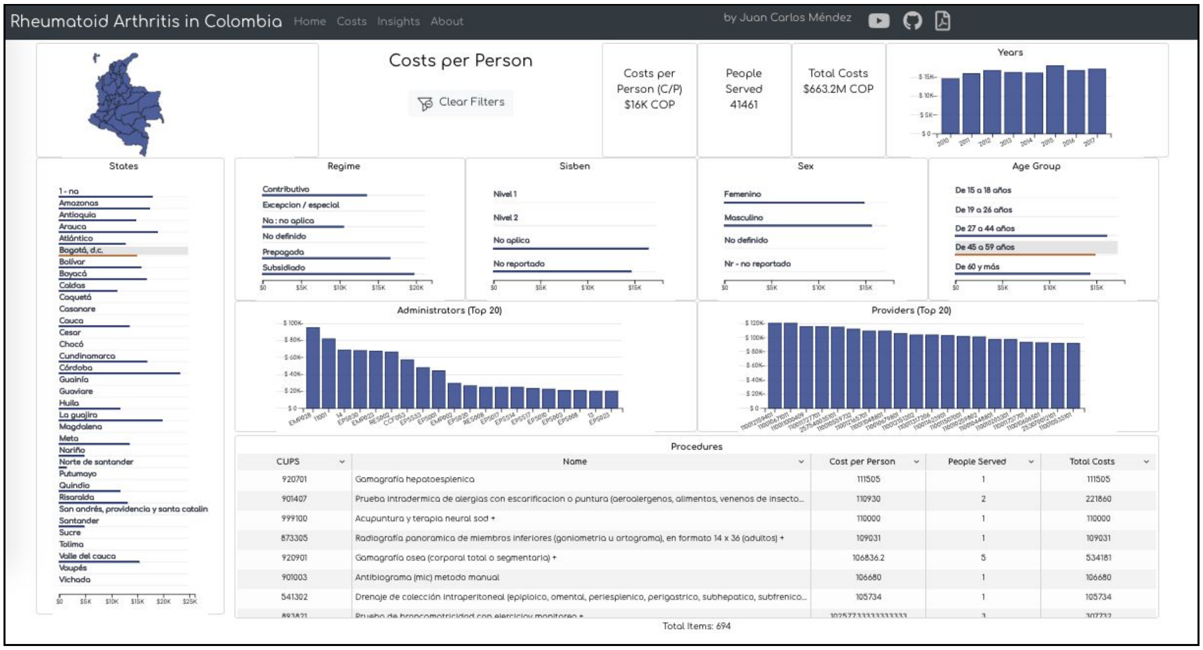
Applying Visual Analytics in Health: Costs of Rheumatoid Arthritis in Colombia

Méndez M. Juan C.

Universidad de los Andes

jc.mendez@uniandes.edu.co



Procedures

CUPS	Name	Cost per Person	People Served	Total Costs
920701	Gamagrafía hepatoesplenica	111505	1	111505
901407	Prueba intradermica de alergias con escarificacion o puntura (aeroalergenos, alimentos, venenos de insecto...	110930	2	221860
999100	Acupuntura y terapia neural sod +	110000	1	110000
873305	Radiografía panoramica de miembros inferiores (goniometria u ortograma), en formato 14 x 36 (adultos) +	109031	1	109031
920901	Gamagrafía osea (corporal total o segmentaria) +	106836.2	5	534181
901003	Antibiograma (mic) metodo manual	106680	1	106680
541302	Drenaje de colección intraperitoneal (epiploico, omental, periesplenico, perigastrico, subhepatico, subfrenico...	105734	1	105734
893821	Prueba de broncomotricidad con ejercicio y monitoreo +	102577.33333333333	3	307732

Total Items: 694

Fig. 1. An interactive Dashboard that helps to analyze the costs of Rheumatoid Arthritis (RA) in Colombia.

Abstract—Rheumatoid arthritis (RA) is an autoimmune disease that can cause joint pain and damage throughout your body. There's no cure for RA, but there are treatments that can help you to manage it. In addition to to physical and emotional pain, the economic costs associated are high. In general, it is considered as a high-cost disease. The following work tries to bring a visual analytics tool that could help to understand the impact of Rheumatoid Arthritis (RA) in Colombia in terms of the economical costs associated with it. The cost of procedures vary from state, regime, age, administrator, provider, etc. Using a visual tool could help the experts to explore and understand the available data.

Index Terms—Visual Analytics, Rheumatoid Arthritis, Health Information Systems

1 Introduction

Rheumatoid arthritis (RA) is the most common autoimmune inflammatory arthritis in adults, affecting almost 1% of the world's population. At present, fewer than 30% of patients show robust responses to treatments. These treatments are associated with a number of adverse side effects, including disease relapse and bone deformation of individual joints[1] [14]. Although there is no cure, early intervention has made RA a less disabling disease and if treatment is instituted right from the onset, no functional impairment may occur and structural integrity may be preserved[2]. The impact of the disease is wide, not only resulting in decreased health- related quality of life, but also a loss of productivity and a major increase in healthcare costs [3] [13].

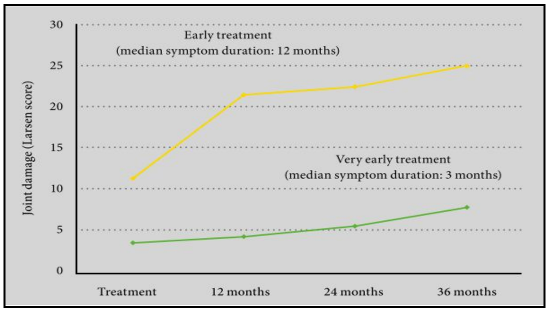


Fig. 2. The importance of starting rheumatoid arthritis therapy very early.[2]

Patients with established rheumatoid arthritis (RA) may incur important resource utilisation and work productivity loss, resulting in high costs of illness. Impairment in physical function, which increases with disease duration, is the main variable driving all aspects of these costs. The large variation of costs across administrations is a complex issue and results not only from differences in access to and provision of care but also from absolute differences in the prices for health-care or loss of paid work [4].

Direct costs increase over proportionally during the course of the disease. The most important driver of direct costs is hospitalization, especially in moderate and severe RA. Costs of medication represent a comparatively small proportion of direct costs. Indirect costs caused by work disability can be substantially higher than direct costs, particularly in working-age patients, The total costs of RA to society, and the different cost components such as direct and indirect costs, are broadly comparable in industrialized countries by their order of magnitude. Major confounding factors for international comparison are different study methodologies and patient samples [5].

There are studies that try to standardize the estimations of costs of RA for the industrialized countries [5], On the other hand, Latin America has undergone an epidemiological shift from acute to chronic disease as the major contributor to morbidity and mortality, while still confronting problems such as poverty and malnutrition. Consequently, these changes impact the allocation of health-care resources from acute diseases, such as infections, to chronic diseases such as RA[6].

Challenges for Latin American countries in the treatment of RA include making RA a public health priority, knowing its socioeconomic impact in terms of its high cost and burden on the health-care system, and increasing access to prompt diagnosis, treatment by rheumatologists, and availability of effective low cost medications [6].

2 State of The Art

Health care data is expected to grow exponentially large in the following years. The goal among researchers and health care providers is to integrate, digitize the value of such kind of *"big data"* in different health care sectors which includes hospital, small nursing homes, independent physician offices and other healthcare organizations[7].

The complexity of dealing with "big data" forces to apply some kind of framework that includes complex data management, algorithms and knowledge discovery tasks. As shown in figure 3, a visualization component is key in such environments.

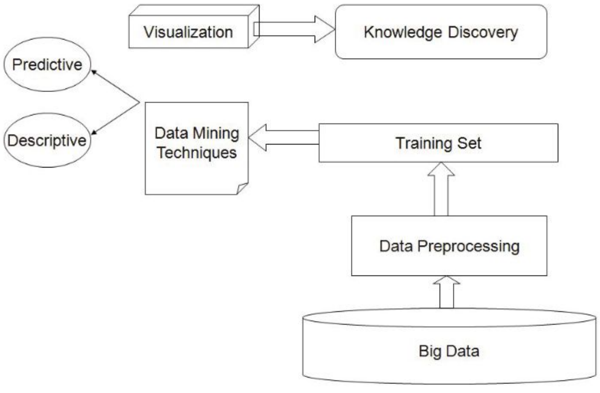


Fig. 3. Framework for big data analytics.[7]

Visualization and visual analytics can contribute to substantial technological advances to support the reliable, effective, safe, and validated systems required in the healthcare domain for personal health, clinical healthcare, and public health policymaking [8] [9].

Studies like [10] use interactive data visualization tools to allow users to deepen the comprehension of the dynamic changes of medical data. A more structured approach is described in [11] where a whole process-driven framework is presented. (See figure 4). That methodology states that visual analytics normally commences with a pre-determined task – then goes through an iterative process to get the required data, choose appropriate visual structure (e.g. chart/graph), view the data, formulate insight and then act. This process involves users moving around between different steps as new data insights (and new questions) are revealed.

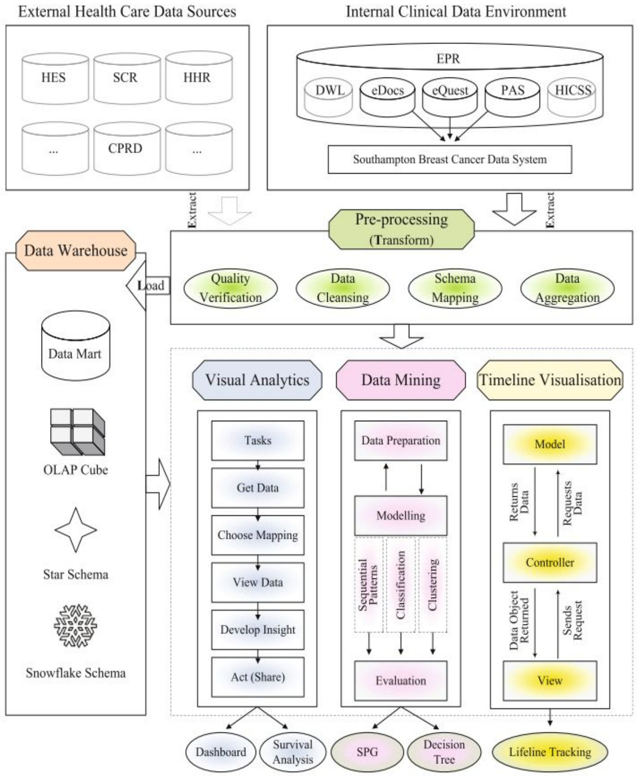


Fig. 4. Process-driven framework for health data analytics.[11]

In such context, dashboards are typically used as a means of displaying live data and each dashboard is a collection of individual indicators, designed in such a way that their significance can be grasped quickly [11] .

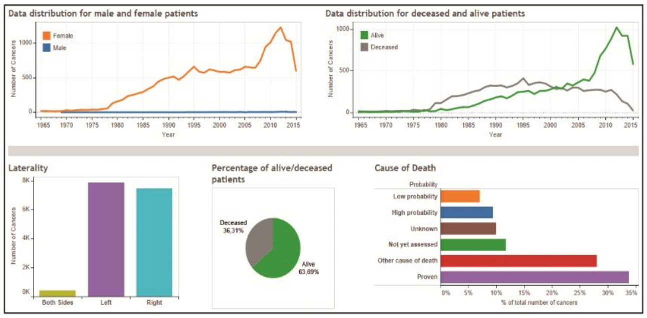


Fig. 5. SBCDS Dashboard example [11].

3 A Dashboard that helps to analyze the costs of Rheumatoid Arthritis (RA) in Colombia.

In order to address the complexity of RA data, an interactive dashboard is proposed. Such visual analytics tools is intended to allow physicians and health professionals to understand the costs associated with RA in Colombia.

3.1 Characterization

The characterization of the information visualization system is based on the framework of Tamara Munzner [12] which states that visualization usage can be analyzed in terms of ***who*** the user is, ***why*** the user needs it, ***what*** data is shown, and ***how*** the idiom is designed [12].

3.2 Who

The visualization is intended for Physicians and Health professional interested in occurrence of Rheumatoid arthritis (RA) in Colombia.

3.3 What

Main Dataset: SISPRO

Description: Administrative Database with Medical services given to patients in Colombian health system, filtered by Diagnostic codes for Rheumatoid Arthritis.

Source: SISPRO (*Sistema Integral de Información de la Protección Social*) [15]. It has been designed as a Data Warehouse. Includes information about Health (epidemiological and demographic data, drug unit costs, health service use), social protection, social promotion, occupational risks. Data is provided by external and internal sources at the Ministry of Health.

Source Type: Microsoft analysis services data cube Cube: CU - Prestación Servicios de Salud

Dataset Type: Table, Temporal

Attributes

- States : Categorical. States of Colombia
- Year: Categorical, ordered, sequential. Year of the procedure.
- Regime : Categorical. Type of health regime to which the patient belongs
- Sisben : Categorical. Subtype of Subsidized regime.
- Sex : Categorical.
- Age Group: Categorical, ordered. Age Groups classified by human cycle.
- Administrator: Categorical. Administrators of the Social Security System
- Provider: Categorical. Company that provides a medical service.
- Procedure: Categorical. Procedures and medical services performed in Colombia
- Procedure Cost: Quantitative, ordered, sequential. Cost of a procedure applied to a patient.
- People Served : Quantitative, ordered, sequential. Number of people

Derived Data: Categorical

Most categorical attributes from raw data are long strings that repeat many times. The size of original CSV file is 456 MB. Such kind of "big" file can generate a lot of latency during downloads for "normal" web clients. To avoid that kind of problem, data is derived in two files:

- A lookup table (domains.json - 1 MB)
- Encoded rows (costs.tsv - 20.4 MB)

Encoding was made as follows:

- Extract ids / codes from original strings for States, Administrators, providers and procedures
- Generate ids for regime, sisben, sex and age.

Derived Data: Geographic

The geo data of the States used for the map (colombia_index.geojson) is a simplification of the original polygons from OSM [16]. The derived file tries to reflect a "*Grid Map*"[17] for Colombia that allows the user to easily identify a State for interactive widgets based on the bounding boxes of the 25k scale grid of Colombia. The grid map was made with Postgis and QGIS.

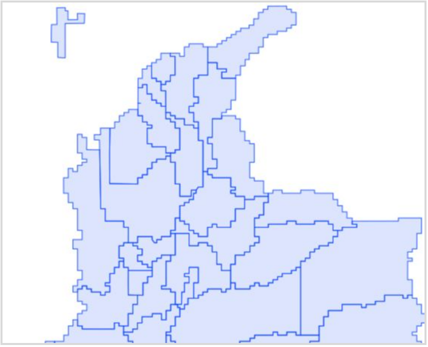


Fig. 6. A proposed Grid Map for Colombia.

3.4 Why

Main Task:

- ***Discover*** the ***distribution*** of Costs per Person (CP) of RA procedures in Colombia by state, year, regime, sisben, sex, age group, administrator and provider.

Secondary Tasks:

- ***Derive*** attributes from raw data as ***features*** to be used in the final visualization.
- ***Identify Outliers*** in costs.
- ***Identify*** the ***Features*** of a specific procedure in the dataset.
- ***Summarize*** the ***distribution*** of Costs per Person, people served and total costs of RA procedures in Colombia.

3.5 How

The dashboard uses different idioms / widgets with Different encodings for all data with Linked filtering (Crossfiltering).

3.5.1 Idiom : Horizontal / Vertical Bar Charts

Encode:

- Attributes: Year, State, Regime, Sisben, sex, age, administrator, provider
- Mark: Line
- Channel
- Position: Key attribute. Horizontal / Vertical.
- Color: Selection / Hover
- Encode: Separate, Order, Align.

Manipulate

- Select and Highlight: Click / Hover
- Navigate: Attribute Reduction, Slice
- Change with Animated Transitions

Facet

- Juxtapose
- Linked Filtering (Crossfiltering)

Reduce

- Filter Items / Attributes
- Aggregate Attributes

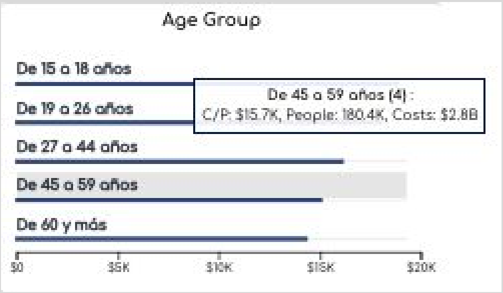


Fig. 7. Horizontal Bar Chart

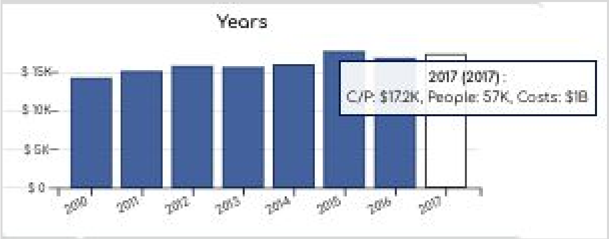


Fig. 8. Vertical Bar Chart

3.5.2 Idiom : List

Encode:

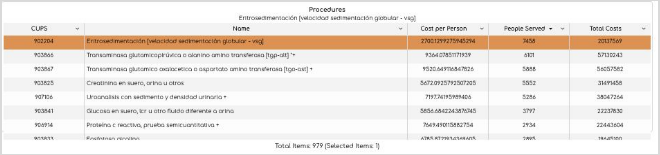
- Attributes: Procedure Name, Cups, Cost per Person, Persons attended, total costs. Total costs per person, total persons attended, total costs.
- Mark: Area
- Channel
- Position: Vertical, Key attribute (Cups). Horizontal: other attributes
- Color: Selection
- Encode: Separate, Order

Manipulate:

- Select
- Reorder

Facet

- Juxtapose
- Linked Filtering (Crossfiltering)



Procedures
Eritrosedimentación [velocidad sedimentación globular - vsg]

CUPS	Name	Cost per Person	People Served	Total Costs
902204	Eritrosedimentación [velocidad sedimentación globular - vsg]	2700.1299275945294	7458	20137569
903866	Transaminasa glutamicopirúvica o alanino amino transferasa [tgp-alt] *+	9364.07851171939	6101	57130243
903867	Transaminasa glutamico oxalacetica o aspartato amino transferasa [tgo-ast] +	9520.649116847826	5888	56057582
903825	Creatinina en suero, orina u otros	5672.0925792507205	5552	31491458
907106	Uroanalisis con sedimento y densidad urinaria +	7197.74195989406	5286	38047264
903841	Glucosa en suero, lcr u otro fluido diferente a orina	5856.6842243876745	3797	22237830
906914	Proteina c reactiva, prueba semicuantitativa +	7649.490115882754	2934	22443604
903833	Fosfatasa alcalina	6785.87219343469605	2895	19645100

Total Items: 979 (Selected Items: 1)

Fig. 9. List

3.5.3 Idiom : Grid Map

Encode

- Attributes: State
- Mark: Area
- Channel
- Spatial Region
- Color: Selection

Manipulate

- Select

Facet

- Juxtapose
- Linked Filtering (Crossfiltering)

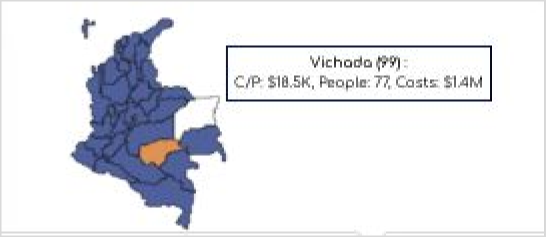


Fig. 10. Grid Map

4 Data Extraction and Validation

Data source is available as a datacube implemented with Microsoft Analysis Services. In order to extract data for visualizations the following process was made:

- Create a Tableau [18] workbook on a windows machine
- Connect the datasource to the datacube on Microsoft analysis services.
- Create a Worksheet with the filters of interest and add dimensions (rows) and measures (columns)
- Export data as CSV using "Analysis / View Data / Export All"
- Process csv (ie. filtering, cleaning) using Python / Jupyter Lab
- Save results as tsv, csv or json

The original extracted file had 660161 rows and 22 columns. 8 of those columns where removed after validating the first functional validation of the application. The rows without costs values also were removed. The final dataset had 343413 rows. The file was processed using Pandas [19] and Jupyter Lab [20].

Some really weird data points were found in costs data. See Fig. 11. Those rows were removed using the "interquartile range" (IQR) technique [21]. See Fig. 12.

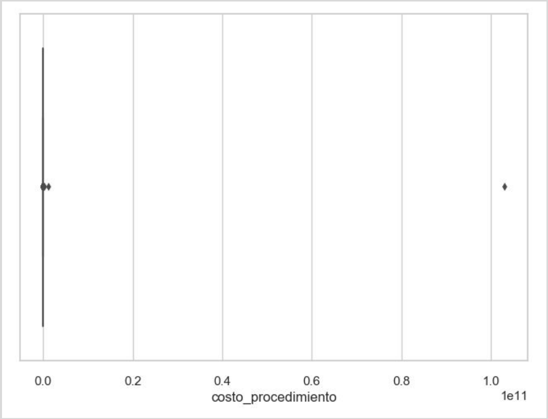


Fig. 11. Boxplot of costo_procedimiento

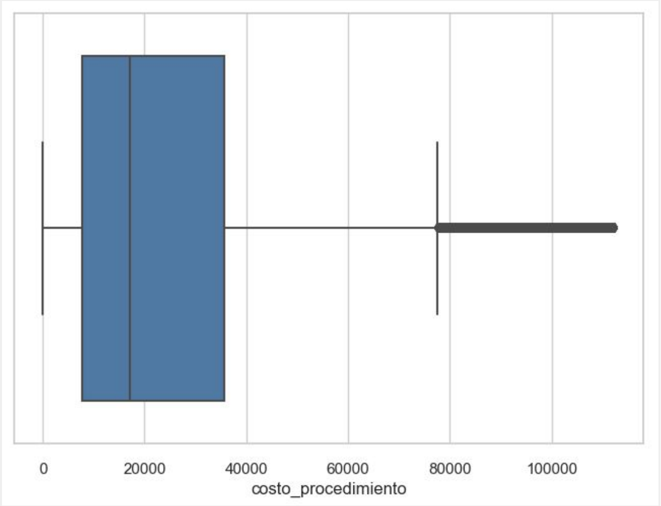


Fig. 12. Boxplot of costo_procedimiento after outliers removal

5 Process

The project had the following phases:

- **Proposal and approach of the project:** The initial interview with the expert help to define the main task of the visualization. A first mockup was made using draw.io [22]. See Fig 13.

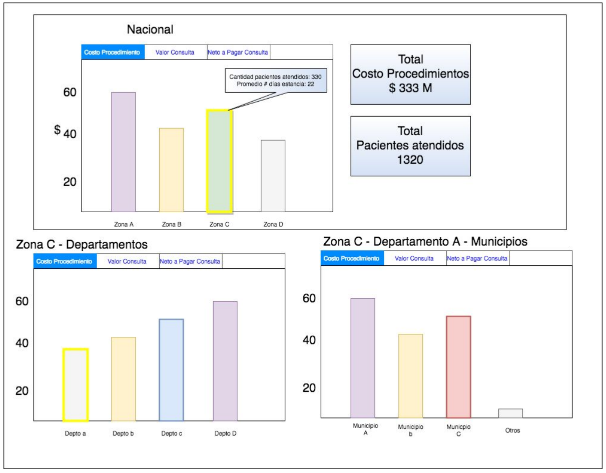


Fig. 13. Initial mockup

- **Progress report:** A new mockup was created after initial validation with the experts. See Fig 14.

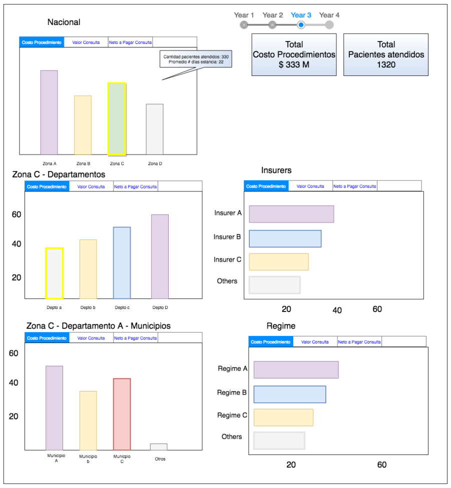


Fig. 14. Second mockup

- **Usability Study:** Using the second mockup a usability study was made with users. As a result, a third mockup was made. See Fig. 15.

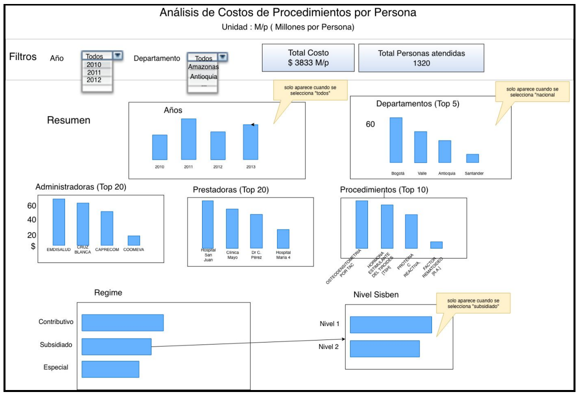


Fig. 15. Third mockup

- **Functional Prototype:** A functional prototype was made including findings of the progress report phase. . See Fig 16.

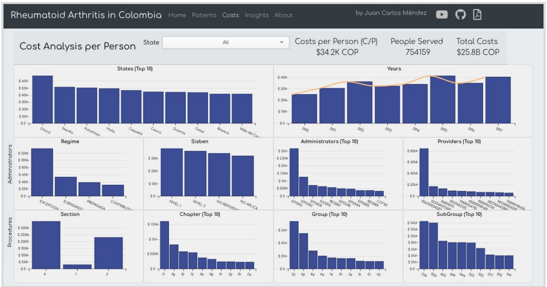


Fig. 16. Functional Prototype

- **Final Implementation:** The final implementation was made having into account the feedback of the expert after using the functional prototype. Visualization was also improved looking for optimization of available space, the integration of animated transitions on data change, adding of the grid map and list widgets. See Fig 17.

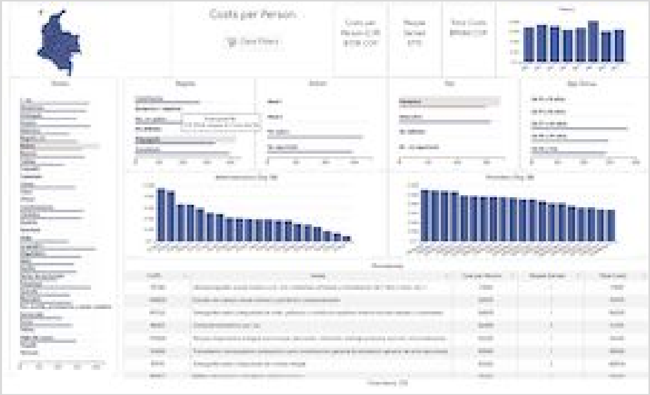


Fig. 17. Final implementation.

6 Evaluation

After the Progress Report phase a usability study was made. To visualize the results of the study, two *Likert Scale 5* charts were created using a *Centered stacked bar chart*. The first one including neutral category (See Fig 18) and the second one excluding "neutral" as suggested in [23]. See Fig 19.

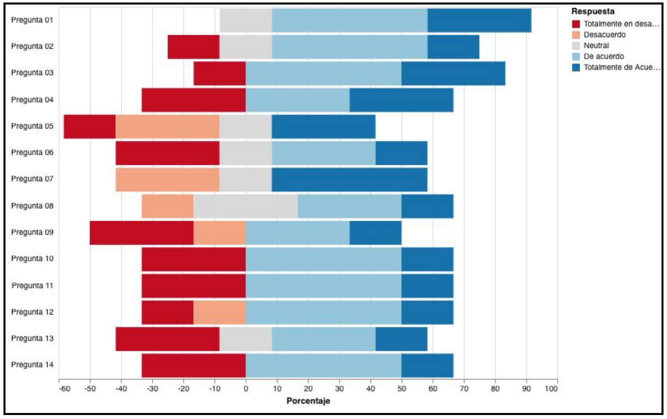


Fig. 18. Results of usability study using a Likert Scale 5 including "*neutral*" category.

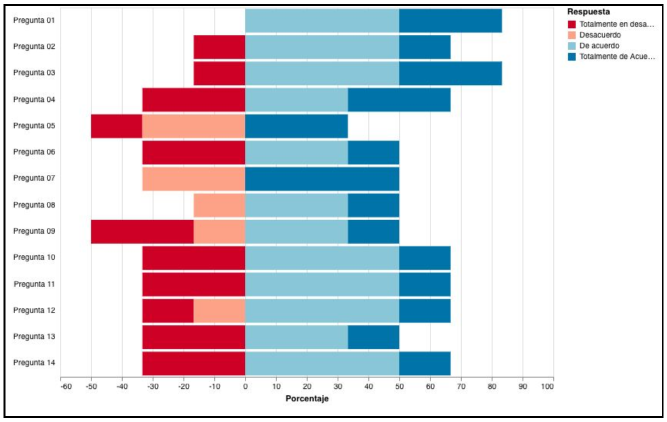


Fig. 19. Results of usability study using a Likert Scale 5 excluding "*neutral*" category.

7 Results

The implemented interactive dashboard allowed the user to develop the proposed tasks using a relatively *"big"* dataset. The latency of interaction of the application had low latency. The "*Horizontal Bar chart*" implemented in the final iteration improved the readability of labels and the identification of change using animated transitions.

8 Insights

There are general data quality problems like these:

- Some states do not have data for one or more years (e.g Amazonas, Arauca, Casanare, Guainía, Guaviare)
- There are "Not Reported - NR" and "Not Available - NA", values in some of the attributes. Such kind of "data loss" problem should be mitigated by data publishers in order to improve the general data quality of the dataset.
- The expert found different NIT providers with the same name.
- There are anomalies in procedure costs that reflect problems during data collection (e.g. procedures with cost of $1 COP)

The domain experts also found the following:

- The overall Costs per Person (C/P) for procedures is higher for younger people.
- The overall C/P is higher for well known [24] isolated states like Chocó, Guainía, La Guajira , Putumayo , Arauca, Vaupés and Vichada.
- For every year of the dataset Córdoba is the State with the highest C/P.
- According to the expert, the costs of some procedures in the database differ a lot from the official rates established for the country. (SOAT DECRETO 2423 DE 1996] and its yearly updates)
- There is a huge number of persons (65k) with RA that belong to the subsidized (subsidiado) regime. According to the expert, usually that group of people doesn't have access to the most advanced or modern procedures due to the high costs of them.
- There are more women affected with the decease, but costs for men are higher.

9 Technologies

The following technologies were used to develop the dashboard:

- D3 v5.7.0 https://github.com/d3/d3/releases/tag/v5.7.0
- Crossfilter2 v1.4.6 https://github.com/crossfilter/crossfilter
- AngularJS 1.6.6 https://angularjs.org/
- Angular UI Grid 4.6.6 http://ui-grid.info/
- Bootstrap 4.1.3 http://getbootstrap.com/
- Google Fonts
- Font Awesome

The final source code of the application can be found in the following url: https://github.com/dersteppenwolf/isis4822_final_project

10 Conclusions

The visual analytics approach used in this work improves the workflow between final users and visualization experts.

The data of SISPRO should be used with care due to the quality problems of the data found during the ETL processes and later when the users interacted with the dashboard.
Health Information Systems in Colombia need to radically improve in order to achieve a more efficient and effective service to citizens.

11 Acknowledgments

The author wish to thank to Adriana Beltrán Ostos M.D. MsC. and Sergio Mauricio Moreno López Statician. MsC. Hospital militar central as main customers of the project. Paula Andrea Méndez M.D. and Viviana Ariza PhD in Biotechnology for their feedback after dashboard exploration. Jofre Manchola and Jairo Ruiz teaching assistants, and at last but not least to John Alexis Guerra Gómez professor of the class *ISIS 4822: Visual Analytics [25]* at Universidad de los Andes de Colombia.

12 References

[1] Mason, V., King-, U., & Diseases, S. (2018). Rheumatoid Arthritis. (S. Liu, Ed.) (Vol. 1868). New York, NY: Springer New York. http://doi.org/10.1007/978-1-4939-8802-0

[2] Smolen, J. S. (2015). Atlas of Rheumatoid Arthritis. Atlas of Rheumatoid Arthritis. http://doi.org/10.1007/978-1-907673-91-7

[3] van Vollenhoven, R. F. (2016). Biologics for the Treatment of Rheumatoid Arthritis. FEMS Microbiology Letters (Vol. 66). Cham: Springer International Publishing. http://doi.org/10.1007/978-3-319-13108-5

[4] Fautrel, B., Verstappen, S. M. M., & Boonen, A. (2011). Economic consequences and potential benefits. Best Practice and Research: Clinical Rheumatology, 25(4), 607–624. http://doi.org/10.1016/j.berh.2011.10.001

[5] Pugner, K. M., Scott, D. I., Holmes, J. W., & Hieke, K. (2000). The costs of rheumatoid arthritis: An international long-term view. Seminars in Arthritis and Rheumatism, 29(5), 305–320. http://doi.org/10.1016/S0049-0172(00)80017-7

[6] Burgos-vargas, R., Jose, L., Galarza-maldonado, C., Ostojich, K., & Cardiel, M. H. (2013). Current therapies in rheumatoid arthritis : A Latin American perspective. Reumatología Clínica (English Edition), 9(2), 106–112. http://doi.org/10.1016/j.reumae.2013.01.007

[7] Chauhan, R., & Kaur, H. (2015). Computational Intelligence for Big Data Analysis, 19, 165–179. http://doi.org/10.1007/978-3-319-16598-1

[8] Shneiderman, B., Plaisant, C., & Hesse, B. W. (2013). Improving healthcare with interactive visualization. Computer, 46(5), 58–66. http://doi.org/10.1109/MC.2013.38

[9] Gotz, D., & Borland, D. (2016). Data-Driven Healthcare: Challenges and Opportunities for Interactive Visualization. IEEE Computer Graphics and Applications, 36(3), 90–96. http://doi.org/10.1109/MCG.2016.59

[10] Yamada, S., Yamamoto, Y., Umezawa, K., Inokuchi, S., Miyachi, H., Hashimoto, M., & Asai, S. (2017). Exploratory data analysis for medical data using interactive data visualization. International Conference on ICT and Knowledge Engineering, 7–11. http://doi.org/10.1109/ICTKE.2016.7804091

[11] Lu, J., Hales, A., & Rew, D. (2017). Modelling of Cancer Patient Records: A Structured Approach to Data Mining and Visual Analytics (Vol. 8060, pp. 30–51). http://doi.org/10.1007/978-3-319-64265-9_4

[12] Munzner, T. (2015). Visualization analysis and design. Boca Raton: CRC Press, Taylor & Francis Group, CRC Press is an imprint of the Taylor & Francis Group, an informa business.

[13] Healthline. (2018). Rheumatoid Arthritis: Huge Cost to Patients & Economy. [online] Available at: https://www.healthline.com/health-news/rheumatoid-arthritis-heavy-cost-to-patients-economy#5 [Accessed 6 Dec. 2018]

[14] Mayo Clinic. (2018). Rheumatoid arthritis - Symptoms and causes. [online] Available at: https://www.mayoclinic.org/diseases-conditions/rheumatoid-arthritis/symptoms-causes/syc-20353648 [Accessed 6 Dec. 2018].

[15] Sispro.gov.co. (2018). [online] Available at: http://www.sispro.gov.co/ [Accessed 6 Dec. 2018].

[16] OpenStreetMap. (2018). OpenStreetMap. [online] Available at: https://www.openstreetmap.org/ [Accessed 6 Dec. 2018].

[17] Forumone.com. (2018). Good Data Visualization Practice: Tile Grid Maps | Forum One. [online] Available at: https://forumone.com/ideas/good-data-visualization-practice-tile-grid-maps-0 [Accessed 6 Dec. 2018].

[18] Tableau Software. (2018). Tableau: Business Intelligence and Analytics Software. [online] Available at: https://www.tableau.com/ [Accessed 6 Dec. 2018].

[19] Pandas.pydata.org. (2018). Python Data Analysis Library — pandas: Python Data Analysis Library. [online] Available at: https://pandas.pydata.org/ [Accessed 6 Dec. 2018].

[20] GitHub. (2018). jupyterlab/jupyterlab. [online] Available at: https://github.com/jupyterlab/jupyterlab [Accessed 6 Dec. 2018].

[21] Stapel, E. (2018). Interquartile Ranges (IQRs) & Outliers | Purplemath. [online] Purplemath. Available at: https://www.purplemath.com/modules/boxwhisk3.htm [Accessed 6 Dec. 2018].

[22] Draw.io. (2018). Flowchart Maker & Online Diagram Software. [online] Available at: https://www.draw.io/ [Accessed 6 Dec. 2018].

[23] Petrillo, F., Spritzer, A. S., Freitas, C. D. S., & Pimenta, M. (2011). Interactive analysis of Likert scale data using a multichart visualization tool. Proc. IHC+CLIHC '11, (July 2015), 358–365. Retrieved from http://dl.acm.org/citation.cfm?id=2254436.2254496

[24] Article (2018). Una región desconectada | ELESPECTADOR.COM. [online] ELESPECTADOR.COM. Available at: https://www.elespectador.com/noticias/medio-ambiente/una-region-desconectada-articulo-449013 [Accessed 6 Dec. 2018].

[25] Johnguerra.co. (2018). ISIS 4822: Visual Analytics. [online] Available at: http://johnguerra.co/classes/visual_analytics_fall_2018/ [Accessed 6 Dec. 2018].