InfoSAS: a data mining system for production control of SUS [Brazilian public healthcare system]

Osvaldo Carvalho
has a PhD in Computer Science from the Pierre et Marie Curie University, France, and is an associate professor at the Federal University of the State of Minas Gerais (UFMG). He worked with distributed algorithms and currently teaches Computer Programing and develops information systems.

Marcos Prates
has a B.A. in Computational Mathematics from the Federal University of the State of Minas Gerais (UFMG), an M.A. in Statistics from the same institution and a PhD in Statistics from the University of Connecticut. He develops statistic methods and algorithms for the analysis of spatial statistics and machine learning.

Raquel Minardi
has a B.A. in Computer Science and a PhD from the Federal University of the State of Minas Gerais (UFMG). She also has a postdoc from the French Alternative Energies and Atomic Energy Commission (CEA). She develops models and algorithms in data visualization and in computational biology. She has published over 20 articles in international journals and in national and international conferences.

Wagner Meira Jr.
has a PhD from the University of Rochester and is a Computer Science professor at the Federal University of the State of Minas Gerais (UFMG). His areas of interest are data mining, parallel and distributed systems and their applications.

Renato M. Assunção
has a PhD in Statistics from the University of Washington and is a professor in the Computer Science Departament of UFMG. He is also a specialist in algorithms development and methods for the analysis of statistic data, especially those with georeferencing.

José Nagib Cotrim Árabe
has a PhD in Computer Science from the University of California, Los Angeles (UCLA), U.S.A., and is a professor at the Federal University of the State of Minas Gerais (UFMG). He is head of the Computer Science Department.

ABSTRACT

This paper introduces InfoSAS, a system for detection of statistical anomalies in SUS production records. InfoSAS finds *per capita* inhabitant service rates that are much higher than the Brazilian average or hospitalization prices way over those charged by most institutions for the same procedure. Results show that hundreds of millions of Brazilian Reais spent by SUS are destined for treatments considered anomalous based on conservative criteria. Statistical anomalies may result from fraud, but also from healthcare intensive programs, epidemics, or poor distribution of healthcare services. In any case, serious anomalies should be investigated or explained.

Keywords: SUS, Brazilian public healthcare system, Anomaly detection, Data mining, Service rate per inhabitant, Hospitalization average price, Log-normal distribution, Audit prioritization.

1. INTRODUCTION

The large size of the healthcare sector and its huge amounts of funds turn healthcare into an attractive target for fraud in the entire world. In the United States, for instance, over 270 billion dollars are lost annually due to fraud (THE ECONOMIST, 2014). There is no reason to believe that the scenario in Brazil is different. In this paper, we introduce InfoSAS, a system for mining and viewing SUS information that, like many other systems (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996), can find varied data models and standards and could be very useful for SUS management. In particular, for identifying statistical anomalies that might indicate deviations. This paper focuses on its use for audit purposes.

InfoSAS can find many facts requiring managers' attention, including service rates per inhabitant much higher than Brazilian average levels, as shown in Figure 2; or hospitalization prices way over those charged by most institutions, as shown in Figure 4. As seen in Section 5, we have reason to believe that using such system in audit prioritization may produce reimbursements to SUS of BRL400 million per year. It is no exaggeration to estimate that the amount recovered could reach billions of Brazilian Reals as a result of future audits of production data for a period of 8 years (from 2009 through 2016), as well as from reducing anomalous behaviors in future production.

2. STATISTICAL DISCREPANCIES

Services delivered by SUS (jointly called SUS production) are recorded by means of several instruments and those records are entered into the SIA[1] and SIH[2] databases. SUS also keeps the CNES[3] database, a record of healthcare institutions. InfoSAS searches those SUS databases as well as population information provided by IBGE [Brazilian Institute of Geography and Statistics], and produces what we call factsheets.

These databases are massive and those factsheets provide focus when examining them. They match a mining target, an examination period and an institution, showing charts and tables found when examining those databases. A mining target is a SUS production subset, defined by one or more procedures listed in the SUS Table[4] and, sometimes, by patient age group as well. It is worth noting that facts from those sheets can also be verified on an independent basis by checking other sources of information provided by SUS, such as TabWin[5].

Figure 1 shows an example of a factsheet chart. In the chart, we can see that, for the target "Treatment of Visual Apparatus Diseases", all services in Curitiba from September 2013 through August 2014 met expectations based on the Brazilian average, that the numbers of those services are higher than the average of the State of Paraná, and that their provider contributed to less than half of those services. This information is consistent with a regular scenario.

However, that is not always the case. Some facts call the attention of audit professionals. In Figure 2, some cities have service rates per inhabitant much higher than Brazilian and State rates. The institution is practically the only one servicing those cities, a clear

Figure 1:
Examination of the participation of an institution in services in Curitiba for the target "Treatment of Visual Apparatus Diseases"

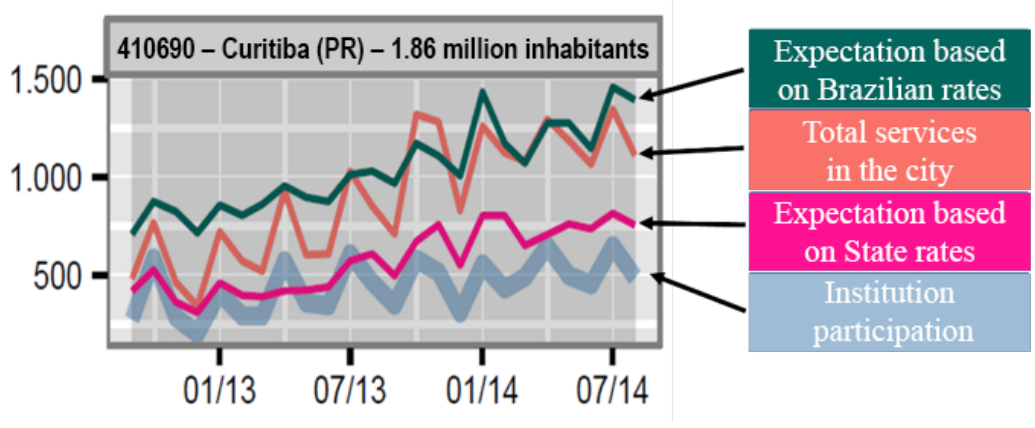


Figure 2:
Analysis of service rates per inhabitant for six cities serviced by an institution for the target "Treatment of Visual Apparatus Diseases"

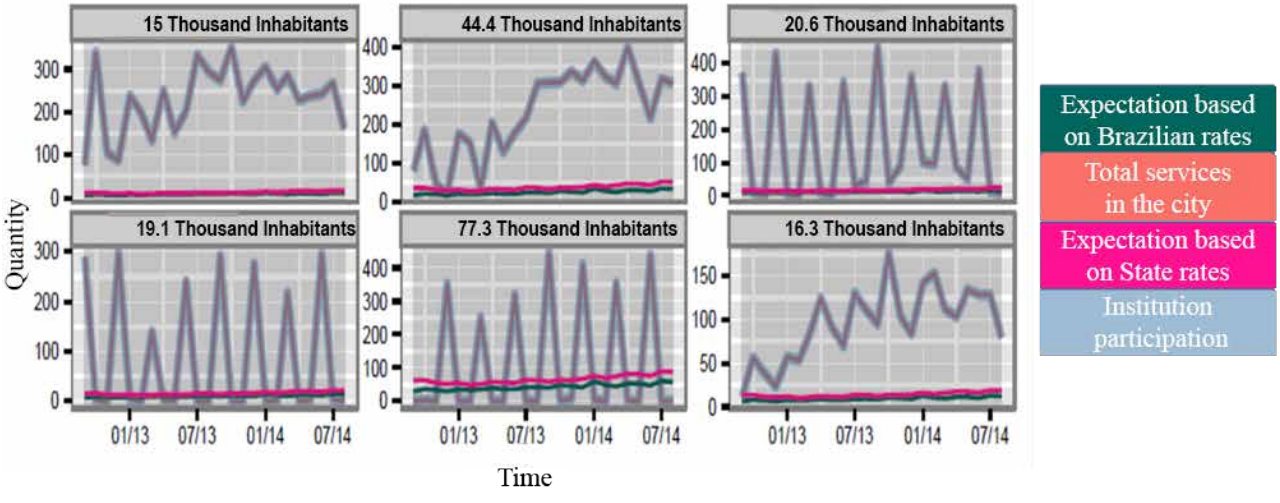
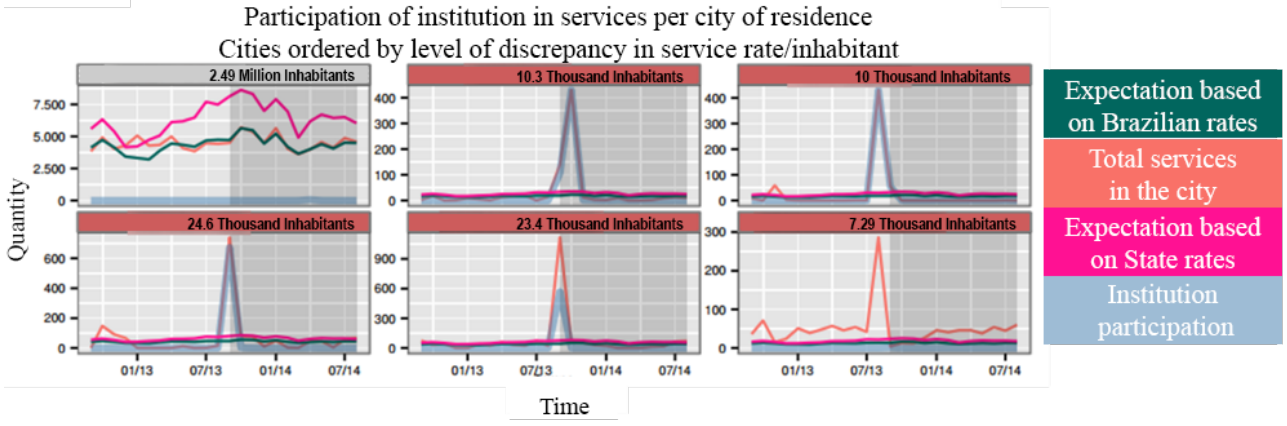


Figure 3:
Statistical discrepancies explained by target characteristics, "Mammogram bilateral screening", and by provider, a Women's Mobile Medical Unit

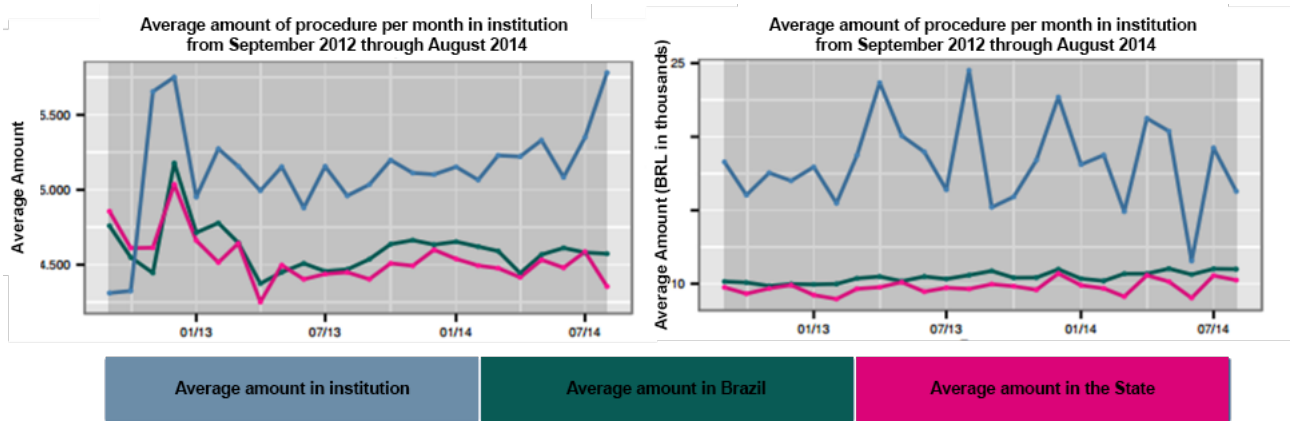


example of what we consider a statistical anomaly or discrepancy.

Statistical discrepancies should be considered cautiously, as shown in the example in Figure 3. Each of the six charts refers to one city and all of them refer to only one provider. We note that only the top left chart (which refers to the city where the relevant institution is headquartered) does not have unusual peaks. It has amounts much higher than expected for Brazilian rates and for the State where those cities are located. Is that fraud evidence? No. Since the target is "Mammogram" and the provider is a mobile medical unit, the statistical anomaly is explained as follows: the provider is a bus equipped with a mammography machine that performs intensive healthcare programs in small towns, leading to service peaks when visiting those towns.

It is worth pointing out that, in general, a statistical anomaly can actually be fraud evidence, but it may also result from correct actions, as shown above, or from incorrect information, or epidemics. It is also possible that cities with high service rates are actually those few ones that do provide good services in the country. Thus, good judgment should be exercised when analyzing a factsheet, which should be done by a public healthcare professional aware of the local context.

Figure 4:
Analysis of amount charged by one institution for hospitalizations for targets "Cementless/hybrid primary total hip arthroplasty", on the left, and by another institution for "Cardiac surgery – Pacemaker cardiac surgery", on the right



In Figure 4, there are other examples of statistical discrepancies, wherein institutions charge average hospitalization prices that are much higher than prices charged in the remainder of Brazil for two mining targets. Once again, the manager shall exercise caution in each case of statistical discrepancy, since the institution could have a profile of patients with above average complications, or follow practices that can increase hospitalization cost but also reduce the mortality rate.

3. AUTOMATIC DETECTION OF STATISTICAL DISCREPANCIES

Finding factsheets that are of interest to oversight in a Healthcare Secretariat is a convincing factor for the decision to investigate. On the other hand, searching relevant discrepancies randomly is like finding the needle in the haystack. We have over 5000 targets and approximately 6000 providers. Considering only 12-month periods, for 3 years of production, there would be 36 possible time windows. A simple calculation shows that, literally, billions of factsheets may be extracted from the databases examined.

That is why InfoSAS is so invaluable. It uses various algorithms to search discrepancies, producing scores that can sort and prioritize factsheets. InfoSAS also allows its user to focus in areas defined by geographic filters and by exam period and medical target, since an audit and evaluation professional many times directs his/her attention to particular healthcare fields, such as cardiology or orthopedics and, of course, tends to have more interest in the region where they work.

InfoSAS analyses the time series of average monthly value per procedure and of monthly production for each target desired. These time series are calculated per institution and patients' city of residence. Various anomaly detection algorithms are used and each of them calculates a score. Those scores are combined subsequently. Since there is not enough space here to detail all scores, we will define only two of them below. A more detailed description of another algorithm can be found in (CARVALHO et al., 2015).

The purpose of an algorithm is to detect sudden variations in a provider's production. Therefore, the institution production is compared to its own historic production series (or average cost). In the event of any abrupt deviation from its historic series, the algorithm assigns a discrepancy score to the institution. More specifically, in for each month i and for each institution l, we use the formula $score_{li}=(t_{li}+1)/(mean_{li}+1)$, where t_{li} is the interest level (that may be the cost or number of services) for institution l in month *i*, $mean_{li}$ is the mean of the last m months for the interest level of institution *l* in month *i*. Typically, we adopt m=12 or m=6 months.

Another algorithm seeks to find discrepancies in service rates per inhabitant. Based on Brazilian production per 100 thousand inhabitants, the production time series carried out by all institutions for residents within one city are examined. In the event this city has a production that is higher than threshold, the algorithm assigns a discrepancy score to it. Afterwards, the score is assigned to institutions proportionally to the participation of each one in delivering services to that city.

An institution score is the result of the addition of all scores assigned to it in each city serviced. More formally, a city production score is computed as the last 12 months cumulative sum, $score_{li}=\sum_{j=i-11}^{i} dif_{lj}$, where $dif_{li}=max\{0,tBayes_{li}-threshold\}$ and $threshold=k*Brazilianrate_i$. The value of $Brazilianrate_i$ is Brazilian monthly production rate per 100 thousand inhabitants in month *i*, $tBayes_{li}$ is the empirical Bayesian rate per 100 thousand inhabitants in month *i* for the city *l*, and *k* is a constant previously defined. In our studies, we adopt k=3. The empirical Bayesian rate (ASSUNÇÃO et al., 1998; MARSHALL, 1991) is a statistical

technique to calculate rates and ratios not affected by statistical fluctuations due to small populations.

4. INFOSAS OPERATION

Figure 5 shows InfoSAS dataflow. Every month, SIA, SIH and CNES databases, and IBGE population information, are entered into the data mining server, which is guided by a table of mining targets and runs algorithms to find statistical discrepancies. The output of this mining stage is referred to as the mined fact cube. This cube is explored by InfoSAS users through a BI tool. The user selects a report model and specifies parameters to filter target sets, analyses periods and geographical cuts. A report is issued using the selected model and filters, and it bears discrepancy scores calculated by mining algorithms. The user extracts from the report those factsheets drawing his/her attention to carry out more in-depth analyses.

5. CONCLUSION AND FUTURE WORKS

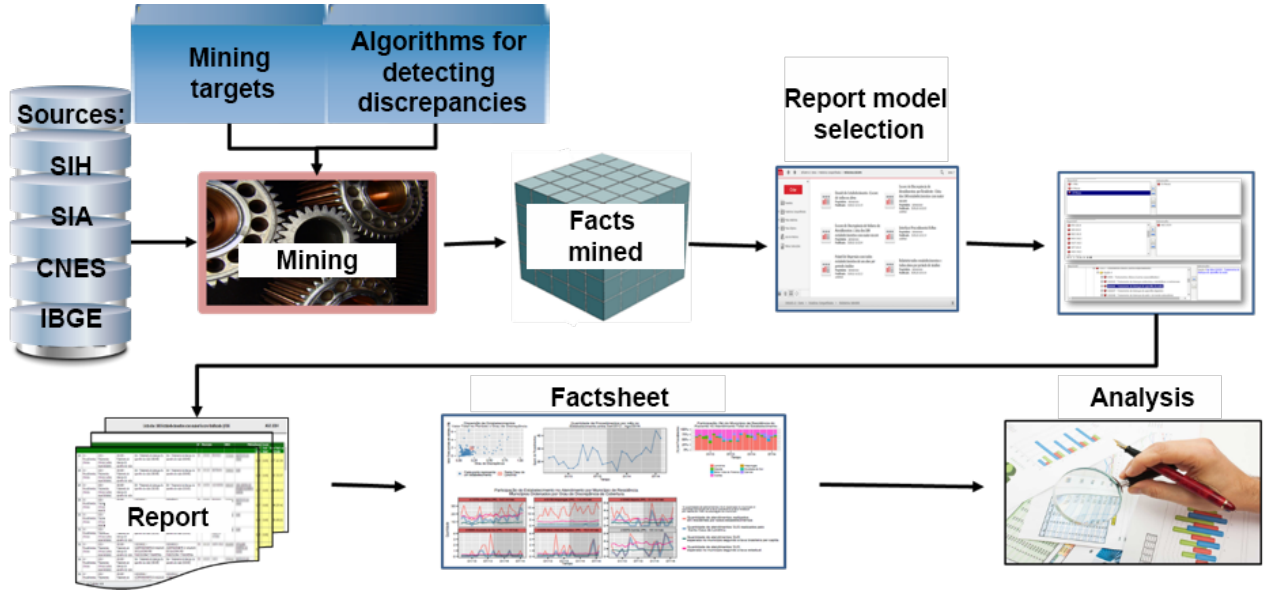
Currently (October 2016), InfoSAS is installed and running on DATASUS. A distance-learning course to train managers to use InfoSAS is being prepared and expected to be ready by November 2016 and offered by the end of this year. InfoSAS system has already been presented at several Brazilian events (DRAC SAS, 2015). In all of them, public health specialists found the results very interesting and useful.

Results obtained by project InfoSAS so far are important and are a step forward in modernizing selection processes for audit and control items. In addition to the need for system maintenance, continuous updates, fixes and fine calibration of algorithms already used, there are many courses for future developments. We envision a new interface with a more flexible and interactive preview of results; a statistical characterization of targets identifying healthcare gaps; the use of SUS population in statistical computation; the calculation of upper and lower limits in order to match a city anomalous production to its providers; the adapted selection of geographic cuts with proper sizing of the frequency of occurrence of each target; and the use of parallelism in preview and mining processing.

In addition, we point out two developments that, in our opinion, are the most important ones. The first one would be feasible by conducting audits, perhaps oriented by InfoSAS alerts, classifying each service as regular or fraudulent. Such classification would make possible the use of supervised learning algorithms, enabling alerts to be issued on a more accurate and complete basis.

The second development consists of a consolidated analysis of anomalies in service rates per inhabitant. The current version of InfoSAS uses several algorithms that try to capture statistical anomalies. We also have a mining target framework where one target may in-

Figure 5:
Data analysis stages by InfoSAS system



clude other targets. Thus, we obtain many findings but we cannot classify institutions by target set nor obtain global estimates of values for anomalous production.

To produce a new report overcoming these problems, we decided to have as targets only those sets of procedures from the same *organization form* of the SUS Table. This causes the intersection between any two targets to be empty. We have also decided to use only one algorithm with a statistical methodology to estimate excessive service rates per inhabitant. This makes possible to find global estimates (in all targets) for anomalous production for each provider, allowing prioritization of fact finding by SUS managers.

The starting point of the methodology we intend to use and already applied to data for 2013 is the fact that the distribution of service rates per inhabitant observed in cities follows a log-normal distribution, as shown in Figure 6. This makes possible to define, for any target and any city, a cut-off point to distinguish what is normal from what is not in service rates per inhabitant.

We regard service rates per inhabitant as abnormal when located to the right of the point dividing the area under a curve in a "normal" part, with 99% probability; and under a curve in an "abnormal" part, with 1% probability, which we deem as a very prudent criterion. In the example shown in Figure 7, rates above 3.2 services per 1,000 inhabitants/month are considered abnormal. Assuming, for that target, that a city with 100,000 inhabitants received 400 services in a given month. By using the maximum rate of 3.2 services/1,000 inhabitants, the population in that city should have received, at most, 320 services. Therefore, we consider that the city had 80 anomalous services, and this excess amount is allocated among providers in the city, unchanging each provider service shares.

The results of applying this analysis to the production entered in SUS in 2013 are very strong. Out of a total of BRL19,912,491,904.00, services considered abnormal totaled BRL413,920,365.56, corresponding to 2.08% of the total. Excess in targets entered into SIH-SUS was BRL350,354,969.25, and in targets entered into SIA was BRL63,565,396.30.

Another result of this study that may be used to prioritize audits is an analysis of allocations among providers of services regarded as anomalous. Throughout Brazil, only 5 providers account for almost 9% of total anomalies; 100 providers account for over 50% of that total. One single provider received almost 10 million Brazilian Reals in services deemed as anomalous in 2013.

6. ACKNOWLEDGMENTS

First of all, we would like to thank the Department of Regulation, Evaluation and Control of the Office of Healthcare Attention, Ministry of Health, which requested and funded the whole project.

Figure 6:
Density of probability of service rates per inhabitant: theoretical log-normal vs. observed data

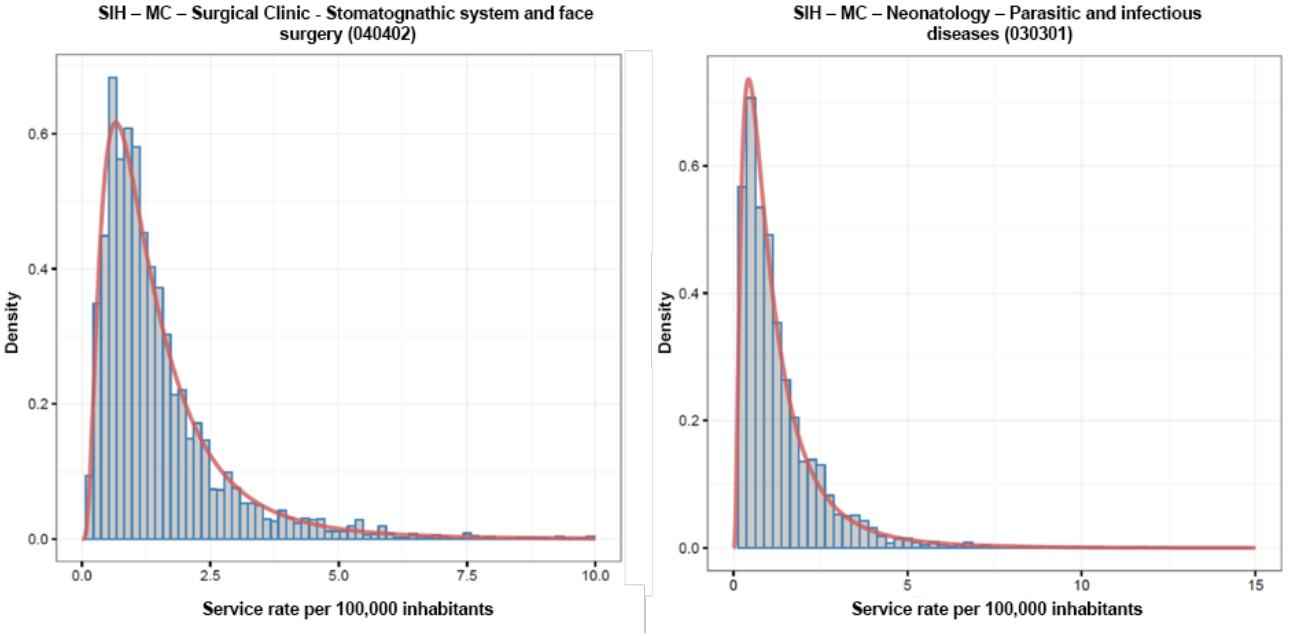
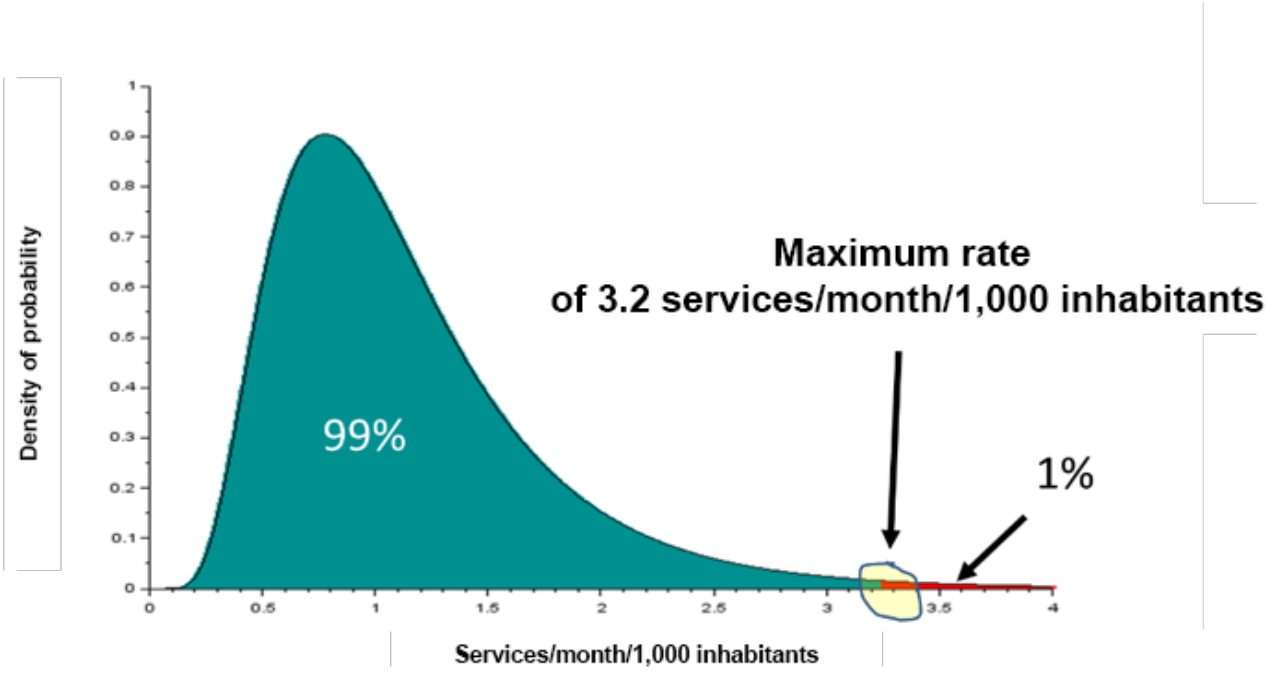


Figure 7:
Definition of maximum service rate per inhabitant at the cut-off point of 1% of log-normal distribution



InfoSAS was built by a large multidisciplinary team. Maria Helena Brandão conducted the project for DRAC. Ester Dias, Marcelo Campos, Sônia Gesteira and Suzana Rattes, from SMSA-BH, Luciana Morais, from Funed, and Mônica Castro, from Unimed-BH, were our health consultants. Fabiana Peixoto and Letícia Neto were the managers, and Tomas Schweizer was the technical leader. Edré Moreira, Carlos Teixeira, Douglas Azevedo, Larissa Santos, Luiz Fernando Carvalho, Luiz Gustavo Silva, Maurício Nascimento Jr., Milton Ferreira, José Carlos Serufo Jr., Pablo Fonseca, Renan Xavier and Raquel Ferreira, graduate scholarship holders participated proactively in the research, building and implementation of various algorithms. Felipe Caetano, Ícaro Braga, Geraldo Franciscani, João Paulo Pesce, João Victor Bárbara, Raphael de Faria and Wicriton Silva were in charge of development, tests, previews and databases. Our sincere thanks go to all the members of the team, the ones actually responsible for building the system.

REFERENCES

ASSUNÇÃO, R. M. et al. Maps of epidemiological rates: a Bayesian approach. Cadernos de Saúde Pública, Rio de Janeiro, v. 14, n. 4, p. 713–723, Oct/Dec 1998.

CARVALHO, L. F. M. et al. A simple and effective method for anomaly detection in healthcare. In: 4TH WORKSHOP ON DATA MINING FOR MEDICINE AND HEALTHCARE, IN CONJUNCTION WITH THE 15TH SIAM INTERNATIONAL CONFERENCE ON DATA MINING, 2015, Vancouver, May 2015. Available at http://homepages.dcc.ufmg.br/~carlos/papers/sdm/dmmh2015.pdf. Web: Nov 29, 2016.

DRAC SAS. Ministério da Saúde. Ciclo de Oficinas do DRAC – Controle de Avaliação. Brasília, DF, Sept 1, 2015. Available at: <https://www.youtube.com/watch?v=vIaR_Q7T-Us>. Web: Jul 4, 2016.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. AI magazine, California, v. 17, n. 3, p. 37-54, 1996.

MARSHALL, R. J. Mapping disease and mortality rates using empirical Bayes estimators. Applied Statistics, Auckland, v. 40, n. 2, p. 283–294, 1991.

THE ECONOMIST. The $272 billion swindle. London, May 31, 2014. Available at http://www.economist.com/news/united-states/21603078-why-thieves-love-americas-health-care-system-272-billion-swindle. Web: Nov 29, 2016.