Journal of
Computer Science & Systems Biology

Elsebakhi et al., J Comput Sci Syst Biol 2015, 8:6
DOI: 10.4172/jcsb.1000211

Research Article Open Access

Data Mining Risk Score Models for Big Biomedical and Healthcare Data

Emad Elsebakhi[1*], Ognian Asparouhov[2], Anton Berisha[2], Kris Latenski[2], Eric Schendel[1], Anwar Haque[1] and Rashid Al-Ali[1]

[1]*Sidra Medical and Research Center, Qatar Foundation, Doha, Qatar*
[2]*MEDai LexisNexis, Elsevier, Orlando, Florida, USA*

Abstract

Recently, big data is becoming the key to improve the future healthcare. The era of big biomedical data comes with significant challenges in querying, storage, visualize, and analyze the available petabytes of biomedical data, which makes healthcare industry a data-driven field. Currently, the available Concurrent Risk Model (CRM) is limited to the availability of patient episodes that are sensitivity to its cost. Herein, we propose a novel hierarchical data mining based on functional networks to develop a new CRM. This new risk score evaluates the last twelve-month period of patients' expected risk/cost/severity/illness burden/disease intervention using both medical and drugs claim-based predictors: diagnoses, medications (yes/no), and demographics. Our novel CRM predicts $50,000 per-member-per month (PMPM) tracks risk trends over time for any particular group, especially severe chronic diseases. Our CRM model has R^2=0.57 in comparison with the best results of Society of Actuaries.

Keywords: Big clinical data; Clustering; Large-scale machine learning; Data mining with data-driven; Functional networks; Episode treatment group; Electronic medical record; Clinical data

Introduction

The U.S. healthcare spending is 15.3% of its GDP and is projected to grow on an average of 6.7% annually over the period 2007-2017, thus reaching 19.5% of the U.S. GDP by 2017 [1-3]. The current healthcare and biomedical industry have greatly benefitted from the recent development in information technology and medical devices, since it has to process massive quantities of biomedical data on a daily basis.

Yet, the chronic diseases (diabetes, cancer and oncology, asthma, and cardiovascular, etc.) are the major causes of death of millions of individuals around the globe [4,5]. The digital electronic clinical/drug data are increasing exponentially and it is expected increase to reach to 25 Exabyte by 2020 [6,7]. In addition, due to the recent development in biomedicine and healthcare industry, we are currently facing great challenges in dealing with the generated or gathered massive biomedical data about patients through EMR systems, which have emerging demand to deal with it towards the essential need of biomedicine at bedside. However, we can only get the benefit of this available big biomedical data when we have adequate computational analytics tools to build future decision systems.

The current Episode Treatment Groups (ETG) has been widely utilized to understand and compare episodes of care across patients, hospitals, and populations. Moreover, ETGs need to evolve to reflect the increasing growth of care, especially, for chronic disease interventions and management. According to the OPTUM (2014) [8], improving care quality while reducing cost are essential to have reliable healthcare and accountable care act with evidence-based medical performance.

At present, performance measurement poses one of the most common challenges in the U.S. healthcare and biomedical industry, where a major challenge is determining the most important factors within EMR and ETG to assess the whole treatments provided to specific patient. These kinds of challenges motivate us to advance the current ETG mechanism and enhance it with the recent advance of big data and data mining predictive modeling paradigms. Therefore, it is mandatory to improve both quality and utilization of ETG that will drive to reliable and stable computations for improving the healthcare quality and minimize its expenditure.

The paper is designed in numerous of sections: Literature review about concurrent risk score is introduced in Section 2. Section 3 proposes the data acquisition and its properties with the repository style. The data mining concurrent risk score predictive models based on a novel functional networks are proposed in Section 4. The new CRM development with numerous of clusters is proposed in Section 5. Section 6 contains the results and interpretations with proper visualizations. Finally, the conclusions and future outlook is presented in Section 7.

Literature Review

The episode treatment group is a commercial software package that organizes claims into episodes of care by combining information from diagnosis, procedure codes, and drug *National Drug Code* (*NDC*) codes from patient claims [9-11]. The most essential factors in grouping ETG within any payment system depends on its diagnoses codes/procedures and the other clinical data that is used by grouping algorithms to establish the episodes and the attached cost. The chronic episode ETG is defined by its durations and it represents around 65% of the costs. It is known that the chronic conditions episodes (diabetes, cancer and oncology, cardiovascular, etc.) have no clear specific dates for its ETG processes, then episodes grouper affords powerful paradigm for hospitals pay for performance systems by assigning treatment costs for each treating health condition, then use benchmark different hospitals cost distribution and reward the most efficient [12-15].

The "current episodes are developed using integrations of diagnoses, procedures, and drugs, which are the patient's clinical services for a specific condition from the onset of symptoms until treatment is

***Corresponding author:** Emad Elsebakhi, Sidra Medical and Research Center, Qatar Foundation, P.O. Box 26999, Doha, Qatar, Tel: +00974-6698-3779; E-mail: eelsebakhy@sidra.org

Received October 21, 2015; **Accepted** November 12, 2015; **Published** November 16, 2015

Citation: Elsebakhi E, Asparouhov O, Berisha A, Latenski K, Schendel E, et al. (2015) Data Mining Risk Score Models for Big Biomedical and Healthcare Data. J Comput Sci Syst Biol 8:6 365-372. doi:10.4172/jcsb.1000211

complete" [16-19]. "One episode is related to only one patient, but patients can be in multiple episodes where grouping episodes is an essential step of the treatment processes and the risk score was built using associated average cost in ETG benchmark database that is frequency published based on the existing procedure and/or diagnosis codes" [16-20]. However, there are numerous limitations, namely: (i) the limitation of available ETGs that a patient can have (20 or less), (ii) the available ETGs are sensitive to their cost, and (iii) some patients have zero number of ETGs, hence they have zero risk score. Therefore, it is essential to propose a novel breakthrough data mining/machine learning risk score model that overcome these drawbacks and covers more predictors with reliable and dynamical performance (Figure 1).

The authors in ref. [7,13,21,22] have shown the summary statistics of the top spending episodes and the corresponding clinical services within the segment of time ending in 2006. In addition, they provide very useful graphs that explain how spending on the top five most expensive medications, shown in Figure 1. Recently, the identification of (i) clinical episodes and burdens of illness, (ii) services involved in knowing diagnosis for specific individual, and (iii) disease intervention, management, and treatment begin to be an essential key factor indicator for our daily-care use. The episodes of care provide a great tool for healthcare analytics and add motivations for high performance. The OptumInsight within the Optum healthcare system has announced that the use of ETG will advance the U.S. healthcare system and potentially improve the treatment quality, while reducing its corresponding cost [8].

The current ETG has been widely utilized to understand and compare episodes of care across patients, hospitals, and populations. At present, performance measurement poses one of the most common challenges in the U.S. healthcare and biomedical industry. These kinds of challenges motivate us to advance the current ETG mechanism and enhance it with the recent advance of big data and data mining predictive modeling paradigms. Therefore, it is essential to improve both quality and utilization of ETG that will drive to reliable and stable computations for improving the healthcare quality and minimize its expenditure. Moreover, the quality of hospitals depends on its resources, quality of care, and physician performance, which is calculated based on ETG risk adjusted based on big biomedical data and comparable episodes of treatment a specific individual [15]. Both quality and physician performance are identified based on assigning clinical data to ETG and then rank the episodes for abut more than 600

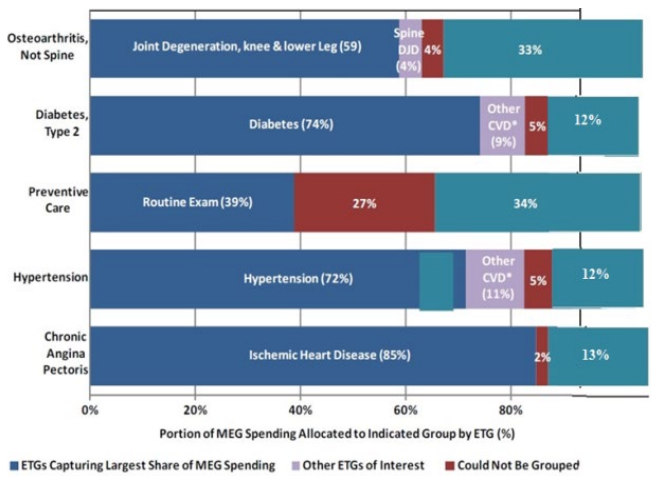

Figure 1: Spending on the five most expensive medications (Allocated by ETG, Allison et al. [3]).

categories health conditions. Therefore, by assigning costs to each ETG and compute the average risk for each patient. Finally, the episodes are assigned to specific hospitals and then risk scores are computed to identify hospitals or physicians ranking among their similar hospitals [12-15].

This research presents novel generalized data mining predictive models based on hierarchical iterative computational algorithms to identify the severe risk patients and predict the corresponding concurrent risk score using the available diagnosis, medications (yes/no), and demographics within EMR repository big data of numerous of American hospitals and health plans. Both the training algorithm and computational processes are carried within numerous clusters based on the organ systems, chronic diseases, age, gender, and enrollment period. The core of this new CRM is to predict the per-member-per-month (PMPM) truncated to $50,000 and the details of the development and deployment are shown below in the following sections. The comparison studies and results interpretations are carried out using SOA data and their comparative results with (i) in-sample and out-sample validation of the accuracy of both Calibrated and Non-Calibrated CRM and (ii) compute variety of statistical accuracy measures, namely, correlation, R-Squared, Sensitivity/Specificity, and Root Mean-Squared-Errors (RMSE). Although we know that a fair comparison requires the usage of one and same data set, the results for annual costs (truncated to $100,000 and $250,000 respectively) are also calculated and it has shown that our CRM hierarchical data mining predictive based on the generalized nonlinear functional networks models match the best result (DxCG DCG) and significantly outperforms the models of all other vendors.

Data Acquisition and Universal Repository

The CRM predicts the current year (last twelve months) individuals' severity risk without regard to any cost/utilization/procedures. We develop the CRM using medical and Rx claim-based big data: diagnoses, drugs and demographic variables. We used MEDai's (now LexisNexis Risk Solutions) big data repository of medical and pharmacy claims. Training set contains 3,809,349 lives (3,024,310 users and 785,039 non-users) with different line of business (LOB). The validation was done with many different big data sets - in total more than three million lives. The MEDai's data sets (from different clients/health plans), used in the repository for training and testing, contained claims incurred from July, 1, 2008 to June 31, 2009. In addition, we used data from the SOA 2007 comparative study [23] and compare our new CRM with the models of the US healthcare predictive modeling vendors participated in this study. This study used data from MedStat Marketscan, and the claims incurred from January 1st, 2003 through December 31st, 2003.

Novel Data Mining Concurrent Risk Score Predictive Models

Dealing with big biomedical data becomes challenges, especially within sparse and high dimensional data. Therefore, the attributes contributions in identifying a specific healthcare event will be meaningless. In addition, due to the sparse within input-space features, the predictive modeling computations will face a challenge problems in determining the importance and meaningful of the high volume of data within a specific cluster during the visualization and predictions. Furthermore, the expansion of the high-dimensional of the input biomedical feature space will lead to ill-condition feature-space matrix; which leads to over fitting problems. In this research, we hope to build a large-scale machine learning predictive modeling algorithm to deal

with such massive of big biomedical data. In this section, we briefly proposes a new large-scale data mining algorithm for predictive modeling platform in solving one of the most common healthcare problem, such as, adjusted risk for Episode treatment group and then be able to identify the severity of illness burden and disease intervention using both medical and drugs-based predictors: diagnoses, medications (yes/no), and demographics.

In the past few years, functional networks model became popular frameworks of predictive modeling in different real-life applications, such as, healthcare, software and petroleum engineering, business, etc. [24-28]. Based on the delivered outcomes, there is a great remarkable that functional networks can be considered a remarkable data mining knowledge discovery paradigm for predicting both continuous and categorical outcomes. The architecture of functional networks is a problem driven, which can be considered as a non-parametric predictive models that takes into considerations both expert domain and data-driven information during its training and deployment. This new machine learning paradigms are capable for large-scale big data classification. However, due to the existence of petabyte biomedical data within biomedicine and healthcare industries, efficient single predictive model will act poorly in predicting or classifying a specific healthcare outcome. Therefore, there is an essential need of advanced and efficient large-scale machine learning schemes to query, store, analyze, and build a suitable medical decision from the available mountains of biomedical data [29-37].

The motivation behind this research is to propose large-scale machine learning based on functional networks to build a suitable decision maker and overcome the most common challenges within biomedical-big and fulfill the essential needs of personalized medicine at bedside. To make it simple to the reader, we use the proposed functional network example from ref. [35-37], Figure 2. It provides a functional networks model for data $D=\{X_1 \ldots X_5, \text{ and } y\}$ of five input attributes and one continuous target, The corresponding architecture was presented in three steps: (i) The architecture constitutes of numerous of layers to keep the neurons: One input layer that contains the features (input variables); output layer for the specific output(s) and (one or more) intermediate layer(s) that store(s) intermediate information ; for instance, x_1 and x_2; (ii) many layers for intermediate units; and finally, (iii) A set of links between neurons and layers.

The training algorithm of the functional networks predictive modeling is based on a given training set $D=\{(y_i, x_{i1},\ldots,x_{ip})\}$, for $i=1,\ldots,n$; where $x_i=(y_i, y_{i1},\ldots,x_{ip})$; the goal is to estimate y_i using the available data, that is, $y_i = F(x_{ij},\Theta)+\varepsilon_i;\ \forall i=1,\ldots,n;$ from which it follows

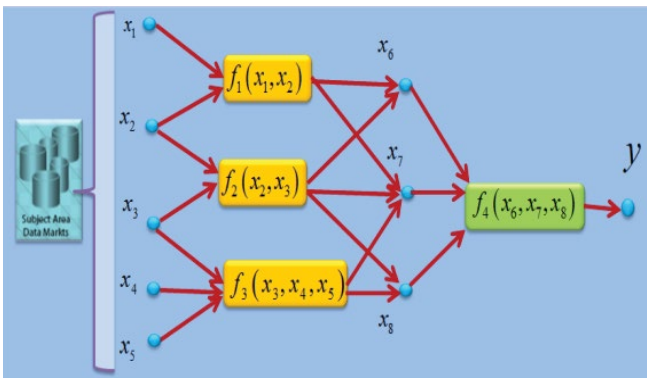

Figure 2: Functional networks for five attributes.

that $-\infty < y_i < \infty$; for $i=1,\ldots,n$. The functions, $F(\mathbf{x}_{i.},\boldsymbol{\Theta})$ could be linear or non-linear according to the domain of the problem-in-hand. The goal is to find $\hat{F}(x_{ij},\hat{\Theta})$ and $\hat{\Theta}$ using one of the common optimization techniques, such as, ***Least Squares***: $Q_{LS}=\sum_{i=1}^{n}\left[y_i-\hat{F}(\mathbf{x}_{i.},\hat{\boldsymbol{\Theta}})\right]^2$; or any other optimization formulae see the same procedures as in ref. [25]. Assume that the neuron functions of this predictive model is written as a combination of sets of independent families (bases), $_{m_{rj}}l=1,\ldots,m_{rj}$ for $l=1,\ldots,m_{rj}$; $r=1,\ldots,2^p-1$,; $j=1,\ldots,p$; that is, $\hat{f}_{rj}(x_{ij})=\sum_{l=1}^{m_{rj}}a_{rjl}\psi_{rjl}(x_{ij})$. Therefore, $\{a_{rjl}\}$ is estimated using mix of popular families: $\{1,X,\ldots,X^m\}$ or $\{1,e^{X},e^{-X},\ldots,e^{mX},e^{-mX}\}$;which represents the activation functions in any machine learning and data mining predictive modeling algorithms. If we write $\hat{f}_{11}(x_{i1})$ as a combination of $\psi_{11l}(x_{i1})=\{1,x_{i1},\ldots,x_{i1}^m\}$; where $\hat{f}_{rj}(x_{ij})$ is written as a combination of other set, but no constant term, $F(x_{ij},\Theta)=\mathbf{W}_{ij}\boldsymbol{\theta}_j$. Therefore, $\hat{F}(x_{ij},\hat{\Theta})=\mathbf{W}_{ij}\hat{\boldsymbol{\theta}}_j$; $\hat{\boldsymbol{\Theta}}=\mathbf{A}^{-1}\mathbf{B}$, where $\mathbf{A}=W^TW$; $\mathbf{B}=W^TY$; then $\hat{y}=\sum_{r=1}^{2^p-1}\prod_{j\in\{1,2,..,p\}}\left[\sum_{l=1}^{m_{rj}}\hat{a}_{rjl}\psi_{rjl}\left(x_{ij}\right)\right]$; for the square matrix, **w**; where **w** is the extended coefficient matrix **w** see ref. [29-37] for more details.

The New CRM - Methodology

The main goal of this research is to present novel generalized data mining predictive models based on hierarchical iterative computational algorithms to evaluate/predict the corresponding concurrent risk score using the available diagnosis, drugs (Yes/No), and demographics within EMR big data repository (medical and Rx claims of different US hospitals and health-plans). The dependent variable (target) is current year cost *per-member-per-month (PMPM)* truncated to $50,000.

The patient's risk score is calculated as the ratio of individual's expected (risk adjusted) cost over average expected cost of last December or prior preferred base period for the specific line of business (LOB). This makes it possible to track the risk trend over time for any particular groups, especially, those with sever conditions and chronic diseases. The non-users Concurrent Risk Score is set to zero although some clients (health plans) prefer a base population risk to be assigned to these members according to their demographics. The desire predictive modeling process consists of the following major steps (Tables 1-6):

- **Step 1:** Split the universal repository data into clusters (see Table 1), according to presence or absence of drugs, member's severity (BodySystem), gender and months of enrollment. BodySystem (integer) represents the number of organ systems where the patient has at least one diagnosis code. The organ system split of the codes follows closely the International Statistical Classification of Diseases, [38], ICD 9 Diagnoses hierarchy. For example if patient has one or more respiratory system diagnosis and one or more diagnosis then BodySystem=2.
- **Step 2:** Build universal (non-calibrated) concurrent risk score model with dependent variable=PMPM (truncated to $50,000) for each cluster using the universal repository data;
- **Step 3:** Customize (calibrate) the universal (non-calibrated) CRM for each client (health plan and LOB), using only this client data.

#	Cluster Definition	% of the total members	Average PMPM ($50K truncation)
1	Drugs=NO; BodySystem<2	8.89	$25.40
2	Drugs=NO; 2 ≤ BodySystem<3	6.13	$57.75
3	Drugs=NO; 3 ≤ BodySystem<5	8.34	$120.95
4	Drugs=NO; 5 ≤ BodySystem	7.03	$498.82
5	Drugs=YES; BodySystem<1	3.60	$29.28
6	Drugs=YES; 1 ≤ BodySystem<2	6.40	$66.09
7	Drugs=YES; 2 ≤ BodySystem<3	9.08	$108.57
8	Drugs=YES; 3 ≤ BodySystem<4	10.40	$162.46
9	Drugs=YES; 4 ≤ BodySystem<5	10.18	$232.63
10	Drugs=YES; 5 ≤ BodySystem<7; Male	6.61	$400.93
11	Drugs=YES; 5 ≤ BodySystem<7; Female	9.35	$366.29
12	Drugs=YES; 7 ≤ BodySystem; Male	4.79	$1,232.69
13	Drugs=YES; 7 ≤ BodySystem; Female	9.20	$960.16

Table 1: Repository data's distribution according to 13 Clusters for the 3,024,310 users.

LOB	Count	Correlation	R^2	RMSE	Top 10% Sens./ Spec.	Average Predicted$	Average Actual$
Commercial HMO	794,687	0.73	0.54	929.4	57.94/95.33	373	373
Commercial Non HMO	2,165,285	0.75	0.56	804.3	61.26/95.70	336	336
Medicare	219,903	0.75	0.56	1365.1	68.48/96.50	366	366
Medicaid	554,005	0.63	0.40	1031.9	54.19/94.91	335	335
State Program	75,469	0.67	0.46	653.4	50.28/94.48	258	257
Total	3,809,349	0.73	0.53	913.96	60.53/95.61	379	379

Table 2: Universal/Non-calibrated CRM in-sample performance for different LOB (PMPM Truncation: $50K).

Measure of goodness	PMPM $50K truncation	Annual Cost $100K truncation	Annual Cost ($250K truncation)
Correlation	0.74	0.77	0.76
R^2	0.54	0.59	0.57
Top 10% Sensitivity	64.67	64.80	64.70
Top 10% Specificity	96.07	96.50	96.25
Average Predicted$	$273	$3,211.8	$3,265.4
Average Actual$	$276	$3,148.8	$3,268.5
RMSE	648.6	5,278.1	6,632.9

Table 3: Calibrated CRM out-of-sample performance for 1,631,088 commercial members.

Data set/ LOB	Member Count	R^2			Top 10% Sensitivity/ Specificity (the same for all three models)	RMSE		
		PMPM ($50K)	Annual$ ($100K)	Annual$ ($250K)		PMPM ($50K)	Annual$ ($100K)	Annual$ ($250K)
Commercial HMO	392,954	0.53	0.50	0.52	61.22/95.69	830.1	6,070.3	8,005.8
Commercial Non HMO	1,780,079	0.51	0.52	0.53	60.39/95.60	926.4	6,217.3	8,797.1
Medicaid	239,782	0.41	0.42	0.42	59.46/95.50	959.2	6,608.5	9,075.3
Total	2,412,815	0.51	0.52	0.53	62.19/95.80	812.5	5743.7	7729.2

Table 4: Universal\Non-Calibrated CRM out-of-sample performance for different LOBs and overall.

Measure of goodness	PMPM $50K truncation	Annual Cost $100K truncation	Annual Cost ($250K truncation)
Correlation	0.7053	0.7423	0.7188
R^2	0.4965	0.544	0.515
Top 10% Sensitivity	63.09	63.09	63.09
Top 10% Specificity	95.90	95.90	95.90
Average Predicted$	273	3251.2	3274.31
Average Actual$	298	3482.09	3563.59
RMSE	596.34	5462.36	6569.87

Table 5: CRM performance on SOA 2007 Data (308,210 commercial members).

The chosen clustering methodology is based on three variables, which is simple but meaningful according to the clinical domain or case manager expert and financial standpoint of view, because:

- BodySystem is a good proxy of members' severity/risk/illness burden;

- Drugs (Yes/No) reduce significantly the predictors' count and

make the modeling process easier. Gender also contributes in this direction;

- Average PMPM cost per cluster significantly different for different clusters.

Results and Discussion

We calculated a variety of statistical measures of model's accuracy and goodness: Correlation, R-Squared, Sensitivity and Specificity, Average Actual and Average Predicted Cost and Root Mean-Squared-Errors (RMSE). We compared the Actual Cost with the Predicted Cost both truncated to $50,000 in Table 2. There are two approaches for model's performance estimation according to the data set used for accuracy's evaluation: in-sample or optimistic estimator where the training data set is used also for model performance evaluation, and out-of-sample or pessimistic estimator where model performance is based on a new data set that did not participate in the training process. Out-Sample results: Similarly as above, the comparative experimental study was done out-of-sample (results are reported in Tables 3 and 4, respectively.

The SOA comparative studies: We applied the new CRM to the data set used in the most recent SOA Comparative Study (Winkelman and Syed) - Tables 5 and 6 [23]. The SOA 2007 data set was split into two subsets: the first data set was used as training for the prospective model but was not used for the CRM (results are reported in Table 5). The second set was used as testing for both the prospective model and CRM and the results are given in Table 6. Our novel CRM matches the best result (DxCG DCG) and significantly outperforms the concurrent models of all other vendors.

We present two different graphs for the developed concurrent risk score with both age and the monthly expenditure of each individual in Figures 3 and 4, respectively. These two Figures represent the impact and benefit of the new developed CRM and its essential need to identify the burden of illness and disease intervention plan.

Figure 3 shows that the developed risk score versus monthly expenditure in ($US) for 1,028,361 individuals with 766,878 (74.6%) users: It shows that the higher individuals monthly expenditure, the higher concurrent risk score. Moreover, Figure 4 shows that the developed risk score versus age for the same 766,878 individuals. The relationship shows that higher the age, the higher the risk score, which is expected/convenience with the physicians or case manager opinion in the real-life. Our results for R^2 vary from 0.5 to 0.57 which is compared to the best result of the 2007 SOA comparative study. Although we know that a fair comparison requires the usage of one and same data set. Although, the *DxCG* uses more than 500 clusters with a well-prepared model, the proposed novel predictive CRM is similar or better than it. In addition, our new CRM outperforms the rest of all other predictive models.

We observed in Figure 3 that the predicted concurrent risk score versus monthly expenditure, there are two isolated dot results: (i) the red triangle when risk factor is close to 1.1, 1.5, and beyond; (ii) the blue circle when the risk factor is close to 0.7. The abnormal predicted values can be interpreted based on looking at the actual data and the definition of the risk score (predict $50,000 truncated per-member-per-month and scales patient risk score using risk-adjusted cost of benchmark year from January 1st to December 31st. In addition, it tracks risk trends over time for any particular group, especially those with severe chronic diseases. Therefore, who are below average risk factor and having multi-conditions of severe chronic diseases will have such kind of abnormal behavior; for example, the low-cost spending (individuals with 0.7 risk score) are healthy with $200 spending per-month and no identification of the burden of illness. On the other hand, the individuals with high cost spending because of their burden of multi-factor of chronic disease, COPD, cancer, diabetes, and asthma, then they behave improperly with the rest of the group, which convince us to customize our predictive model to be able to catch such critical cases to fit with our models and get the proper immune therapy plan.

In Figure 4, we observed that there are few points close to 100 years old, while the risk factors for those points are 0. By looking at the actual data based on both type of medications and health-conditions, we find that these patients are healthier than the rest of the group. In addition, they are taking care of themselves through coaching and proper immune-therapy plan, then there risk factor almost close to zero.

Conclusions

We conclude that the proposed novel generalized data mining predictive models based on hierarchical iterative functional networks computational algorithms was able to efficiently identify the severe risk patients and predict the corresponding concurrent risk score using the available diagnosis, drugs (yes/no), and demographics within EMR

R-squared and MAPE offered Concurrent Nonlagged by Claims Truncation Level							
Risk Adjuster Tool	**Inputs**	**R-Squared**			**MAPE%**		
		100K	**250K**	**None**	**100K**	**250K**	**None**
ACG	Diag	29.4%	29.7%	27.4%	73.0%	75.0%	75.4%
CDPS	Diag	35.5%	32.9%	31.0%	79.0%	80.6%	81.0%
Clinical Risk Groups	Diag	47.1%	43.3%	39.9%	68.6%	70.5%	70.9%
DxCG DCG	Diag	57.2%	51.8%	49.8%	61.6%	65.0%	65.4%
DxCG RxGroups*	Rx	N/A	N/A	N/A	N/A	N/A	N/A
Ingenix PRG*	Rx	N/A	N/A	N/A	N/A	N/A	N/A
MedicaidRx	Rx	32.1%	28.1%	24.6%	77.2%	79.1%	79.6%
Impact Pro*	Med+Rx+Use	N/A	N/A	N/A	N/A	N/A	N/A
Ingenix ERG	Med+Rx	46.5%	42.4%	38.6%	65.8%	67.7%	68.2%
ACG w/ Prior Cost**	Diag+$Rx	N/A	N/A	N/A	N/A	N/A	N/A
DxCG UW Model**	Diag+$Total	N/A	N/A	N/A	N/A	N/A	N/A

New Model	**Inputs**	**R-Squared**			**MAPE %**		
		$100K	$250K	PMPM ($50K)	$100K	$250K	PMPM ($50K)
	Diag. + R_x	54.4%	51.5%	49.83%	67.9%	62.3%	64.7%

Table 6: The SOA comparative studies for concurrent risk score model.

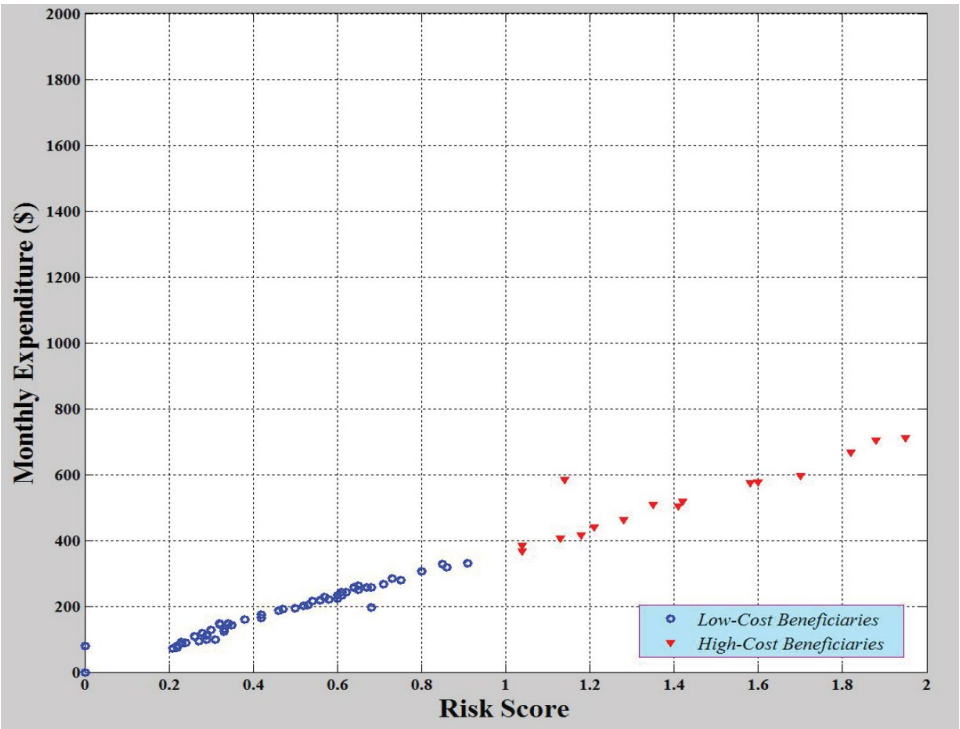

Figure 3: The predicted concurrent risk score versus Monthly expenditure in $US.

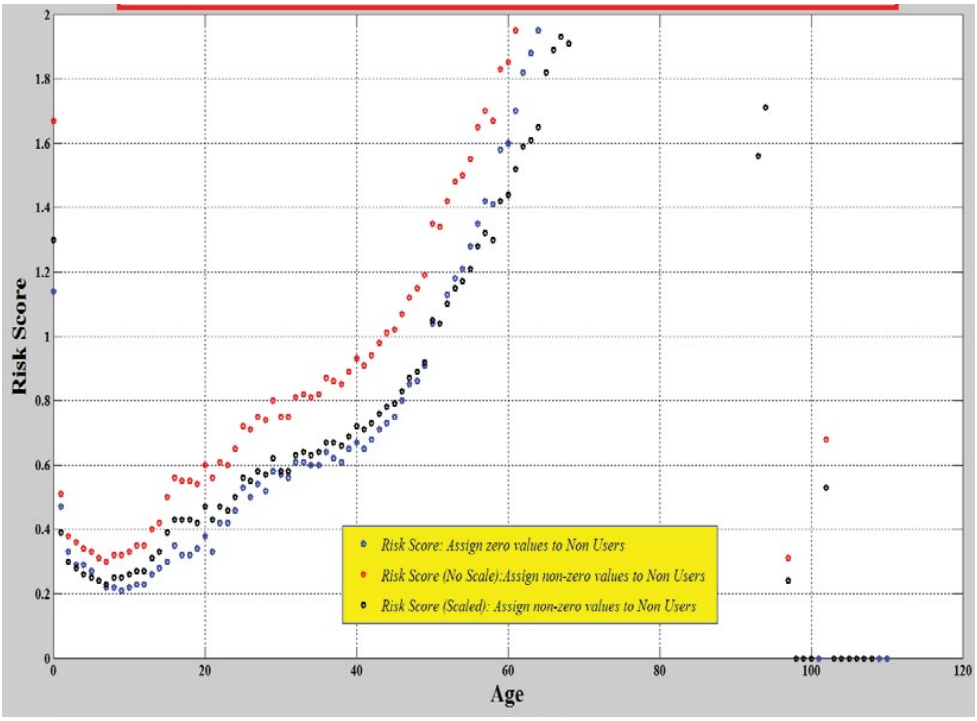

Figure 4: The predicted concurrent risk score versus age.

repository big data with reliable and stable performance. Thus, we can track the risk trend over time for any particular groups, especially, those with sever conditions and chronic diseases. The development and results of our finding can be summarized as follows:

- We created a repository of 3.809,349 million individuals (3.024, 310 million users and 785,039 non-users), described with 362 measures (diagnoses and drugs only).

- We developed a universal (non-calibrated) Concurrent Risk Model with thirteen clusters for commercial LOB and twelve clusters for all other line of business. These clusters were defined for active members with full (12 months) enrollment by the following measures: presence/absence of drugs, member's severity body system (BodySystem) and gender. After that, the final result was adjusted for different number of months enrolled.
- We developed a customized (non-calibrated) Concurrent Risk Model for each health plan.
- The value of R^2 for the out-of-sample validation of the new developed CRMs (calibrated and non-calibrated) vary from 0.5 to 0.57 (with one exception of 0.41 for Medicaid population). The comparison with other US healthcare outcome predictive modeling vendors based on the SOA 2007 study data shows that our CRM matches the best result (DxCG DCG) and significantly outperforms the concurrent models of all other vendors.
- Future research should focus on development of new concurrent risk score models for hospitalization (LOS) and re-admission, genomic wide association studies, individual response to specific drug, emergency visits; and physician performance, or any other healthcare events and biomedical informatics research targets.

Acknowledgements

The authors acknowledge the support of Sidra Medical and Research Center, Doha, Qatar for the facilities utilized to perform the presented work. They are also grateful to MEDai Inc. (now LexisNexis), Orlando, Florida, USA for their support and implementations. This research was funded by MEDai, Inc. and carried out with MEDai's data repository.

References

1. Roski J, Bo-Linn GW, Andrews TA (2014) Creating value in health care through big data: opportunities and policy implications. Health Aff (Millwood) 33: 1115-1122.
2. Cleveland Clinic (2012) Cleveland Clinic unveils top 10 medical innovations for 2012. Cleveland (OH).
3. Rosen A, Liebman E, Aizcorbe A, Cutler D (2012) Comparing Commercial Systems for Characterizing Episodes of Care. Bureau of Economic Analysis.
4. Signal and Forecasts Map: Mapping the Landscape of Challenges and Responses: Looking ahead over the next decade, the future of health and healthcare seems more uncertain than ever before. The possibilities for change are endless (2009) Institute for the Future.
5. Aizcorbe A, Nestoriak N (2011) Changing mix of medical care services: stylized facts and implications for price indexes. J Health Econ 30: 568-574.
6. Alan Weil (2014) Using Big Data to Transform Care. Health Affairs Briefing.
7. Rosen AB, Liebman E, Aizcorbe A, Cutler DM (2012) Comparing Commercial Systems for Characterizing Episodes of Care. Department of Quantitative Health Sciences, University of Massachusetts Medical School.
8. OPTUM (2014) Optum healthcare system to improve care delivery, quality and cost-effectiveness: Episode Treatment Groups.
9. Davis K (2007) Paying for care episodes and care coordination. N Engl J Med 356: 1166-1168.
10. Hackbarth G, Reischauer R, Mutti A (2008) Collective Accountability for Medical Care - Toward Bundled Medicare Payments. New England Journal of Medicine 359: 3-5.
11. Welch WP, Yang W, Flynn J (2010) Prescription Antihypertensive Drug Use in Children. Academy Health, Annual Research Meeting, June 27-29, Boston, USA.
12. Dudley RA, Rosenthal MB (2006) Pay for Performance: A Decision Guide for Purchasers. Rockville, MD: Agency for Healthcare Research and Quality. AHRQ Pub. No. 06-0047.
13. Sandy LG, Rattray MC, Thomas JW (2008) Episode-based physician profiling: a guide to the perplexing. J Gen Intern Med 23: 1521-1524.
14. MaCurdy T, Kerwin J, Gibbs J, Lin E, Cotterman C, et al. (2008) Evaluating the Functionality of the Symmetry ETG and Medstat MEG Software in Forming Episodes of Care Using Medicare Data. Acumen, LLC.
15. MaCurdy T, Shafrin T, Hartmann E (2010) Challenges in the Risk Adjustment of Episode Costs: Acumen LLC.
16. Harriet LK, Feder J, Ginsburg PB (2011) Bundling Payment for Episodes of Hospital Care. Issues and Recommendations for the New Pilot Program in Medicare.
17. Pham H, Ginsburg P, Lake T, Maxfield M (2011) Episode-Based Payments: Charting a Course for Health Care Payment Reform.
18. Rattray MC (2011) Measuring Healthcare Resources Using Episodes of Care. 2008; Accessed September 20, 2011.
19. Rosenthal MB (2008) Beyond pay for performance-emerging models of provider-payment reform. N Engl J Med 359: 1197-1200.
20. BCBSMA (Blue Cross and Blue Shield Association, Massachusetts), (2012), Boston.
21. Sood N, Huckfeldt PJ, Escarce JJ, Grabowski DC, Newhouse JP (2011) Medicare's bundled payment pilot for acute and postacute care: analysis and recommendations on where to begin. Health Aff (Millwood) 30: 1708-1717.
22. Thomas F, Caplan C, Levy JM, Cohen M, Leonard J, et al. (2010) Clinician feedback on using episode groupers with Medicare claims data. Health Care Financ Rev 31: 51-61.
23. Winkelman R, Syed M (2007) A Comparative Analysis of Claims-based Tools for Health Risk Assessment, Society of Actuaries.
24. Castillo E (1998) Functional networks. Neural Process. Letters 7: 151-159.
25. Castillo E, Cobo A, Gómez-Nesterkin R, Hadi AS (1999) A general framework for functional networks. Networks 35: 70-82.
26. El-Sebakhy EA (2004) Functional Networks Training Algorithm for Statistical Pattern Recognition. The 9th IEEE Symposium on Computers and Communications 1: 92-97.
27. El-Sebakhy EA (2009) Data mining in forecasting PVT correlations of crude oil systems based on Type 1 fuzzy logic inference systems. Journal of Computers and Geosciences 35: 1817-1826.
28. El-Sebakhy EA (2009) Software reliability identification using functional networks: A comparative study. Journal of Expert Systems with Applications 36: 4013-4020.
29. El-Sebakhy EA, Faisal K, El-Bassuny T, Azzedin F, Al-Suhaim A (2006) Evaluation of Breast Cancer Tumor Classification with Unconstrained Functional Networks Classifier. The fourth ACS/IEEE International Conference on Computer Systems and Applications 281-287.
30. El-Sebakhy EA, Hadi AS, Faisal KA (2007) Iterative least squares functional networks classifier. IEEE Trans Neural Netw 18: 844-850.
31. El-Sebakhy EA, Asparouhov O, Abdulraheem A, Wu D, Latinski K, et al. (2010) Data mining in identifying carbonate litho-facies from well logs based from extreme learning and support vector machines. In: Proceeding of AAPG. GEO 2010 Middle East Geoscience Conference Bahrain.
32. El-Sebakhy EA, Asparouhov O, Abdulraheem A, Wu D, Latinski K, et al. (2010) On Utilizing Functional Networks Computational Intelligence in Forecasting Rock Mechanical Parameters for Hydrocarbon Reservoirs: Methodology and Comparative Studies. Innovative Geoscience Solutions, Bahrain.
33. El-Sebakhy EA (2010) Flow regimes identification and liquid-holdup prediction in horizontal multiphase flow based on neuro-fuzzy inference systems. Math Comput Simul 80: 1854-1866.
34. El-Sebakhy EA (2011) Functional networks as a novel data mining paradigm in forecasting software development efforts. Expert Syst Appl. 8: 2187-2194.
35. El-Sebakhy EA, Asparouhov O, Abdulraheem A, Majed AA, Wu D, et al. (2012) Functional networks as a new data mining predictive paradigm to predict permeability in a carbonate reservoir. Expert Syst Appl 39: 10359-10375.
36. El-Sebakhy EA, Asparouhov O, Al-Ali R (2015) Novel Incremental Ranking Framework for Biomedical Data Analytics and Dimensionality Reduction: Big Data Challenges and Opportunities. J Comput Sci Syst Biol 8: 203-214.

37. El-Sebakhy EA, Lee F, Schendel E, Haque A, Kathireason N, et al. (2015) Large-scale machine learning based on functional networks for biomedical big data with high performance computing platforms. Journal of Computational Science 11: 69-81.

38. WHO (1992) International Statistical Classification of Diseases and Related Health Problems. Tenth Revision. Vol. 1: Tabular list, 1992. Vol. 2: Instruction Manual, 1993. Vol. 3: Index. Geneva.