



Contents lists available at ScienceDirect

Big Data Research

www.elsevier.com/locate/bdr



Revealing Physicians Referrals from Health Insurance Claims Data[☆]

Vagner Figueredo de Santana^a, Ana Paula Appel^a, Luis Gregorio Moyano^b, Marcia Ito^a,
Claudio Santos Pinhanez^a

^a IBM Research, Brazil

^b FCEN/CONICET, Mendoza, Argentina

ARTICLE INFO

Article history:

Received 13 October 2017

Received in revised form 23 January 2018

Accepted 7 March 2018

Available online xxxx

ABSTRACT

Health insurance companies in Brazil have their data about claims organized having the view only for service providers. In this way, they lose the view of physicians' activity and how physicians share patients. Partnership between physicians can be seen as fruitful, when they team up to help a patient, but could represent an issue as well, when a recommendation to visit another physician occurs only because they work in same clinic. This work took place during a short-term project involving a partnership between our lab and a large health insurance company in Brazil. The goal of the project was to provide insights (with business impact) about physicians' activity from the analysis of the claims database. This work presents one of the outcomes of the project, i.e., a way of modeling the underlying referrals in the social network of physicians resulting from health insurance claims data. The approach considers the flow of patients through the physician-physician network, highlighting connections where referrals between physicians potentially occurred. We present the results from the analysis of a claims database (detailing 18 months of activity) from the health insurance company we partnered with. The main contribution presented in this paper is the model to reveal *mutual referrals* between physicians. Results show the proposed model reveals underlying characteristics of physicians' activity from real health insurance claims data with multiple business applications.

© 2018 Published by Elsevier Inc.

1. Introduction

Health insurance costs are a main issue of concern in almost every country in the world as budget constraints impact directly on the quality of the service. As a result, health insurance companies have been extensively trying to reach a trade-off between offered services and costs as a way to meet budget constraints.

One way for health insurance companies to address those issues is to better understand the complex relationships among the diverse participants of the healthcare systems, including patients, physicians, hospitals, and other service providers. To support this quest, healthcare insurance companies and other health service providers have often a wealth of data from their own operations at their disposal, especially transactional data.

In the case of health insurance companies, an important piece of transactional data involves the claims presented by their ecosystem of providers. In the present work, a claim represents a report from a physician or a healthcare service provider to a health insurance company requesting some form of fee related to a patient's

consultation with a physician, a clinical exam, or a medical procedure. Even though claims data may vary, it generally contains at least the ID of the healthcare professional involved in the procedure (it may also be a group of professionals), the ID of the patient, the type of procedure, and time information related to the event. It may include other types of information such as location of the service, pre-authorization codes, etc.

Traditionally the analysis of claims data is based on applying statistics and Data Mining methods to the individual elements of the system (physicians, service providers, patients) or to the set of claims. However, healthcare is often provided by collaborative teams of physicians, nurses, and technicians which are connected to each other by often strong professional relationships. Physicians that refer patients to other physicians have clear preferences about who they want to team up with for specific procedures and often are involved in master-apprentice structures. Physicians also have preferences for specific service providers such as hospitals and clinical analysis laboratories. Those recommendations could be good for building patient trust or indicate a fraud when this is not the patient's will. Similarly, patients establish bonds of trust and reliance with specific physicians or group of physicians.

In practice, mining claims is difficult because claims are paid to a wide variety of providers, such as hospitals, clinics, or even

[☆] This article belongs to Special Issue: Medical Data Analytics.

E-mail address: santana.vagner@gmail.com (V. Figueredo de Santana).

1 physicians registered as small companies. A single patient can con- 67
2 sult a physician through all those channels. It could be even diffi- 68
3 cult to know who exactly is taking care of a given patient, since 69
4 that there are cases in which the physician ID used in a claim is 70
5 from a professional registered in the health care provider system, 71
6 but the one taking care of the patient is an unregistered physician. 72
7 Moreover, claims contain no information about referrals and of- 73
8 ten the connections among service provider team members are not 74
9 recorded explicitly in claims. It is also well known that claims data 75
10 is riddled with errors and unreliable information. Despite all those 76
11 difficulties, we show in this paper that meaningful and reliable in- 77
12 sights about the flow of patients through the network of physicians 78
13 and how physicians refer each other can be inferred from claims 79
14 data. 80

15 This study took place during a partnership, in a short-term 81
16 project, between our lab and a major Brazilian health insurance 82
17 company involving the analysis of their claims database. The main 83
18 challenge the insurance company brought to our team was to iden- 84
19 tify physicians that excel at medicine by using the claims database. 85
20 After decomposing this challenge considering the health insurance 86
21 workflow, the following components were identified as key factors 87
22 for outstanding professionals, defined by the health insurance com- 88
23 pany: physicians referred by peers, relative importance in the net- 89
24 work of physicians, and returning behavior of patients. The need 90
25 for modeling referrals emerged during interactions with the health 91
26 insurance company staff as a way of identifying physicians that 92
27 excel in a certain specialty and are referred by their peers. These 93
28 interactions occurred weekly, fomenting discussions between our 94
29 team and subject matter experts, IT specialists, analytics team, and 95
30 the superintendent of the health insurance company. Hence, this 96
31 work aims at presenting a way of modeling mutual referrals in 97
32 a physician–physician network, which connects to the hypothesis 98
33 studied in this work: *H1) It is possible to identify underlying physi- 99
34 cians' referrals from claims data.* 100

35 The main contribution of this work is a way of modeling mutual 101
36 referral patterns in the physician–physician network. The method 102
37 can therefore be used by health insurance companies to bet- 103
38 ter manage the physicians they have businesses with, nurturing 104
39 the experience of registered physicians and inviting unregistered 105
40 physicians that collaborate with registered ones. It can also be used 106
41 to support patients to receive more integrated care from a group 107
42 of physicians and service providers. 108

43 This paper is organized as follows: section 2 describes the 109
44 related work, section 3 details the database analyzed, section 4 110
45 presents the proposed model for highlighting mutual referrals, sec- 111
46 tion 5 discusses the obtained results, and section 6 concludes and 112
47 points to future works. 113

2. Related work

48 Healthcare data is heralded as the key element in the quest 114
49 to improve efficiency and reduce costs in healthcare systems [1]. 115
50 This trend is becoming more pronounced as multi-scale data gen- 116
51 erated from individuals is continuously increasing, particularly due 117
52 to new high-throughput sequencing platforms, real-time imaging, 118
53 and point-of-care devices, as well as wearable computing and mo- 119
54 bile health technologies [2]. 120

55 In healthcare, data heterogeneity and variety arise as a result 121
56 of linking the diverse range of biomedical data sources available. 122
57 Sources of quantitative data (e.g., sensor data, images, gene arrays, 123
58 laboratory tests) and qualitative data (e.g., diagnostics, free text, 124
59 demographics) usually include both structured and unstructured 125
60 data, normally under the name of Electronic Health Records. Addi- 126
61 tionally, the possibility to process large volumes of both structured 127
62 and unstructured medical data allows for large-scale longitudinal 128
63 129
64 130
65 131
66 132

studies, useful to capture trends and to propose predictive mod- 67
els [3]. 68

One of the most useful and commonly used datasets are claims 69
databases. Claims data records are often rich in details, as they 70
describe important elements of the events taking place around 71
the healthcare professional and the patient, e.g., timestamps, ge- 72
ographical location, diagnosis codes, associated expenses, among 73
others. The use of claims data in healthcare studies has been scru- 74
tinized in [4] and [5], providing a set of good practices and outlin- 75
ing the shortcomings of claims-based research. 76

Social network analysis has proven to be a useful analysis tool 77
in this context, allowing for insights difficult to reach by traditional 78
descriptive statistics as presented in [5]. For instance, social net- 79
work analysis has been used to study comorbidity, the simultane- 80
ous presence of two chronic diseases or conditions in a patient. By 81
structuring diseases as a network, it is possible to quantify some of 82
the aspects of the complex interactions between conditions in the 83
different patient populations. A number of studies have focused on 84
extensive claims datasets to examine and understand comorbidity 85
networks. In [6], the authors study a diffusion process on a com- 86
orbidity network to model the progressive spreading of diseases 87
on a population depending on demographic data. In [7], the au- 88
thors study how a given chronic disease (diabetes) correlates with 89
age and gender, spanning almost 2 million patients from an en- 90
tire European country. Such comorbidity networks have also been 91
proposed as models to understand the connection between genetic 92
and environmental risk factors for diseases [8]. 93

Beyond clinical purposes, claims data have also been studied 94
to understand the complex interactions of different organizational 95
structures and management relationships involved in patient care 96
processes. For instance, in [9], temporal patterns in electronic 97
health records were modeled in order to present useful informa- 98
tion for decision-making. The authors developed a data represen- 99
tation for knowledge discovery so as to extract useful insights on 100
latent factors of the different processes involving a patient, aiming 101
at improving workflows. 102

Another important trend is the understanding of the relation- 103
ship among healthcare professionals, in particular the physicians. 104
In [10], the authors apply social network analysis to mine networks 105
of physicians which might be used to improve the designation 106
of middle-sized administrative units (accountable care organiza- 107
tions). Sauter et al. [11] use social network analysis to understand 108
networks of healthcare providers which share patients, providing 109
insights in the interplay between general practitioners, internal 110
specialists, and pediatricians. Also, the network structure of differ- 111
ent healthcare providers taking care of a given individual can show 112
important variability of the healthcare system [12]. 113

Social networks have also been used to understand the state of 114
coordination of healthcare actors. In [13], the authors describe a 115
complex network approach applied to health insurance claims to 116
understand the nature of collaboration among physicians treating 117
in-hospital patients and to explore the impact of collaboration on 118
cost and quality of care. Also, in [14], the authors study the social 119
network structure in hospitals among healthcare professionals to 120
understand which variables affect patient care efficiency measures. 121
The idea is further developed in [15] from a statistical point of 122
view in a medium-sized number of hospitals, through the analysis 123
of temporal patterns and costs. 124

The medical referral system in the Canadian healthcare system 125
is studied in [16], where the authors map and analyze the network 126
between general practitioners and specialists. In [17], the authors 127
describe the condition of the basic medical insurance for urban 128
and rural residents in China, then they demonstrate that social net- 129
work analysis can be used in the health insurance claims data to 130
support better understanding of patients transfers among hospi- 131
tals. 132

1 Social network visualization methods can also be powerful to
2 explore and analyze healthcare information, in particular to depict
3 the relationship among healthcare professionals [18].

4 Finally, this work differs from the previous ones as it combines
5 social network analysis, graph mining, and information visualiza-
6 tion. This combination fomented the conversion of a visual pattern
7 identified in the data exploration phase into a network model that
8 reveals underlying mutual referrals from claims database, allowing
9 an in depth scalable analysis of such relationships.

11 3. The insurance claims database

12 The data used in this study was provided by a large Brazilian
13 health insurance company we partnered with. The database con-
14 tains information about services and materials of 108,982,593 in-
15 stances of claims paid by the health insurance company to service
16 providers covering 18 months of activity (about 200,000 claims per
17 day). The database names 279,085 physicians, of which 81% are
18 considered valid, that is, the physician register ID is well formed.
19 Moreover, during the project we had access to information about
20 2,243,198 patients and 26,033 service providers. The database con-
21 tained only claims related to medical consultations performed by
22 physicians, i.e., it does not include claims related to clinical anal-
23 ysis, image-based exams, or hospitalizations. Considering ethical
24 and data protection, the terms of access to database were cov-
25 ered by the contract between our lab and the health insurance
26 company. In addition, given the goals of the project, related to the
27 understanding of physicians' activity, patient sensitive information
28 was not available to our team.

29 Two important aspects related to the quality of the data need
30 to be mentioned. First, approximately 25% of the data do not con-
31 tain information about state (location) in which the service was
32 performed. This proportion increases when the data is modeled as
33 a graph, since that missing state value in one or both nodes (rep-
34 resenting physicians) may hinder an important piece of the analysis.
35 Another important aspect of the data is related to the distribution
36 of physicians' specialties. This attribute is important to correlate
37 with physicians' relationships and crucial to the proper under-
38 standing of the results. However, because of the large amount of
39 missing values, the use of this information in the data analysis had
40 limited scope.

41 As mentioned before, the dataset analyzed contains only claims
42 related to procedures performed by physicians. Therefore, the
43 claims data can be mapped into a social network by considering
44 connections among healthcare professionals or patients. Moreover,
45 additional bits of information in the data may be included in the
46 graph in the form of edges' weights, for instance, the number of
47 events connecting the two nodes, the total expense, etc. For the
48 nodes, the attributes to be included in the network involve de-
49 mography, specialty, location, etc. Both weights and attributes may
50 be represented visually by mapping their values to colors or sizes
51 in the case of nodes, or colors and widths in the case of links.

52 In this work we focus on the relationships between physicians
53 within the health insurance company network. Thus, two physi-
54 cians are considered related if they have a common patient, that
55 is, a patient that had a consultation with both physicians (i.e.,
56 a physician-physician network). This does not indicate a direct
57 relationship, but, for large number of common patients (represented
58 here as an outlier), there is a high likelihood that these physi-
59 cians have some kind of relationship, be it a similar profile, same
60 provider, similar location, similar education background, etc. This
61 signals the possibility that they know each other and have referred
62 their patients to one another.

63 In order to model the flow of patients throughout the physician-
64 physician network, we consider the graph $G(V, E, w)$, where the
65 $|V| = N$ denotes the set of nodes that represent physicians and

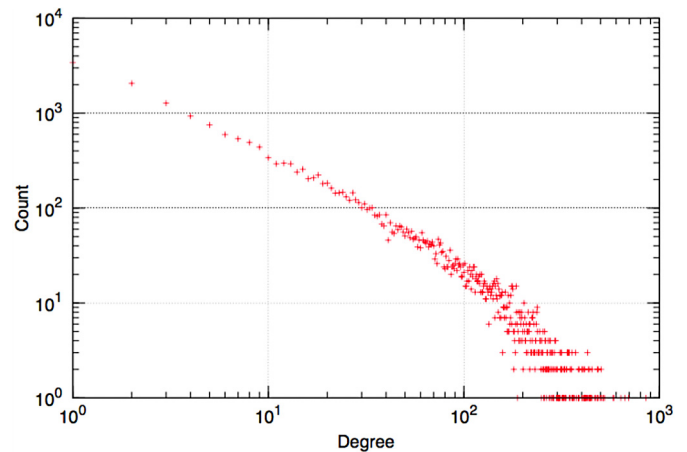


Fig. 1. Node degree distribution of the physician-physician network.

66 $|\mathcal{E}| = M$ denotes the edges that connect physicians that have in
67 common consultations claim of a certain patient. Moreover, node
68 $v_i \in V$ and edges $e_{ij} \neq e_{ji} \in E$. Each edge E has an associated
69 weight $w : E \rightarrow \mathbb{R}^+$ equal to the number of patients that consulted
70 with physician v_i and subsequently with v_j , denoted as w_{ij} .

71 The advantage in using a social network to model the rela-
72 tionships among physicians involves the several ways to quantify
73 the relative importance of the physicians and to visually represent
74 multiple physicians' connections and attributes. For example, social
75 network analysis allows for individualization of physicians that are
76 in a prominent position in the network be it due to the relation-
77 ship with his/her peers, due to her connections with influential
78 physicians, or because without the physician the topology of the
79 network would change substantially (e.g., by increasing the num-
80 ber of connected components).

81 Fig. 1 shows the node degree distribution of the physician-
82 physician network. As we can see, the tail of the distribution fol-
83 lows approximately a power-law, pointing to the fact that the rela-
84 tionship between physicians has a structure also commonly found
85 in other real networks, such as social networks [19]. Here, most
86 physicians are connected to only a few other physicians while a
87 small number of physicians are very well connected in the net-
88 work.

89 In the next sections we describe in detail the model proposed
90 to better understand the physicians' referral behavior in this type
91 of network.

111 4. Modeling mutual referrals

112 This section details the proposed way of modeling mutual ref-
113 errals between physicians from health insurance claims data. It
114 considers healthcare service relationships to identify the underly-
115 ing referral network among physicians.

116 Identifying physicians that work together, especially for consul-
117 tations, is of great business value for health insurance companies.
118 Physicians working together may be due to several reasons. On the
119 one hand, it could be positive for medical care professionals to
120 treat patients together and be able to work as a team, building
121 trust. On the other hand, this pattern could also point to a misuse
122 of health insurance resources. For instance, it could be the case
123 that every time a patient visits a given clinic or physician, she/he
124 is redirected to another physician, even if there is no need to do
125 so. Other situations could involve physicians benefiting certain col-
126 leagues by issuing unnecessary referrals, possibly under terms of
127 reciprocity.

128 The data analysis of the physician-physician network started
129 from the overview of the whole network, followed by zoom and

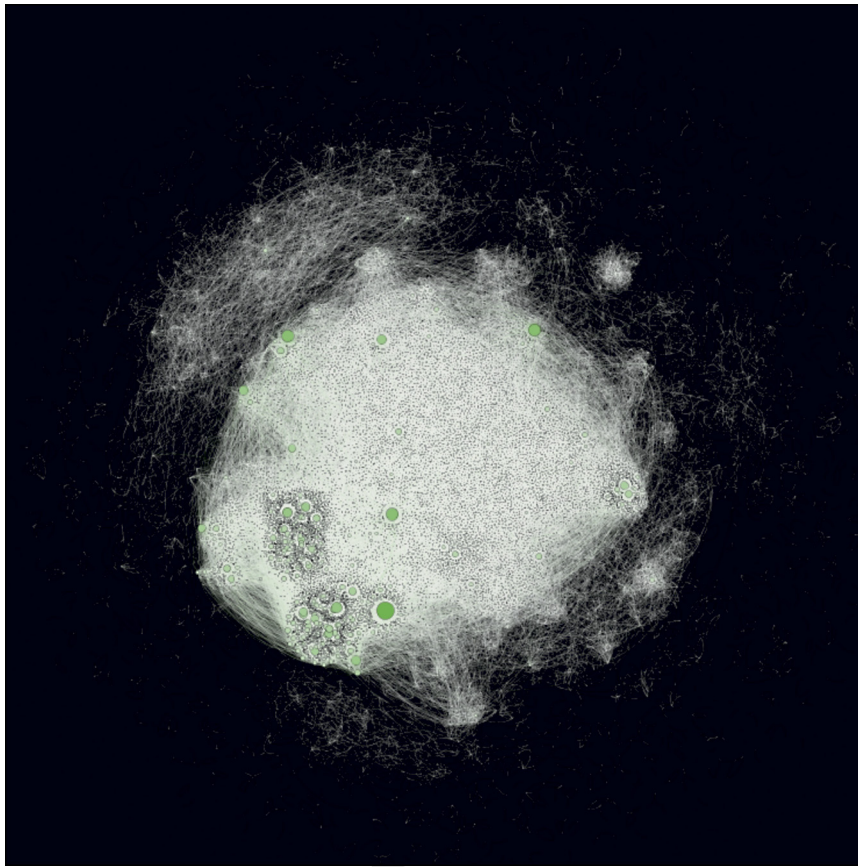


Fig. 2. Directed graph showing the physician–physician network and the flow of consultations. Two physicians (P_1 and P_2) are connected if a patient visited first (P_1 and then P_2). The graph counts on 19,424 nodes, 386,204 edges, has diameter of 20, average path length of 4.71, average degree of 19.34, average weighted degree of 32.12, and density of 0.001. In this visualization, edges' opacity was set to maximum to highlight how the mutual referral visual pattern was identified in the peripheral connected components of the graph, although this opacity hindered connection patterns occurring in the core of the network.

filter, and details on-demand, as proposed by Shneiderman [20]. By visualizing the whole network, it was possible to identify interesting topological characteristics and visual patterns. The mutual referral model focuses on identifying nodes that are part of a visual pattern that emerged during the first visualizations of the network. Fig. 2 shows the whole physician–physician network where connections represent the consultations of shared patients (as detailed in Section 3). In the network it is possible to see the small components at the border of the network and how the connections occur in a reciprocal way between multiple pairs of physicians.

The visual pattern connecting pairs of physicians was clearer when the nodes with small degree (less than the 2nd decile) were filtered (Fig. 3). Fig. 4 shows an example of the same visual patterns for the state of São Paulo, now in a circular layout. In Fig. 4 it is also possible to identify the pairs of mutual connections and degree distribution. These initial results from visual analysis foreshadowed the creation of the proposed model.

Definition 1 (Mutual Referral). The *Mutual Referral* focuses on identifying pairs of nodes v_i and v_j (where $i, j \in [1, N]$ being N the total number of nodes in the network G) connected by edges e_{ij} and e_{ji} , where the weights w_{ij} and w_{ji} are high (the metric increases proportionally to the weights) and, at the same time, as similar as possible to each other (the metric decreases as weights differ):

$$mr(v_i, v_j) = w_{ij} + w_{ji} - |w_{ij} - w_{ji}|,$$

where $mr(v_i, v_j) = mr(v_j, v_i)$. The metric is symmetric to i, j , as expected for a variable describing a property of a given pair

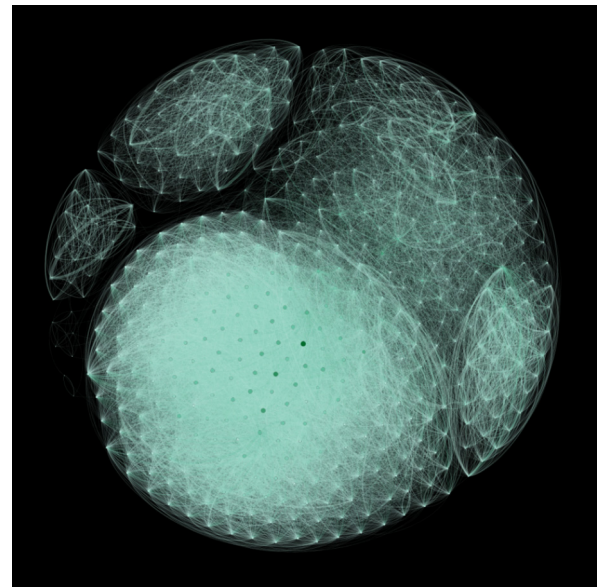


Fig. 3. Directed graph showing the physician–physician network for consultations after filtering nodes with degree smaller than the 2nd decile (270).

of nodes. Note that the second term in $mr(v_i, v_j)$ represents a penalty for those pairs of edges not similar to each other, thus the metric scores higher in case of symmetrical relationships.

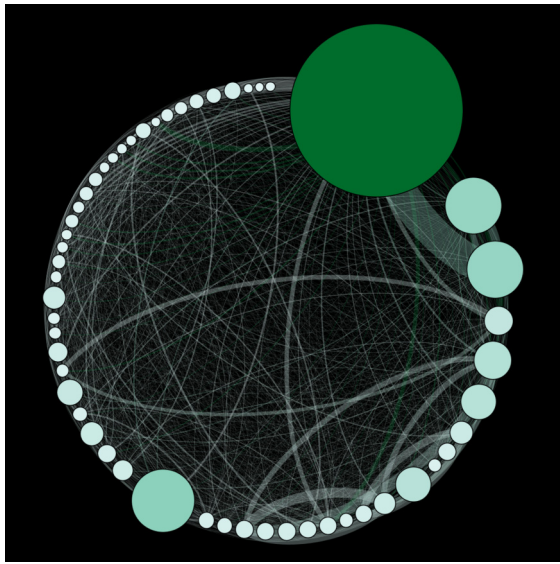


Fig. 4. Directed graph with top 50 physicians in the state of São Paulo in a circular layout. Node color and node diameter represent degree.

To allow for the possibility of a global comparison between pairs of nodes, we defined a mutual referral score (*mrs*), by normalizing by the maximum value of *mr* across the complete graph *G*.

Definition 2 (Mutual Referral Score). The *Mutual Referral Score* measures, in the unit range $[0, 1]$, the relationship between each pair of nodes v_i and v_j , relative to the maximum mutual referral identified in the directed graph *G*, represented by $\max(mr(v_s, v_t))$, $\forall s, t \in [1, N]$. Thus, for any pair of nodes v_i and v_j in *G*, the mutual referral score is defined as:

$$mrs(v_i, v_j) = \frac{mr(v_i, v_j)}{\max(mr(v_s, v_t))}, \forall s, t \in [1, N],$$

where again $mrs(v_i, v_j) = mrs(v_j, v_i)$.

Aiming at identifying connections among physicians when the studied settings change (e.g., population, availability of health services, socioeconomic levels), analyses were performed considering the top 20 and top 50 physician IDs with strongest mutual referral scores from five Brazilian states containing the most claims.

Fig. 5 presents a visualization of five Brazilian states that have different characteristics not only in terms of population, but also in per capita income and healthcare services availability. São Paulo and Rio de Janeiro have the larger population numbers, per capita income, and healthcare services offerings. Nodes represent physicians, nodes' size and color are proportional to node degree, larger/darker nodes represent degrees of higher value. The health insurance company has an important presence in São Paulo and Rio de Janeiro, which can be observed looking at the top 20 physicians in these states. Because of the number of physicians is larger in these states, patients have more options to choose, but they usually choose physicians working with a particular group of other physicians (mainly associated with regions and well-known service providers).

On the other hand, in other states such as Bahia, Pernambuco and Distrito Federal, the presence of the health insurance company is sparser; there are less physicians that work with this company, meaning less options for patients and in consequence more evenly distributed degrees, which can also be identified in the density values of the network (see Fig. 5).

We observe an unusual case in Distrito Federal, where two pairs of heavily connected physicians have the median of days between two consultations equal to zero days, meaning that half or more patients consulted both physicians in the same day. This could represent physicians that work in the same healthcare provider, for example, a cardiologist that executes an electrocardiogram and another cardiologist that analyzes the results right after the exam. However, in both cases we did not have the physicians' specialty information. Actually, most of the physicians that are strongly connected with others do not specify their specialty, which hinders details regarding the context in which the connection occurred.

Table 1 shows the strongest pairs of physicians in the studied database. Except the top pairs from the state of Mato Grosso do Sul, $P_{MS}028$ and $P_{MS}027$, all the other physicians are from the 5 states represented in Fig. 5. Physicians $P_{MS}028$ and $P_{MS}027$ have 205 same-patient consultations (first with $P_{MS}028$ and then with $P_{MS}027$) and 196 consultations (first $P_{MS}027$ and then $P_{MS}028$). None informed their specialty, but with this information, subject matter experts can analyze in detail the highlighted relationships, identifying groups of physicians that work together, misuse health insurance resources by forcing unnecessary referrals, or cases of referral/counter-referral.

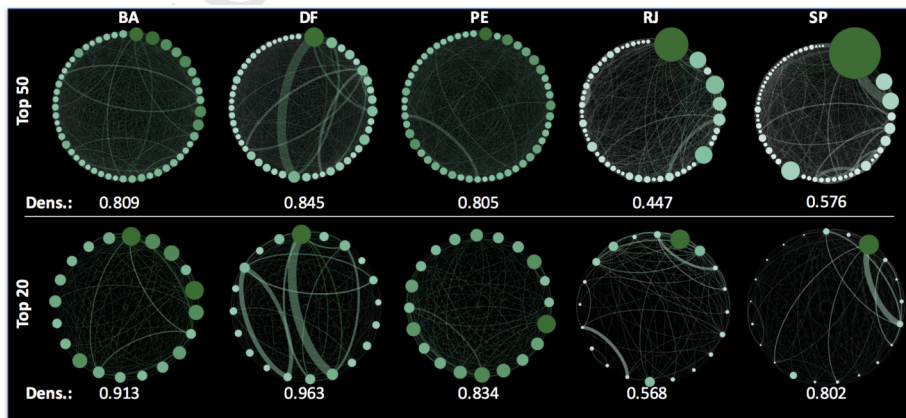


Fig. 5. Directed graphs showing top 50 and 20 physicians with highest mutual referral score from the following Brazilian states: Bahia (BA), Distrito Federal (DF), Pernambuco (PE), Rio de Janeiro (RJ), and São Paulo (SP).

Table 1

The top 10 mutual referral scores in the database; physicians' IDs are anonymized and state codes are used to allow interpretation considering different Brazilian states.

P_1	P_2	Consultations		$mrs(P_1, P_2)$
		$P_1 \rightarrow P_2$	$P_2 \rightarrow P_1$	
$P_{MS}028$	$P_{MS}027$	205	196	1.000
$P_{DF}010$	$P_{DF}009$	267	108	0.551
$P_{SP}022$	$P_{SP}021$	103	102	0.520
$P_{SP}139$	$P_{SP}138$	92	72	0.367
$P_{DF}057$	$P_{DF}056$	73	71	0.362
$P_{SP}141$	$P_{SP}140$	72	70	0.357
$P_{SP}139$	$P_{SP}140$	73	66	0.337
$P_{SP}024$	$P_{SP}023$	92	63	0.321
$P_{SP}143$	$P_{SP}142$	73	63	0.321
$P_{SP}145$	$P_{SP}144$	70	62	0.316

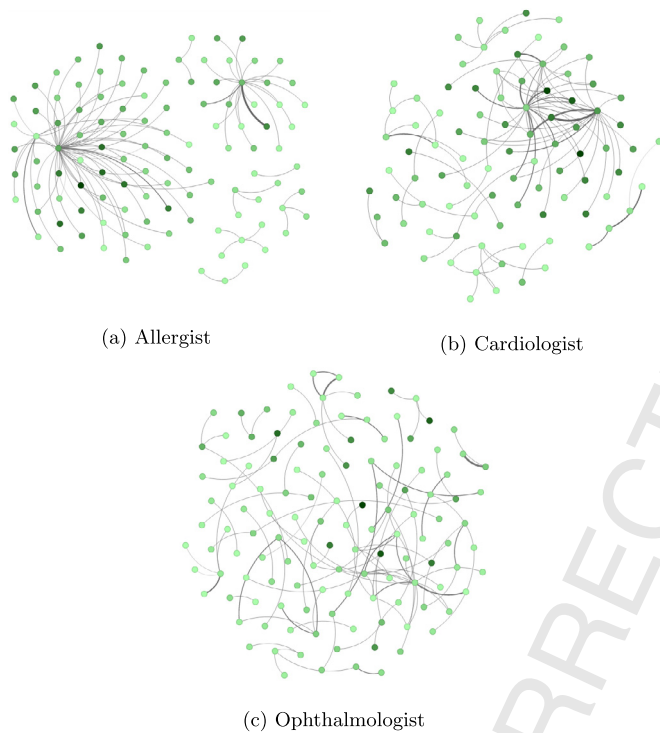


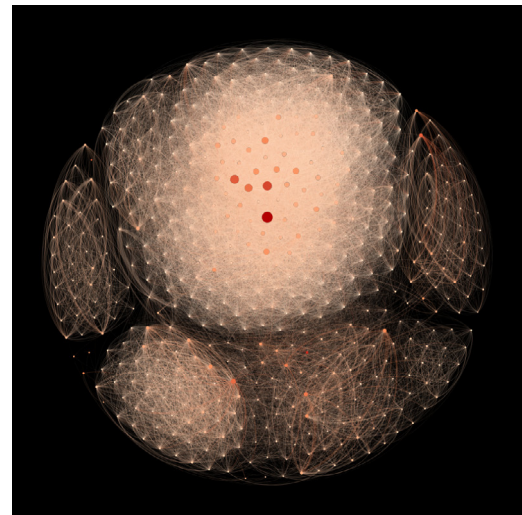
Fig. 6. Undirected graphs showing the top 100 mutual referral scores connected to the following specialties: allergist, cardiologist and ophthalmologist.

Considering the whole network, analysis of those pairs of physicians with highest mutual referral score and with informed specialties revealed the following common specialties referrals:

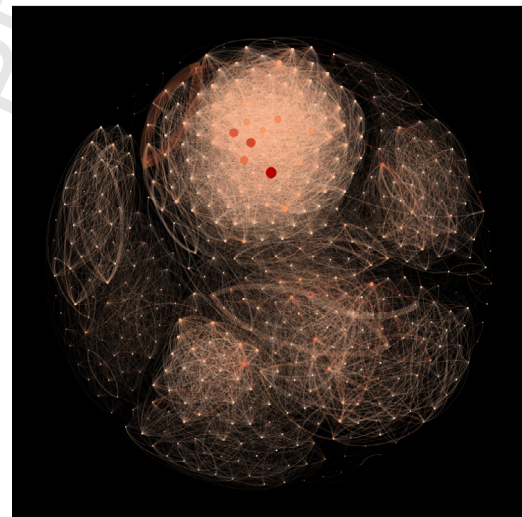
- Ophthalmologist \leftrightarrow ophthalmologist;
- Cardiologist \leftrightarrow cardiologist/vascular surgery¹;
- Acupuncture/pediatrician \leftrightarrow allergist;
- Cardiology \leftrightarrow hematology/clinical pathology;
- Dermatology \leftrightarrow cardiology.

Based on the identified connections considering specialties, the next step considered verifying differences regarding the mutual referrals for physicians with informed specialties. Interesting aspects were revealed considering connections of allergists, cardiologists and ophthalmologists. Fig. 6 shows the network for the top 100 mutual referral scores for different specialties. We can see that in the studied database the referral behavior is different when considering specialty.

¹ Physicians can have more than one specialty informed.



(a) Mutual referral



(b) PageRank

Fig. 7. Directed graphs of top 500 nodes selected by the mutual referral (a) and the top 500 nodes selected by PageRank algorithm (b), both from the physician-physician network generated from the claims database. The graph in (a) counts on 38,597 edges and has density of 0.158 while the graph in (b) counts on 18,099 edges and has density of 0.072. Node size represents in-degree and node color represents PageRank value.

For allergists (Fig. 6a) there are two main physicians connected with several other physicians from different specialties, most of them pediatricians. Thus, it is possible to infer that these two allergists are recommended by most of their pediatrician colleagues. Hence, for this physician-physician network losing these two allergists would cause a negative impact on the network and likely on the health insurance company.

For cardiologists (Fig. 6b) there are two physicians that are strongly connected with several other physicians and nine other cardiologists that have a small number of other physicians connected to them. Finally, the network is different for ophthalmologists (Fig. 6c) in which almost no other physician is strongly connected to multiple physicians. They work mostly in pairs (usually the other ophthalmologist is a surgeon, or a peer responsible for running certain types of exams). Thus, the health insurance company now knows that approaching an ophthalmologist either for registering or improving his/her experience with the company means approaching also his/her peer.

Fig. 7 shows the comparison between our proposal for modeling mutual referrals and the PageRank algorithm [21], both applied

to the same physician–physician network. The PageRank algorithm aims at providing a proxy for nodes' reputation, allowing this reputation to be propagated when nodes with higher reputation “refer” to nodes with smaller reputation values. In the context of health insurance claims data, PageRank would support identifying a new physician in the network being referred by a physician with high reputation in the network. However, PageRank algorithm does not support selecting pairs of physicians highly connected, even in cases that they are disconnected from physicians with high reputation in the network, but, to the best of our knowledge, this is the closer technique to compare considering referrals. This effect can be seen in Fig. 7a. It shows the resulting graph using mutual referral model with a higher number of pairs strongly connected and with higher density (0.158) than in PageRank graph (0.072). On the other hand, Fig. 7b shows the resulting graph using the PageRank algorithm, in which fewer edges are connecting the nodes, but, in turn, existing connections allow for the selection of physicians without mutual relationships to physicians highly referred.

Lastly, the mutual referral model suggests that some physicians consider their own connections to indicate patients. Hence, in situations where the health insurance company wants to improve the relationship with physicians, it could consider an approach involving groups of doctors that already act together, increasing the involvement of the group as a whole with the health insurance company and its services. More generally, this metric could support decision-making processes involving increasing or reducing the network of accredited physicians/service providers. Connecting to our hypothesis (i.e., *it is possible to identify underlying physicians' referrals from claims data*). The pairs of physicians emerging from the visual analysis, the connections between specialties, and the relationship involving location of the service provider are all in accordance to facts highlighted by subject matter experts.

5. Discussion of results

The results show interesting aspects related to social network analysis, information visualization, and the healthcare insurance business. Considering the main contribution of the paper, the proposed model represents a step forward towards revealing underlying characteristics of physician–physician networks and physicians' referral behavior in the context of private healthcare services.

The study was conducted considering multiple problems the health insurance company presented us in the context identifying physicians that excel in their daily work. After multiple interactions with subject matter experts, this goal was decomposed in minor challenges, one of them being the referral. The importance for the health insurance company in improving the relationship with professionals involves increasing the quality of service as whole.

The model for mutual referrals proposes a proxy for identifying how physicians refer peers to their patients. The act of referring a physician occurs informally and is not coded anywhere in the data. The presented approach not only was able to detect those relationships but also provided insights considering social ties, time between referrals, location, and specialty attributes. Moreover, the mutual referral can be used as a metric in any type of service involving people at both ends (provider and consumer) and where this informal referral between peers might occur.

The comparison between the mutual referral and the PageRank algorithm [21] in the selection of the top 500 physicians highlighted the main differences between them. On the one hand, mutual referral supports selecting pairs of physicians that are not connected to the core of the physician–physician network. This case is interesting for health insurance companies with country-wide operations, when the resulting graph might result in a network with multiple connected components with multiple locally important connections. On the other hand, PageRank supports the identifi-

cation of physicians that are new in the network and that are referred by a physician highly influent in the network. However, in the context of physician–physician network, such connections could also occur by chance, scenario that is reduced in the case of mutual referral given the nature of the model to value reciprocity. Finally, the use of mutual referral approach or PageRank algorithm in the context of health insurance data depends on the question to answer. If the question involves identifying pairs of mutual relationships, the mutual referral can be used as a starting point for highlighting such underlying relationship. If the question involves identifying how physicians highly important in the network refer physicians recently added ones, then PageRank seems an appropriate first step for such analysis.

6. Conclusion

In this work we have shown how information visualization and social networking techniques can be applied to the analysis of health insurance claims data, mainly by mapping physicians using shared patients as a proxy for a relationship between them. The resulting model provided useful insights to the health insurance company we partnered with. The way of identifying mutual referrals improved the understanding of important characteristics both from physicians and the flow of patients considering consultations, specialty, location, and time between the consultations. Those insights can have multiple business applications such as to detect frauds or to improve the relationship between the health insurance company and important physicians for the company.

We demonstrated, by the analysis of a real database of claims from a large Brazilian healthcare insurance company, how health insurance data can be modeled as a social network supporting a structural analysis as opposed to traditional transactional approaches. Moreover, our results point out that the complex physician–physician network derived from the health insurance claims database has characteristics similar to a social network, opening multiple paths for this research to follow, in particular, considering theories and techniques from Social Network Analysis.

The data analysis and the value of the model for the business of the health insurance company we partnered with were validated through weekly meetings involving our team and the health insurance company staff, including physicians, process analysts, and IT specialists. Those interactions with subject matter experts were key for the proper understanding of the database, processes, and the identification of the most important cases in terms of decision-making support and business value. The company was already performing analysis of high cost procedures and flow of patients throughout the health insurance services. However, our results provided them with new tools for performing analysis at scale, considering the data they have at their own disposal.

Future works involve a more detailed study of the proposed approach, including authorization processes and clinical exams/surgeries. Moreover, building on the presented results, we plan to apply Social Network Analysis algorithms (e.g., community detection, link prediction, among others) in order to deepen the analysis involving groups of physicians that work together.

References

- [1] U. Srinivasan, B. Arunasalam, Leveraging big data analytics to reduce healthcare costs, *IT Prof.* 15 (6) (2013) 21–28.
- [2] J. Andreu-Perez, C.C. Poon, R.D. Merrifield, S.T. Wong, G.-Z. Yang, Big data for health, *IEEE J. Biomed. Health Inform.* 19 (4) (2015) 1193–1208.
- [3] P.B. Jensen, L.J. Jensen, S. Brunak, Mining electronic health records: towards better research applications and clinical care, *Nat. Rev. Genet.* 13 (6) (2012) 395–405.
- [4] B. Burton, P. Jesilow, How healthcare studies use claims data, *Open Health Serv. Policy J.* 4 (2011) 26–29.

- [5] V. Chandola, S.R. Sukumar, J.C. Schryver, Knowledge discovery from massive healthcare claims data, in: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '13, ACM, New York, NY, USA, 2013, pp. 1312–1320, <http://doi.acm.org/10.1145/2487575.2488205>.
- [6] A. Chmiel, P. Klimek, S. Thurner, Spreading of diseases through comorbidity networks across life and gender, *New J. Phys.* 16 (11) (2014) 115013.
- [7] P. Klimek, A. Kautzky-Willer, A. Chmiel, I. Schiller-Frühwirth, S. Thurner, Quantification of diabetes comorbidity risks across life using nation-wide big claims data, *PLoS Comput. Biol.* 11 (4) (2015) e1004125.
- [8] P. Klimek, S. Aichberger, S. Thurner, Disentangling genetic and environmental risk factors for individual diseases from multiplex comorbidity networks, arXiv preprint arXiv:1605.09535.
- [9] N. Lee, A.F. Laine, J. Hu, F. Wang, J. Sun, S. Ebadollahi, Mining electronic medical records to explore the linkage between healthcare resource utilization and disease severity in diabetic patients, in: 2011 First IEEE International Conference on Healthcare Informatics, Imaging and Systems Biology (HISB), IEEE, 2011, pp. 250–257.
- [10] B.E. Landon, J.-P. Onnela, N.L. Keating, M.L. Barnett, S. Paul, A.J. O'Malley, T. Keegan, N.A. Christakis, Using administrative data to identify naturally occurring networks of physicians, *Med. Care* 51 (8) (2013) 715.
- [11] S.K. Sauter, L.M. Neuhofer, G. Endel, P. Klimek, G. Duftschmid, Analyzing healthcare provider centric networks through secondary use of health claims data, in: 2014 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI), IEEE, 2014, pp. 522–525.
- [12] K.D. Mandl, K.L. Olson, D. Mines, C. Liu, F. Tian, Provider collaboration: cohesion, constellations, and shared patients, *J. Gen. Intern. Med.* 29 (11) (2014) 1499–1505.
- [13] F. Wang, U. Srinivasan, S. Uddin, S. Chawla, Application of network analysis on healthcare, in: 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), IEEE, 2014, pp. 596–603.
- [14] S. Uddin, L. Hossain, et al., Social networks in exploring healthcare coordination, *Asia Pac. J. Health Manag.* 9 (3) (2014) 53.
- [15] S. Uddin, Exploring the impact of different multi-level measures of physician communities in patient-centric care networks on healthcare outcomes: a multi-level regression approach, *Sci. Rep.* 6 (2016).
- [16] W. Almansoori, O. Zarour, T.N. Jarada, P. Karampales, J. Rokne, R. Alhadj, Applications of social network construction and analysis in the medical referral process, in: 2011 IEEE Ninth International Conference on Dependable, Automatic and Secure Computing (DASC), IEEE, 2011, pp. 816–823.
- [17] H. Guo, F. Wei, S. Cheng, F. Jiang, Find referral social networks, in: 2015 International Symposium on Security and Privacy in Social Networks and Big Data (SocialSec), IEEE, 2015, pp. 58–63.
- [18] L.G. Moyano, A.P. Appel, V.F. de Santana, M. Ito, T.D. dos Santo, Graphys: understanding health care insurance data through graph analytics, in: Proceedings of the 25th International Conference Companion on World Wide Web, International World Wide Web Conferences Steering Committee, 2016, pp. 227–230.
- [19] M. Newman, *Networks: An Introduction*, OUP, Oxford, 2010.
- [20] B. Shneiderman, The eyes have it: a task by data type taxonomy for information visualizations, in: Proceedings, IEEE Symposium on Visual Languages, 1996, IEEE, 1996, pp. 336–343.
- [21] S. Brin, L. Page, Reprint of: The anatomy of a large-scale hypertextual web search engine, *Comput. Netw.* 56 (18) (2012) 3825–3833.