Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 7.056

IJCSMC, Vol. 9, Issue. 4, April 2020, pg.49 – 60

ICT Diffusion and Primary Care Methodological Contribution on Clustering Methods to Partition Medical Practices in the USA

Professor Christine C Huttin[1,2]
[1]Professor, Aix Marseille University, Aix en Provence, France
[2]CEO and researcher, Endepusresearch, Inc, Cambridge, USA

Abstract:
OBJECTIVES: *This project presents an analysis of IT processes on variations of prescribing patterns for patients diagnoses with diabetes type II. It follows a first study on electronic billing and analyses various stages of IT processes in clinical practices.*
METHODS: *A sample of 610 patients is extracted from the CDC physician survey (Huttin/Wong dataset 2010).*
RESULTS: *Two Hierarchical clustering methods: Average Linkage (AL) and Ward were used and lead to a partitioning of medical records in three clusters, showing significant differences in levels of computerized clinical information (Savage Test).*
CONCLUSIONS: *more research is needed to include other clustering techniques and a generalization with a matrix of similarity, possibly with an optimization of multi objectives. This type of research can be used for analysis and management of propagation of IT processes inside clinical systems and control for their effects on physician prescribing behaviors.*

Keywords: clustering, medical records, claims, categorical data, ICT diffusion

I. Introduction

This paper aims to continue a research stream on ICT diffusion in primary care, it is part of a more global agenda on digital health and the digitalization of both supply and demand in medical markets; it discusses an empirical study using clustering methods to analyze ICT diffusion on medical practices; the study is based on a statistical analysis on US diabetic patients, from a dataset designed with the National Ambulatory Care Survey (NAMCS) from the Center for Disease Control (CDC).

It can be used as a technical note to also complement the methodological paper entitled "stated and revealed preference issues: product, patient and physicians 'attributes in choice experiments" [1]. In this 2017 paper, several advances for study designs for economic studies on physicians' choices were described, in the move towards precision medicine. This note aims to provide a description of the methods used to examine the ICT diffusion process in primary care practices. The 2006 CDC data set may serve as a baseline for further development with additional data elements on computerization of practices with time-series medical databases. The statistical modeling of physicians' choices can then control some impact of ICT diffusion process in primary care. The paper first describes the data elements specific to the survey design instrument to measure and analyze the effects of the computerization of medical practices. The second section presents the clustering methods used to partition the physicians into various groups; results of the diabetic study are then used to explain the methodological choices and their limitations. This clustering analysis of medical claims and records is useful to explain some sources of variations in prescribing patterns and impacts of electronic billing and computerization of medical data on physicians' treatment choices.

II. Data Elements used for Measurement of ICT Diffusion in Medical Practices from the CDC Physician Survey

In order to measure and analyze the implementation of IT process in practices, the CDC designed new elements from its National Ambulatory Care Survey. This note only describes the elements integrated in the 2006 dataset, with questions related to the ways IT affects both billing operations and medical records (types of electronic medical records (EMRs) with or without patient demographic information, computerized order for prescriptions, for tests and for lab results, electronic nurse and physician notes, inclusion of reminders in the EMRs). This move to digitalization may impact patterns of use and expenditures, with heterogenous effects according to the type of public and private insurance.

Data elements on computerization of financial information:

The main element of the survey on financial information aims to measure the switch from paper to electronic billing of claims in medical practices; the type of information system in use is not directly described in the survey. So, the following question was added in the survey design:

"Does this practice submit claims electronically?"

It mainly reflects a characteristic of the practice rather than a mode of payment associated to health plans' characteristics. The CDC statisticians opted to reduce the list of questions referring to various types of payments and add a survey question to measure the major change referring to the processing of claims via an electronic versus a non-electronic system. For an analysis of medical records, this type of question may also require controlling nested effects of ICT diffusion with computerization of records from sampled patients possibly seen in the same practice.

Data elements on computerization of clinical information:

The medical record is the information transformed to an electronical medical record (EMR); the data elements added in the survey however are more comprehensive than for financial information in this CDC physician' survey; they tend to also measure whether or not a practice opts for a computerization of its basic electronic medical records and add several survey questions to examine what information is computerized. Table 1 provides descriptive statistics from the diabetic study, using 2006 CDC data elements, and after treatment of missing values for electronic billing and EMR use and recoding of the number of EMR use. It clearly identified a slow adoption of computerization for clinical information but showed that electronic billing was the most diffused IT process implemented in the practices.

Table 1 Descriptive statistics on electronic processes for billing systems and medical records

	Yes	No	Other categories (blank, unknown, missing)	Number of Obs Total sample
Ebilling				
Percent	84.16 %	13.60%		610
N	526	85		
EMR use (rec)				
Percent	28.69 %	71.31%		610
N	175	435		
Of which:				
All electronic				
N of Obs	83			
Part paper, part electronic				
N of Obs	92			
Three main Criteria used on EMR (1)	For Clustering	Analysis	of Medical Records and Computerization	
C1: Computerized orders for Rx				
Percent	16.11%	83.89%		
N of Obs	97	505	8	610
C2: Computerized orders for test				
Percent	12.58%	87.42%		
N of Obs	76	528	6	610
C3: Computerized orders for lab results				
Percent	17.52%	82.48%		
N of Obs	106	499	5	610

(1) Source: Prof Huttin CC, Endepusresearch, Inc, Cambridge, USA, Ten Criteria for different types of computerized information have been used for the clustering analysis, available upon request.

III. Clustering Methods for Partitioning Medical Practices

As the results tended to show the slow but also heterogeneous diffusion/ propagation of computerization of types of information in Electronic Medical Records, it is useful to discuss the types of useful methods to analyze this diffusion process, and not only whether the medical practice adopted or not the electronic records. This technical note mainly focused on the methodological approach used for partitioning medical practices in different stages of computerization, in complement to the statistical analysis of the Diabetic type II studies [1,2]. A selection of clustering methods was tested and two of them lead to very significant results to control the effect of computerization on patterns of drug use and expenditures. Further research is currently under progress for more methodological improvements; but it needs to incorporate the additional data elements provided by CDC statisticians on the computerization process of primary care practices over the years. Clustering methods have proved to be useful in many fields of science since the early 60's (business, medical/biomedical sciences, computer sciences) (e.g.3,4,5,6,7]. They are also useful algorithms to analyze natural patterns and partition datasets into clusters on some similarity/dissimilarity metrics, or supervised learning for classifiers [8], where the number of clusters may or may not be known a priori. Such techniques appear very useful for partitional medical record databases [9,10]. In the scope of this technical note, the clustering analysis is a component to the dataset used for predictive economic models on diabetic care; it helps to partition medical records in different levels of propagation of IT process (for both economic and clinical information) in sampled primary care groups.

The two selected methods use different approaches of hierarchical clustering: Between groups (HB) and Ward method; they agglomerate pairs of points. Both use natural partitioning and do not fix a priori the number of clusters. Both lead to significant variations between three clusters of medical records, with different levels of computerization. Results with other methods such as the K Means or the simple linkage method did not appear as significant. In the K Means method especially, the number of K clusters is set before running the analysis; while in the two selected methods, a natural positioning is used, either with an ascending or descending order. The following section provides a more detailed descriptions of these methods also called "unweighted pair-group" methods using arithmetic averages.

III.1 Linkages Methods and Medical Records' Analysis

The linkage method, called average method, was successfully applied on the dataset, using the SAS proc Cluster command. In this method, "The selection of clusters with the average linkage method is to take the intergroup dissimilarity to be the average dissimilarities between all pairs of objects" [4,10]. This average linkage method clusters by the average of all links within a cluster.

If SUMi is the sum of all pairwise similarities among entities within a cluster i

If SUMi is the sum of all pairwise similarities entities within cluster j

Sij a pairwise combination

Ni the number of entities in cluster i

Nj the number of entities in cluster j

The Average within group similarity for the clusters i and j is

$$\frac{SUMi + SUM\ j + sij}{(Ni+Nj)\ (Ni+Nj-1)/2}$$

"This method does not depend on extreme values for defining clusters, no statement can be made about the minimum or maximum of each cluster". It is quite suitable for medical datasets used for health care statistical demand model since usually distributions are very skewed and the method allows to use clustering event with a few outliers. This average linkage clustering analysis helped to identify three clusters representing different stages of IT processes:

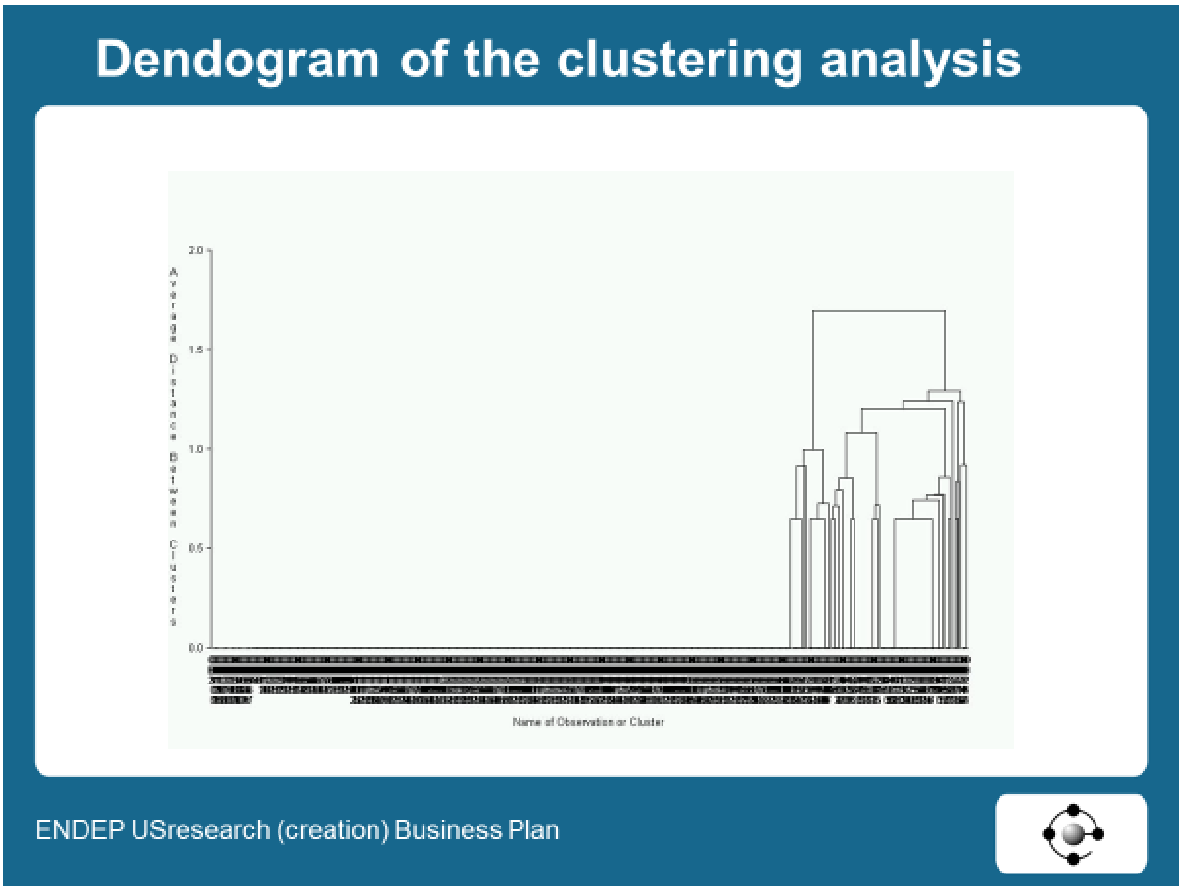
Stage 1: physicians using no IT at all (Cluster 1)

Stage 2: physicians using Electronic Medical Records (EMRs) (Cluster 2)

Stage 3: physicians using IT for EMR including some of this additional information: patient demographics information, computerized order for Rx, computerized order for tests, lab results (Cluster 3).

Graphic 1 provides an example of a dendrogram, showing the significant clusters with the Average Linkage method.

Graphic 1



Dendogram of the clustering analysis

ENDEP USresearch (creation) Business Plan

For a longitudinal clustering analysis of such medical data, the complete linkage criterion seems to be preferred or to outperform average linkage (AL) criterion for the interpretability of dendrograms in time-series data [10].

III.2. The Ward Clustering Method and Medical Record Analysis

Contrary to the previous linkage method, this clustering method is based on within group variance instead of linkage (1). Such an approach also seems quite suitable on the sampled physicians, in order to analyze some patterns of prescribing (e.g. number of medications). Similarity matrices on individual pairs based on computerization criteria is then performed, in a descending order. In statistics, the Ward method is also described as a general agglomerative hierarchical clustering procedure, where the criterion for choosing the pair of clusters to merge each step is based on the optimal value of an objective function. The dendrogram generated with the Ward method clearly showed also three clusters as with the AL method. In Further development, such algorithm may be used for optimization of multiple objectives.

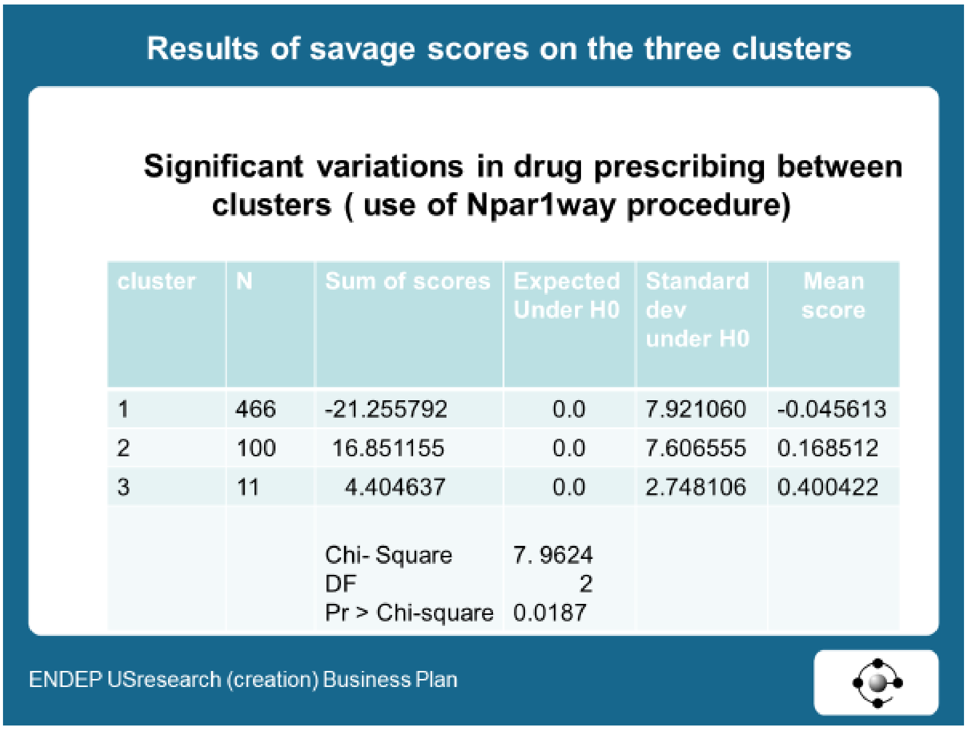
Clustering analysis could also be used to compare similarity matrices with Medications characteristics (e.g. old/new) versus matrices on individual pairs on computerization. It could be compared then to another study design, using two sets of criteria: on innovation in medication use versus innovativeness in communication technologies [11].

IV. Limitations and Further Research

Usually, several clustering algorithms are compared in order to validate the results; the comparison of similarity or dissimilarity matrices between the methods is then performed, if possible, on different datasets. Software vendors like SPSS have invested in such validation issues, especially on the minimum number of methods to validate results of a clustering analysis. Therefore, the dataset was re formatted for runs also with SPSS; the two clustering algorithms were used average linkage and Ward method. Similar number of clusters were identified, and a validation can be performed. For further development on time series medical data or other disease datasets, it may be necessary to explore further whether these two clustering methods are sufficient. As additional options were also available in the SPSS command, two types of distances were tested: the squared Euclidean distance and the Minkowski distance, on the average linkage method.

It can also be suggested to link subsequently the two methods showing significant results in a subsequent way. Starting first with the Average Linkage method, it helps to eliminate or control the effects of outliers; then using the second Ward method helps to control possibly better some forms of variations of prescribing patterns between groups of practice according to their levels or pace of computerization. It can be proposed to construct the two similarity matrices with the two different approaches first, and then to link the matrices subsequently (see two-step command in SPSS). Nonparametric results comparing physicians' prescription of old versus new medications confirmed at this stage that the clusters help to control the effect of ICT adoption and significant variations of drug utilization patterns by cluster. For instance, Table 2 shows the results of the Savage scores, comparing differences in number of medications between the three clusters (SAS nonparametric test).

Table 2



Results of savage scores on the three clusters

Significant variations in drug prescribing between clusters (use of Npar1way procedure)

cluster	N	Sum of scores	Expected Under H0	Standard dev under H0	Mean score
1	466	-21.255792	0.0	7.921060	-0.045613
2	100	16.851155	0.0	7.606555	0.168512
3	11	4.404637	0.0	2.748106	0.400422
		Chi- Square DF Pr > Chi-square	7. 9624 2 0.0187		

ENDEP USresearch (creation) Business Plan

V. Conclusions

One of CDC' objectives has been to provide to the health service research community, more data elements over the years; this will help to examine at what pace and how ICT is implemented in medical practices and what type of information is computerized. It also supports the progressive engagement of other providers of care and patients, with more information sharing. The move from a market economy to information exchanges in health care (medical insurance market), has created new needs for information systems. This paper is a methodological contribution to understand the transformation of the information system in medical groups; clustering algorithms applied on medical data may be very useful to understand modifications in information processing of health systems and impact on physicians' choices.

This application of an AL clustering analysis may be used in addition to various analysis of clinical and administrative tasks in high-tech medical groups, with advanced tracking systems using time and motion or video data [12]. Such analysis is usually at community or health system level, while the CDC survey is a nation-wide physician survey. It is timely to follow the increase of types of computerized information, in current multi-stakeholder engagements including patients (e.g. patients' preferences in development of clinical trials for drugs and devices, under the initiative of the Medical Devices Innovation Consortium, www.mdic.org). Clustering analysis can also be used for supervised learning and be useful as a category of classifiers for prototypical development [8, 13].

Acknowledgement: Prof Christine Huttin is grateful to S Atwood director of research computing, Partners Health Care (BWH) and to Countway medicine library, Harvard University, for their help in this study.

References:

[1]. Huttin CC Stated and Revealed Preference Issues: Product, patient and physicians' attributes in choice experiments, Technology and Health Care 25 (2017): 1005-1020.

[2]. Huttin CC and Atwood S. IT process in clinical practices for diabetes Type II. Value in Health, May 2011:14(3): A103.

[3]. Ward JH, Hierarchical grouping to optimize an objective function, Journal of the American Statistical Association. 1963; 58: 236-244.

[4]. Anderberg, MR, Cluster analysis for applications, New York: Academic Press, 1973.

[5]. Hartigan JA, Clustering Algorithms, New York: John Wiley & sons, 1975.

[6]. Bezdek JC and Pal SK Fuzzy models for pattern recognition, IEEE Press, New York. 1992 (Chapters 2&3).

[7]. Wong MA and Schaack C Using the nearest neighbor clustering procedure to determine the number of sub populations, American Statistical Association Proceedings of the Statistical Computing section, 1982: 40-48.

[8]. Maulik U, Bandyopadhyay S Integrating clustering and supervised learning for categorical data analysis, IEEE Transactions on systems, Man and Cybernetics- July 2010; Part A, 40 (4).

[9]. Tsumoto S, Hirano S A comparative study of clustering methods for long time-series medical databases, Springer-Verlag Berlin, Heidelberg. 2004.

[10]. Hirano S, Tsumoto S, empirical comparison of clustering methods for time series database. Active Mining, 2005; 3430 :268-286.

[11]. Huttin C Access to Drugs and the Impact of Financials on Clinical Decision Making. Academy Health Research Meeting Proceedings, Seattle, 2011.

[12]. Tai-Seale M, Olson CW, Albert SC, Morikawa C, Durbin M, Wang W, Luft HS. Electronic Health Records Logs indicate that physicians split time evenly between seeing patients and Desktop Medicine, Health Affairs, April 2017; 36 (4): 655-662.

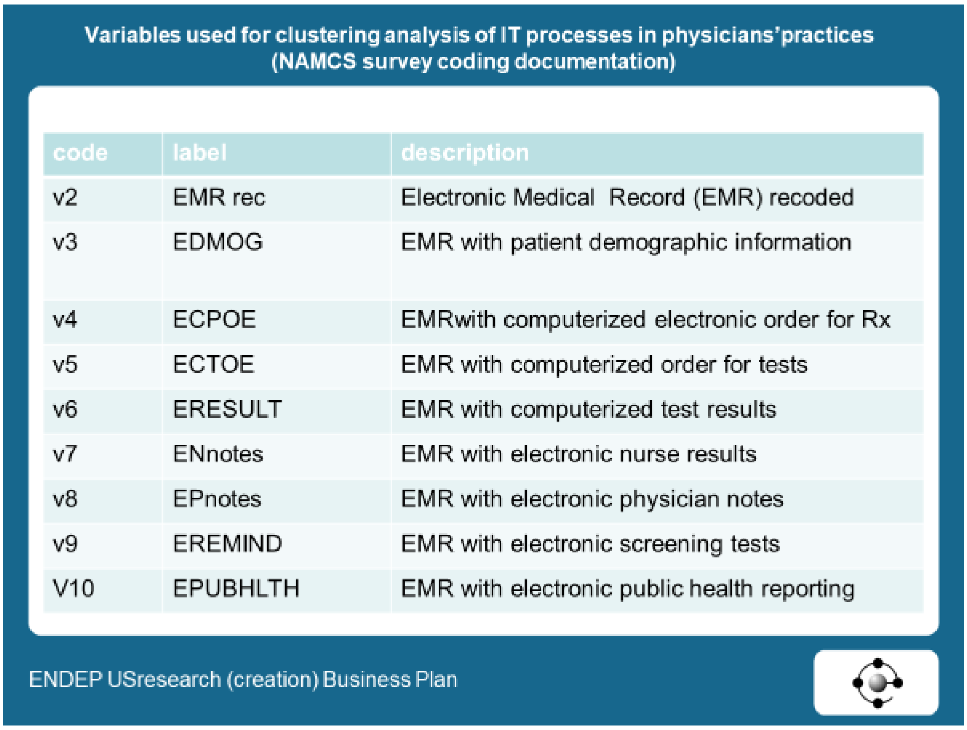
[13]. Bezdek JC, Kuncheva LI. Nearest Prototype Classifier Designs: An Experimental Study. International Journal of Intelligent Systems, 2001; 16: 1445-1473.

Related references for use of clustering algorithm for chemicals:

[14]. Massart DL and Kaufman L, The interpretation of analytical chemical data by the use of chemical analysis, New York, John Wiley & Sons, 1983.

Appendix 1: List of variables used for the clustering analysis

The following table provides the list of variables used in the clustering analysis with the coding documentation from CDC for the National Ambulatory Medical Care Survey (NAMCS)



Variables used for clustering analysis of IT processes in physicians' practices (NAMCS survey coding documentation)

code	label	description
v2	EMR rec	Electronic Medical Record (EMR) recoded
v3	EDMOG	EMR with patient demographic information
v4	ECPOE	EMRwith computerized electronic order for Rx
v5	ECTOE	EMR with computerized order for tests
v6	ERESULT	EMR with computerized test results
v7	ENnotes	EMR with electronic nurse results
v8	EPnotes	EMR with electronic physician notes
v9	EREMIND	EMR with electronic screening tests
V10	EPUBHLTH	EMR with electronic public health reporting

ENDEP USresearch (creation) Business Plan

The recoding of the variables used in the clustering analysis was performed in two steps. First, the main variable representing the computerization: Electronic Medical Report (Variable "EMEDREC") was recoded to be consistent with the codes of the other variables used for the analysis. Secondly, the series of the 10 selected variables used for the clustering was recoded in the same way. In the original CDC physician survey, the coding of the "EMEDREC" variable is the following:

1= Yes

2= Yes part of EMR is paper and part is electronic

3= No

V2 was first recoded in order to be consistent with the other selected variables for the clustering analysis (V3…., V10) and the "No" and "Yes" answers to the survey

The recoding of variables V2 to V10 was the following: 1= "Yes"

2 = "No" include value 2,3,4, 8 (not applicable)

Appendix 2: Issues of computer storage in medical practices

In order to implement clustering methods on medical claims, computing or storage for matrix elements may raise capacity issues, according to IT infrastructures. Clustering analysis was limited in the past because of limitation of computer storage especially for the linkages' methods such as the full linkage method. Computation may not be an issue any more with medical informatics infrastructures of hospitals and cloud services. However, the analysis of medical claims discussed in this paper, is for medical practices outside hospitals; the medical informatics infrastructure in small/medium sized practices which may depend in some states or regions, on refurbished or older computers may need to be addressed, for example in Societies such as HIMSS (according to various chapters).