risks involved in wrong handling of EMR data sets. Mathai N et al. (2017) highlights various potential issues which might occur during EMR processing. Further the paper gives recommendations and preventive actions to avoid improper handling of EMR VijayaKumar.K, Arun.C et al. (2017) discussed about the framework for the cloud based application for the medical research and provide them the continuous classification of diseases.

3. Architecture Design

Relational forecast limiter algorithm is a novel technique which has multiple stages of anonymizing the EMR dataset. Raw EMR dataset is cleaned and sanitized by removing the identifiers like patient's name.

Figure.2 shows the architecture of RFL algorithm. ICD 10 Version: 2016 classification table is used for calculating the forecast relation height. The classification table has 22 chapters of disease classifications.

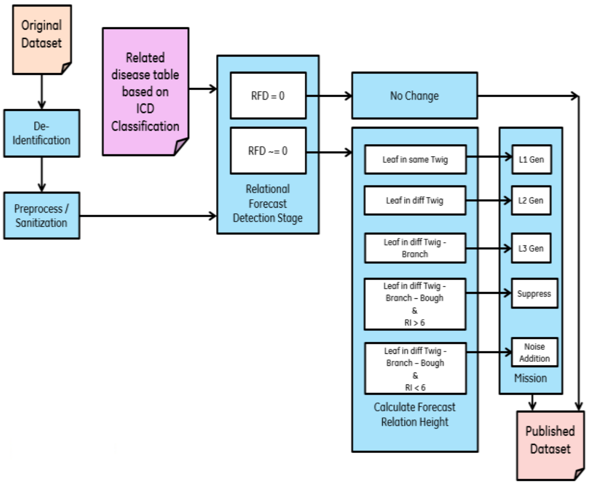
Original Dataset
Related disease table based on ICD Classification
De-Identification
Preprocess / Sanitization
RFD = 0
RFD ~= 0
Relational Forecast Detection Stage
No Change
Leaf in same Twig
Leaf in diff Twig
Leaf in diff Twig - Branch
Leaf in diff Twig - Branch – Bough & RI > 6
Leaf in diff Twig - Branch – Bough & RI < 6
Calculate Forecast Relation Height
L1 Gen
L2 Gen
L3 Gen
Suppress
Noise Addition
Mission
Published Dataset

Fig. 2: Architecture Design

Table.1: ICD vs. Algorithm Hierarchy

ICD Hierarchy	Algorithm Hierarchy	Description
ICD-10	Trunk	Version:2016
Chapter	Bough	IV Endocrine, nutritional & metabolic diseases
Disease Group	Branch	E10-E14 Diabetes mellitus
Disease Sub Group	Twig	E10-Type 1 diabetes mellitus
Disease	Leaf	E10.2-Type 1 diabetes with renal complications

Table.1 shows a typical example of levels as per ICD10 classifications. RFL algorithm relates the ICD levels with dedicated naming structure. The mapping of the naming structure (Trunk, Bough, Branch, Twig and Leaf) are shown in Table.1. EMR tuples have more than 1 disease recorded and the diseases are often associated closely with each other. It is necessary to break the association between them to ensure hidden knowledge not transferred to the data analyst. The intensity of the association is calculated based on the risk identification of further useful information from the EMR record and the position / location of associated diseases at various levels as per ICD classification. The de-identification method is also defined based the above conditions.

Table.2: Level based Anonymization

RFD	RFH		RI	Mission
RFD ~=0	Leaf	Leaf in same Twig		L1Generalization
RFD ~=0	Leaf	Leaf in diff Twig		L2Generalization
RFD ~=0	Leaf	Leaf in diff Twig - Branch		L3Generalization
RFD ~=0	Leaf	Leaf in diff Twig - Branch - Bough	RI > 6	Suppression
RFD ~=0	Leaf	Leaf in diff Twig - Branch - Bough	RI < 6	Noise Addition
RFD = 0	Leaf	Any Combination		No Change

Table.2 shows level based anonymization triggered by the RFL algorithm. The relational forecast detector compares each EMR tuple against the related disease table based on ICD classification. If it finds relation between any 2 diseases in the tuple, then it flags RFD (Relational Forecast Detector). If there is no significant relation between diseases in the tuple, then it de-flags RFD and the tuple is not disturbed and is published without any change. This denotes the tuple doesn't have any sensitive hidden knowledge and is free for publication. The flagged RFD tuples are further screened based on RFH (Relational Forecast Height) and RI (Risk Identification) flags. If any 2 diseases in the tuple fall between two leafs in the same twig then the algorithm triggers L1Generalization anonymization method. If any 2 diseases in the tuple fall between two leafs in different twigs then the algorithm triggers L2Generalization anonymization method. If any 2 diseases in the tuple fall between two leafs in different branches then the algorithm triggers L3Generalization anonymization method. If any 2 diseases in the tuple fall between two leafs in different boughs then the algorithm further checks RI flag. If RI >6, denoting high risk identification, the algorithm triggers suppression anonymization method. If RI <6, denoting lower risk identification, the algorithm triggers noise addition anonymization method.

4. Algorithm

Table.3: Relational Forecast Limiter Algorithm

A	***Predictive Algorithm***
1	L[i…..N] = No of line items in the publishing list
2	For i=1
3	If RFD = RFD[0]
4	<No Change>

```
5   else If RFD ~= RFD[0]
6   {
7   If RFH = Leaf in same Twig, then Mission = L1Generalization
8   else if RFH = Leaf in diff Twig, then Mission = L2Generalization
9   else if RFH = Leaf in diff Twig - Branch, then Mission = L3Generalization
10  else RFH = Leaf in diff Twig - Branch - Bough
11  {
12  If RI > 6
13  Mission = Suppression
14  else
15  Mission = Noise Addition
16  }
17  }
18  I = i++, roll back to step 3 until i = N
```

Table.3 shows Relational Forecast Limiter Algorithm implementation. Generalization is applied on related diseases which are in the same or different twig and branches. This is mainly due to ease of anonymizing with respect to ICD naming structure. Further the diseases form a subset which makes it easy for generalization. When the related diseases fall across different boughs then generalization cannot be applied, as it becomes difficult to converge based on naming structure. Related diseases across different boughs are more risker and thus needs more anonymization. Suppression or noise addition is applied in these conditions. If the risk of identifying hidden knowledge is more then suppression is applied. Both the associated diseases are removed. This will reduce the accuracy of published data, but would retain the data privacy. If the risk of identifying hidden knowledge is less, then the quasi identifier (age) is altered to introduce anonymization. In this case the associated diseases are published as it is.

Table.4: Sanitised dataset

Sex	Age	Location	ICD Codes
F	32	60032	**E11.4 , E11.2**, G11.1, S10.1
M	56	60054	J95.9, P52.2, Q10.2
M	12	60021	P05.2, **E11.7, E66.0**
F	73	60098	**N17.3, E11.5,** R30.1
M	25	60044	O31.1, S11.9
M	66	60058	**N17.0**, V01.9, **N18.3**

Table.4 shows the sanitised EMR dataset to be fed to RFL algorithm. Age, Sex and Location are quasi identifiers and disease being sensitive attribute. The highlighted ICD codes are associated with each other in the tuples and they have potential hidden knowledge. The algorithm runs on each EMR tuple and logs the RFD, RFH and RI flags to determine the correct anonymization method to be applied. The mission flag defines the final anonymization method.

Table.5: RFL applied to Sanitised dataset

ICD Codes	RFD	RFH	RI	Mission
E11.4 , E11.2, G11.1, S10.1	RFD(1,2)	Leaf in same Twig		L1Generalization
J95.9, P52.2, Q10.2	RFD(0)	Leaf in diff Twig - Branch - Bough		No change
P05.2, **E11.7, E66.0**	RFD(2,3)	Leaf in diff Twig - Branch		L3Generalization
N17.3, E11.5, R30.1	RFD(1,2)	Leaf in diff Twig - Branch - Bough	RI > 6	Suppression
O31.1, S11.9	RFD(1,2)	Leaf in diff Twig - Branch - Bough	RI < 6	Noise Addition
N17.0, V01.9, **N18.3**	RFD(1,3)	Leaf in diff Twig		L2Generalization

Table.5 shows the impact of RFL algorithm applied on sanitised dataset. Leafs in the tuples 1, 6 and 3 fall in different twigs and branches resulting in Generalization algorithm missioned. Leafs in tuple 4 and 5 fall in different boughs. As a result suppression and noise identification method is missioned. As RI > 6 for tuple 4, suppression method is applied and noise addition method applied for tuple 5, as RI < 6. Tuple 2 is left undisturbed as there are no associated diseases and hidden knowledge identified.

Table.6: Anonymized dataset

Sex	Age	Location	ICD Codes
F	32	60032	**E11.* , E11.***, G11.1, S10.1
M	56	60054	J95.9, P52.2, Q10.2
M	12	60021	P05.2, **E*, E***
F	73	60098	R30.1
M	**82**	60044	O31.1, S11.9
M	66	60058	**N1***, V01.9, **N1***

Table.6 shows the anonymised dataset with different anonymization methods applied on the tuples. (E11.*, E11.*), (E*, E*), (N1*, N1*) denotes impact of generalization. ICD codes N17.3 and E11.5 are removed from tuple 4, due to impact of suppression. Age value is modified from 25 to 82 in the tuple 5, due to impact of noise addition. Noise addition is applied on the quasi identifiers and not on the sensitive attributes.

5. Test Setup

EMR data set is extracted from http://www.emrbots.org/. This site holds simulated tuples of 5k size which replicate EMR dataset. The raw data from the website is further cleaned and sanitised for analysis. A standard windows desktop with 3 GB RAM and Windows 7 operating system will run the algorithm.

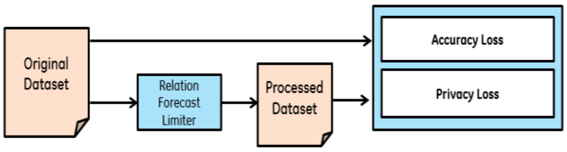

Fig. 3: Test Setup

Figure.3 shows test setup required for calculating accuracy and privacy loss. Both the losses are calculated by comparing the original dataset and sanitised dataset. 3 sets of 5K EMR datasets are used for analysis. First EMR data set had 70% of the tuples with associated diseases falling at different twigs and branches. Thus this set had 70% of tuples imposed with generalization anonymization. Second EMR dataset had 70% of the tuples with associated diseases falling at different boughs with RI>6. Thus this set had 70% of tuples imposed with suppression anonymization. Third EMR dataset had 70% of the tuples with associated diseases falling at different boughs with RI<6. Thus this set had 70% of tuples imposed with noise addition anonymization.

6. Experimental Results

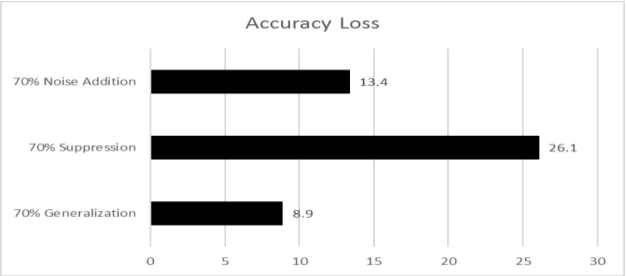

Fig. 4: Accuracy Loss

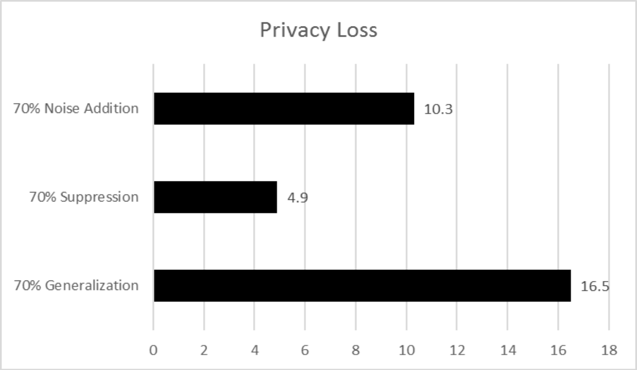

Fig .5: Privacy Loss

Figure 4 & 5 shows the accuracy and privacy loss plotted for various cases of anonymized dataset. Dataset with 70% Generalization anonymization has better accuracy compared to privacy. Generalization hampers the data to lesser extent compared to other two suppression techniques. This is evident by seeing the results of suppression and noise addition techniques. Accuracy loss is high in noise addition and higher in suppression methods. As noise addition anonymizes only the quasi identifiers and not the sensitive attributes the impact on accuracy loss is lower.

7. Conclusion

Proposed Relational Forecast Limiter algorithm offers balanced anonymization on the EMR dataset. It combines generalization, suppression and noise addition methods which are mutually exclusive with respect to accuracy and privacy factors. Maintaining a balance between privacy and accuracy would yield better averaged results when compared to conventional methods. This algorithm avoids hidden knowledge related to associated diseases are not transmitted during data publishing stage. However if majority of the population of diseases which holds hidden knowledge falls across different boughs, then the overall accuracy loss increases. Similarly if the majority of the population of diseases which holds hidden knowledge falls across different twigs and branches, then the overall privacy loss increases. Thus the algorithm works based on concentration of hidden knowledge at various hierarchy levels.

References

[1] Acar Tamersoy et al. (2012). Anonymization of Longitudinal Electronic Medical Records. IEEE Trans Inf Technol Biomed. May; 16(3): 413–423.

[2] Adebayo Omotosho and Justice Emuoyibofarhe (2014). A Criticism of the Current Security, Privacy and Accountability Issues in Electronic Health Records. International Journal of Applied Information Systems. Volume 7– No.8, September.

[3] Amit Thakkar et al. (2015). Correlation Based Anonymization Using Generalization and Suppression for Disclosure Problems. Advances in Intelligent Informatics pp 581-592.

[4] Carlos Moque et al. (2012). AnonymousData.co: A proposal for interactive anonymization of Electronic Medical Records. SciVerse ScienceDirect, Procedia Technology 743-752.

[5] Khaled El Emam et al. (2009). Globally Optimal k-Anonymity Method for the De-Identification of Health Data. J Am Med Inform Assoc. Sep-Oct; 16(5): 670–682.

[6] Khaled El Emam (2011). Methods for the de-identification of electronic health records for genomic research. Genome Med; 3(4): 25.

[7] Mathai N et al. (2017). Electronic Health Record Management: Expectations, Issues, and Challenges. Journal of Health & Medical Informatics. DOI: 10.4172/2157-7420.1000265.

[8] Melanie L. Balestra (2017). Electronic Health Records: Patient Care and Ethical and Legal Implications for Nurse Practitioners. The journal of Nurse Practitioners. Volume 13, Issue 2, Pages 105–111.

[9] Raymond Heatherly et al. (2016). A multi-institution evaluation of clinical profile anonymization. Journal of the American Medical Informatics Association, Volume 23, Issue e1, 1 April, Pages e131–e137.

[10] Soohyung Kim et al. (2017). Privacy-preserving data cube for electronic medical records: An experimental evaluation. International Journal of Medical Informatics, Volume 97, January, Pages 33-42.

[11] Wayne Newhauser et al. (2014). Anonymization of DICOM Electronic Medical Records for Radiation Therapy. Comput Biol Med. Oct 1; 0: 134–140.

[12] K. Vijayakumar, C. Arun, Analysis and selection of risk assessment frameworks for cloud based enterprise applications", Biomedical Research, ISSN: 0976-1683 (Electronic), January 2017.

[13] K. Vijayakumar, C.Arun, Automated risk identification using NLP in cloud based development environments Ambient Intel