Prediction of Public Health System Coverage for Senior Adults in Chile using Machine Learning Tools

FOUNDATIONS FOR A MONITORING PLATFORM FOR PUBLIC HEALTH SYSTEM COVERAGE IN CHILE

Víctor Hernández M*, Rodrigo Pérez†, Fernando Henríquez‡, Paulina Arriagada§, Pedro Zitko ¶, Andrea Slachevsky ‖ and Juan D. Velásquez **

Web Intelligence Centre. Department of Industrial Engineering, University of Chile.* † **

Neuropsychology and Clinical Neuroscience Laboratory (LANNEC), Physiopathology Department - ICBM, Neurocience and East Neuroscience Departments, Faculty of Medicine, University of Chile.‡ ‖

Neurology Unit, Regional Hospital Coyhaique, Aysén Health Service.§

Health Service & Population Research, IoPNN. King's College London.¶

Email: victor.hernandez@wic.uchile.cl*, rodrperez@ing.uchile.cl†, fehech@gmail.com‡, arriagadapau@saludaysen.cl§, pedrozitko@gmail.com¶, andrea.slachevsky@uchile.cl‖, jvelasqu@dii.uchile.cl**

ABSTRACT

Chile is an aging country, so monitoring the health coverage of the public system towards its elderly is of interest. We propose a novel approach to eventually predict the level of health coverage that an elderly person could receive, based on a profile based on sociodemographic and health variables, through the use of Machine Learning. Data is collected through a specialized survey and tests are carried out through different supervised classification algorithms. Our results suggest that this approach could complement the information that will be available to the decision makers of the public health system, nevertheless, the collection of data required to be improved and our approach need to be evaluated in future surveys.

I. INTRODUCTION

Chile is a country that has aged rapidly over the last decade, from having 15% of the total population over 60 years in 2009, to 19.3% in 2017. In addition, that same year, 84.9% of the population over 60 years of age was part of the public health system (FONASA). On the other hand, there are health problems of high prevalence for these age segments, such as dementia disorders or neurocognitive disorders, which are an influential factor in the inequality of the population, representing high monetary and social costs in the family of those affected (Ministerio de Desarrollo Social, 2017). This background shows that it is important to have efficient methods to monitor the health coverage of senior adults in Chile, to support decision-making and an efficient allocation of limited resources (Hojman *et al.*, 2015).

Currently, the data that exists in Chile regarding these factors is static and scarce. The closest to the collection of data of this nature is the National Health Survey (Margozzini & Passii, 2018), carried out periodically and on a limited sample of people. The last edition was carried out in the period 2016-2017 covering 6233 people. It is worth mentioning that the processing of its results is a slow process, carried out by expert personnel, which leads to a report. In other words, static results are obtained for each edition of the survey.

Here, we provide an approach to collect information in an expeditious manner, which allows us to monitor the public health system coverage for older adults in Chile. The main idea is to make predictions of the potential health coverage for a elderly, using Data Mining and Machine Learning tools, to generate information that supports the decision maker. This is a work developed jointly between the Web Intelligence Centre of the Faculty of Engineering and the Faculty of Medicine of the University of Chile.

II. METHODOLOGY

A. Data collection

In order to gather the patients information, we designed a survey composed of two sections for each health problem: a module that evaluates if the person has a health care need (normative need) (Bradshaw, 1972), and a module that assesses the health system coverage. The latter, considers the primary care level (care centers), the secondary care level (specialized clinics) and the multiple treatments that could be indicated for a given health care need. Along with this, there are attached instruments that assess the socioeconomic status of the respondents, their level of accessibility to healthcare centers, their occupational career, their level of literacy in health, among others.

The survey includes 15 health modules, associated with problems prevalent in the elderly population. Some of them are: cognitive, affective, cardiometabolic and musculoskeletal. To obtain the normative need, a standardized instrument is used for each corresponding health module. To model the health system coverage, an adaptation of the Tanahashi coverage model (Tanahashi, 1978) has been used. The selected population is located in the city of Coyhaique (Aysén region, southern Chile).

The survey was carried out with the help of a digital platform developed by the Center of Geroscience, Mental Health and Metabolism (Gero)[1] and trained interviewers. 2754 addresses were visited in a period of two months, from which 440 subjects were found eligible for the study, and 137 agreed to participate. 81 surveys were answered in their entirety, beign the length of the survey and the inability of some people to respond due to their health status, between the main causes of uncompleted surveys.

B. Data preprocessing

Once the 81 completed surveys were identified, a preprocessing is performed based on the cleaning of null values, conversion of numerical values with respect to formats and units of measurement, and semi-automatic processing of unstructured text is included in some answers (e.g., how long does it take to get to a health care center from your home?).

The majority of the survey consists of categorical variables, so in addition, these questions were preprocessed as binary variables (or dummy variables) to facilitate their interpretation by the classifier.

C. Data transformation

This stage consists of two steps. First, for each health module the calculation of the normative need is made, to establish whether the person has the problem or not, which is represented with a binary variable. For this analysis, we evaluate the rules for each module, which were defined based on the standards associated with each instrument. For example, cognitive health module uses Montreal Cognitive Assessment (Nasreddine & Phillips, 2005), Pfeffer Functional Activities Questionnaire (Fuentes-García, 2014) and AD8 Informant Questionnaire (Galvin *et al.*, 2005; Muñoz *et al.*, 2010) instruments together to determine the normative need.

The second step computes the health system coverage for each of the treatments associated with the module and seeks to capture whether at each level of care the patient was diagnosed and received the right treatment. It also captures the patient's response to the treatment and the reasons of treatment abandonment. The transformation interprets the response flow and links it with the level of health system coverage, which is saved in a vector of categorical variables. It is worth mentioning that for most health modules, if the person does not have the health problem, the person immediately falls into the first level of health coverage, implying that the person did not require the service. In those health modules where the normative need may be affected by an ongoing treatment, the flow of responses is analyzed anyway.

[1] http://www.gerochile.cl/es/

D. Classification Process

Up to this point there are three types of data: the variables associated with all the answers given by the respondents, the computed normative needs and the level of health system coverage for each treatment, for each health module of the survey.

Because we aim to predict the health system coverage based on a patient's profile, a supervised classification process is applied, where the dependent variable corresponds to the treatment coverage levels found, and the independent variables are the computed normative needs, together with the variables that allowed us to determine that value. In addition, the sociodemographic information, accessibility and occupational trajectory are included within the independent variables to further complement the profile of each person.

Tests were done with Multinomial Naive Bayes, K-Nearest Neighbors and Random Forests algorithms for supervised classification. These were implemented from the library scikit-learn (Pedregosa *et al.*, 2011). Each classification includes a 3-fold cross validation process for the evaluation.

III. Results

Below are the best 5 average results for each classification process.

TABLE I: Multinomial Naive Bayes results

Health module	Precision	Recall	F-measure
Respiratory	0,7802	0,3502	0,4562
Cognitive	0,8099	0,3270	0,4362
Sleep disorder	0,5727	0,3304	0,4064
Vision	0,5168	0,2484	0,2926
Affective	0,4091	0,2338	0,2800

TABLE II: K-Nearest Neighbors results

Health module	Precision	Recall	F-measure
Respiratory	0,8142	0,9020	0,8557
Cognitive	0,7997	0,8924	0,8426
Sleep disorder	0,7027	0,7348	0,6590
Diabetes	0,4579	0,4658	0,4503
Overweight	0,4286	0,4878	0,4458

Random Forests results:

TABLE III: Random Forests results

Health module	Precision	Recall	F-measure
Cognitive	0,8368	0,9042	0,8659
Respiratory	0,8142	0,9020	0,8557
Sleep disorder	0,5388	0,7052	0,6094
Vision	0,4731	0,6329	0,5167
Overweight	0,4865	0,6321	0,5134

IV. Discussion

The results show that the best performance is obtained with the Random Forests algorithm. However, the process should continue to improve given that performance is only high for some health modules. The survey turned out to be an impractical experience for the respondents, so an application outside the experimental environment would be unfeasible. The next step is to identify, through the results obtained, which are the health modules and annexed instruments that add more information, to preserve only those relevant questions in a future instance of data collection. This will allow a greater representativeness of the population to be achieved, through a more complete data set and a larger scale. However, this work reflects a novel contribution to complement the information available to decision makers of the public health system in Chile and that laid a basis for directing the availability of this information towards a dynamic approach, fed by various data sources.

References

Bradshaw, Jonathan. 1972. Taxonomy of social need.

Fuentes-García, Alejandra. 2014. *Pfeffer Functional Activities Questionnaire*. Dordrecht: Springer Netherlands.

Galvin, JE, Roe, CM, & Powlishta. 2005. The AD8: a brief informant interview to detect dementia. *Neurology*, **65**(4), 559–564.

Hojman, D, Duarte, F, Ruiz-Tagle, J, Nuñez-Huasaf, J, Budinich, M, & Slachevsky, A. 2015. The cost of dementia: The case of chile. Results of the cuideme study. *Journal of the Neurological Sciences*, **357**, e11.

Margozzini, P., & Passii, A. 2018. Encuesta Nacional de Salud, ENS 2016-2017: Un aporte a la planificación sanitaria y políticas públicas de Chile. *Ars Medica Revista de Ciencias Médicas*.

Ministerio de Desarrollo Social, Chile. 2017. *Encuesta CASEN 2017 Adultos Mayores - Sintesis de Resultados*.

Muñoz, Carlos, Nunez, Javier, Flores, Patricia, Behrens, P MI, & Slachevsky, Andrea. 2010. Usefulness of a brief informant interview to detect dementia, translated into Spanish (AD8-Ch). *Revista medica de Chile*, **138**(8), 1063.

Nasreddine, Ziad S, & Phillips. 2005. The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. *Journal of the American Geriatrics Society*.

Pedregosa, Fabian, Varoquaux, Gaël, Gramfort, Alexandre, & Michel. 2011. Scikit-learn: Machine learning in Python. *Journal of machine learning research*.

Tanahashi, T. 1978. Health service coverage and its evaluation. *Bulletin of the World Health Organization*, **56**(2), 295–303.