PLOS ONE

RESEARCH ARTICLE

Feature sensitivity criterion-based sampling strategy from the Optimization based on Phylogram Analysis (Fs-OPA) and Cox regression applied to mental disorder datasets

Fatemeh Gholi Zadeh Kharrat[1]*, Newton Shydeo Brandão Miyoshi[2], Juliana Cobre[3], João Mazzoncini De Azevedo-Marques[4], Paulo Mazzoncini de Azevedo-Marques[5], Alexandre Cláudio Botazzo Delbem[1,6]

1 Department of Bioengineering, Universidade de Sao Paulo Escola de Engenharia de Sao Carlos, Sao Carlos, Sao Paulo, Brazil, **2** Center of Information and Informatics of Medical School, Ribeirao Preto, Universidade de Sao Paulo Escola de Enfermagem de Ribeirao Preto, Sao Paulo, Brazil, **3** Department of Mathematics and Statistics, Universidade de Sao Paulo Instituto de Ciencias Matematicas e de Computacao, Sao Carlos, Sao Paulo, Brazil, **4** Department of Social Medicine of Medical School, Universidade de Sao Paulo Campus de Ribeirao Preto, Ribeirao Preto, Sao Paulo, Brazil, **5** Department of Medical Imaging, Hematology and Clinical Oncology of Medical School, Universidade de Sao Paulo Campus de Ribeirao Preto, Ribeirao Preto, Sao Paulo, Brazil, **6** Department of Computer Science, Universidade de Sao Paulo Instituto de Ciencias Matematicas e de Computacao, Sao Carlos, Sao Paulo, Brazil

* f.gholizadeh@usp.br

Citation: Gholi Zadeh Kharrat F, Shydeo Brandão Miyoshi N, Cobre J, Mazzoncini De Azevedo-Marques J, Mazzoncini de Azevedo-Marques P, Cláudio Botazzo Delbem A (2020) Feature sensitivity criterion-based sampling strategy from the Optimization based on Phylogram Analysis (Fs-OPA) and Cox regression applied to mental disorder datasets. PLoS ONE 15(7): e0235147. https://doi.org/10.1371/journal.pone.0235147

Editor: Zezhi Li, National Institutes of Health, UNITED STATES

Received: January 17, 2020

Accepted: June 9, 2020

Published: July 1, 2020

Data Availability Statement: Data contain potentially identifying, or sensitive patient information and health Information and Informatics Center (CIIS) at Ribeirão Preto Medical School - University of São Paulo is the third party who owns the data underlying our study. The CIIS team member who will be responsible for keeping the data and responding to interested researchers in obtaining access to it will be Lariza Laura de

Abstract

Digital datasets in several health care facilities, as hospitals and prehospital services, accumulated data from thousands of patients for more than a decade. In general, there is no local team with enough experts with the required different skills capable of analyzing them in entirety. The integration of those abilities usually demands a relatively long-period and is cost. Considering that scenario, this paper proposes a new Feature Sensitivity technique that can automatically deal with a large dataset. It uses a criterion-based sampling strategy from the Optimization based on Phylogram Analysis. Called FS-opa, the new approach seems proper for dealing with any types of raw data from health centers and manipulate their entire datasets. Besides, FS-opa can find the principal features for the construction of inference models without depending on expert knowledge of the problem domain. The selected features can be combined with usual statistical or machine learning methods to perform predictions. The new method can mine entire datasets from scratch. FS-opa was evaluated using a relatively large dataset from electronic health records of mental disorder prehospital services in Brazil. Cox's approach was integrated to FS-opa to generate survival analysis models related to the length of stay (LOS) in hospitals, assuming that it is a relevant aspect that can benefit estimates of the efficiency of hospitals and the quality of patient treatments. Since FS-opa can work with raw datasets, no knowledge from the problem domain was used to obtain the preliminary prediction models found. Results show that FS-opa succeeded in performing a feature sensitivity analysis using only the raw data available. In this way, FS-opa can find the principal features without bias of an inference model, since the

Oliveira, BSc in Biomedical Informatics, Master, and Ph.D. in Bioinformatics. Her email is larizalaura@gmail.com. Interested researchers can replicate our study findings in their entirety by directly obtaining the data from the CIIS and following the protocol in our Methods section. The authors did not have any special access privileges that others would not have.

Funding: Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

proposed method does not use it. Moreover, the experiments show that FS-opa can provide models with a useful trade-off according to their representativeness and parsimony. It can benefit further analyses by experts since they can focus on aspects that benefit problem modeling.

Introduction

In the last decades, thousands of patients have their data storage into digital datasets. Nevertheless, those datasets have not been analyzed in a way that all the possible relationships among all the data are verified. In general, it would require several experts in different types of knowledge working in an integrated way. The lack of available professionals for such work usually involves the problem domain (e.g., health, energy, finance, agribusiness, etc.) and/or data science areas (statistics, artificial intelligence, optimization, high-performance computing, etc.). Emphasis on innovations to deal with large-scale dataset has increased recently, in some forms called BigData [1], it has motivated the development of new computing methods. In the healthcare domain, this challenge is even greater, since it may involve a lot of distinct knowledge from experts, making hard to develop automated analysis.

On the other hand, *Estimation of Distribution Algorithms (EDAs)* composes an area of investigation of optimization strategies that aim at learning a problem from scratch to create models for search space exploration [2] Although the efficacy verified for some of them, they showed not properly for several large-scale real-world problems, in fact, the majority of the success cases are related to some problem categories. Recently, a new *EDA*, called *Optimization based on Phylogram Analysis) OPA (*[3] overcame those drawbacks, obtaining relevant results for different data types, as well as mixed data types. Moreover, *OPA* scalability was proved for binary problems [4] and it has been experimentally verified for some real-world problems.

Such new scenario open opportunities to extend those results for data mining of complex datasets. This paper presents an *OPA* extension to construct a problem model from raw data. The first version of the model consists in selecting the main variables of features from the dataset. It means to find a set of consistent variables, those with enough information, but without repetition, among other aspects. Note that variable correlation (related to common information among variables) is one of the main aspects learned by *EDAs* to construct probabilistic models.

The challenges in this context of healthcare datasets can also involve other data aspects, beyond size and their data meaning. Usually the factors used for performing analysis by experts motivated the procedures for data acquisition, the data structure of the storage, among other aspects that make them biased by demands that were relevant (a) decade(s) ago. The dynamics of population and technology can make them asynchronous to current demands, which can generate data inconsistency. Such an issue is hard to automate for large datasets. A typical concern is time-space granularity and its consistency with an analysis goal. Another critical aspect of health data is the constrained access to information due to its sensitiveness for patients, which is more complex when considering the whole dataset from large health centers. Fortunately, hospitals and prehospital facilities for mental health disorder in the Ribeirao Preto region in Brazil succeed in generated a public and reliable dataset, as shows the achievements of Barros et al [5] and Miyoshi et al. [6] Living with mental health problems are a serious personal and social challenge that has a profound impact, not only on patients, their family, and the health services but also on the economy, which is given that mental health disorders

are the most costly condition in low-income and middle-income countries [7–9] Whereas the recognition of mental disorders has risen over the last decades, the Length of Stay (LOS) in psychiatric hospitals has increased. So, understanding factors associated with the indicator of LOS for mental health disorders not only can be critical to managing the quality of care and economic reasons but also, insurers, administrators and policymakers are interested in the predictors of the length of each hospitalization [10,11]. Some previous studies have determined that different factors tend to effect on the LOS. In general, they found heterogeneous conclusions. For instance, some of them suggested the demographic variables such as age, gender, marital status, type of admission, place of residence and employment status [7,12–15] On the other hand, other aspects as clinical variables, administrative information [16–18], suicide attempts and homelessness [19] are reported as the most important. Until a certain point, the heterogeneous conclusions also illustrate the discorrelation that may take place involving the types of features stored and a purpose of an analysis, which can increase with the dynamics of population and technology.

Such new scenario open opportunities for investigations of new data mining approaches, as the proposed here based on *OPA*. When constructing a problem model from raw data, *OPA* can automate a series of data mining procedures, without the bias for an analysis goal. In fact, *OPA* generates several models that evolve according to general criteria, related to representativeness and parsimony. The first step of modeling consists of selecting the main variables (features) from the dataset. It produces a set of consistent variables, that should be small but with enough information according to the criteria, without redundancy, among other aspects. Note that variable correlation (related to common information among variables) is one of the main aspects learned by *EDAs* to construct probabilistic models.

In order to deal with a variety of data types in the same dataset as well as the absence of domain knowledge for pre-processing data, *DAMICORE* [20] method is used for the model construction. Basically, it finds variable correlations and represents them in a graph tree (a raw model). Based on such results, some feature sensitivity was proposed, composing a new method that we called Feature Sensitivity based on *OPA* (*FS-opa*).

The following objectives can synthesize the investigation we carried on: 1- Development of a method that can directly work with complete raw datasets; 2- Verification of its capabilities using a real-world datasets and the relevance of information extracted from data; 3- A model that can benefit further analysis by experts, and 4- the development of a *FS-opa* procedure to enable the inclusion of problem domain knowledge. Note that selecting the principal features is a way to obtain parsimonious inference models, which can make easier for further analysis by experts as well as it may reduce collinearity and eventually improve the accuracy of predictors.

The remaining of the paper is organized as follows. First presents the main aspects of the dataset used and proposes *FS-opa*. The second shows the experiments with *FS-opa* for the mental disorder dataset. After that presents results and discussion of Cox model, combine with *FS-opa*. Finally, concludes the paper.

Material and methods

The dataset applied and the main computer methods investigated for developing the proposal are arranged in the sequel. Mental disorder dataset section describes some aspects of the mental disorder dataset studied. Estimation of Distribution Algorithms (EDAs) section presents the Optimization based on Phylogram Analysis, which possesses some properties that are relevant for working with raw data in bases of relatively large amount of features.

Mental disorder dataset

The investigation here presented uses the dataset of mental health care collected by the information system of the Coordination of Hospitalizations in Ribeirao Preto, Brazil, from July 2012 to December 2017 [6,21]. The dataset contains information on 8,755 patients with an average age of 37.6 years. The features of the dataset are related to: 1- socio-demographic aspects, 2- information for people admission in a hospital stay (location, duration, patient origin and destination), 3- diagnoses (at admission, primary diagnosis and secondary diagnosis), 4- services used (such as transfer patients to other hospitals), 5- Information on inpatient discharges from hospitals that provide general or specialized care, 6- date information such as (date of registration, date of discharge and date of death, etc.), 7- codes associated to patients, hospitals and hospitalization procedures, 8- mental diagnostic codes (according to the International Classification of Diseases, 10th Revision—ICD-10).

Next we highlight some groups of those disease presented on the dataset: neurotic and anxiety disorders (F40-F49), psychotic disorders including schizophrenia (F20-F29), bipolar disorder (F30, F31), depressive disorders (F32-F34), personality disorders (F60-F69), alcohol-related disorders (F10), other substance-related disorders (F11-F19), and other mental disorders (mainly F00-F09 and F70-79). Moreover, the records include information about other types of diseases, such as a heart problem, trauma, and stroke. All the fifty-two Portuguese and English names of all the features (also denoted here as variables) are shown in S1 Appendix. The corresponding dataset is referred as both 52-feature dataset (52-FD) and raw-and-Full dataset (raw-Full FD), according to the purpose.

Estimation of distribution algorithms (EDAs)

EDAs are optimization methods based on evolutionary theory that automatically constructs a problem model, which is used to search on the decision space. They require relatively few parameters to setup. Basically, an *EDA* uses samples of variable values from promising solutions of a problem to generate a model of variable correlations and probabilities. Then, new values for each variable (a candidate solution) are generated from the model, sampling the decision space. After selecting the best-found solutions among those new generated, a new model can be produced and start a new cycle of the *EAD* processing. (Fig 1) synthesizes the main steps of general *EDAs*.

The main drawbacks of them have being the computing time to construct the models, since high quality models (that can properly represent a problem) usually requires significantly more computation. For example, Bayesian Networks [22] can construct models with high quality for several complex problems, but the corresponding running time can make the complete processing unfeasible for large instances of them (high number of variables, large groups of correlated variables, mixed data type, etc.).

Optimization based on phylogram analysis

OPA [3] is an *EDA* that can guarantee an adequate tradeoff between computing time and the quality of problem models for some complex problems. Models of *OPA* are phylograms (as the

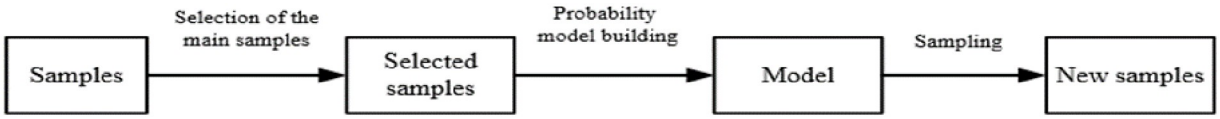


Fig 1. Overview of EDAs.

https://doi.org/10.1371/journal.pone.0235147.g001

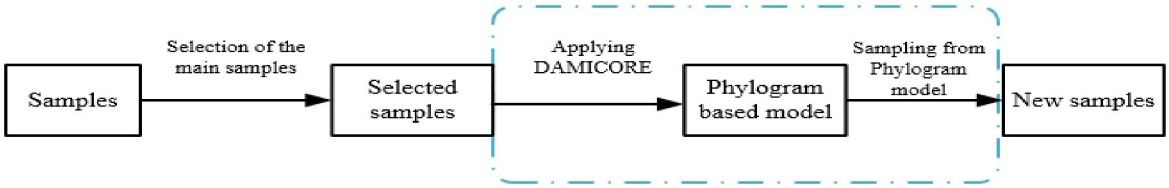


Fig 2. Optimization based on Phylogram Analysis-OPA.

https://doi.org/10.1371/journal.pone.0235147.g002

phylogenetic trees used to describe species evolution) combined with joint probabilities of variables associated to each phylogram subtree (clades). Relatively fast algorithms can construct useful phylograms. *OPA* uses Fast Algorithm of Newman to generate models from large amount of data (that can also be of any type) and a resampling technique to overcome bias from greedy procedures involved in this algorithm and from data samples. Those characteristics make *OPA* proper for dealing with new data sets and problem domains with poor knowledge available (where there is no previous model or expert to orient a modeling process). (Fig 2) shows a flowchart of the *OPA* main steps. The optimization cycle of an *OPA* is like a typical *EDA*, differing mainly by the method of model construction.

Next, we propose an approach, called Feature Sensitivity through criterion-based resampling from OPA (*FS-OPA*), based on *OPA* to identify the principal features (variables) of a problem from scratch. Thus, problem models can be generated for any problem if some significant amount of data is available. This type of result can reduce the time of the investigation until one reaches a useful result or preliminary conclusions, benefiting analysis for different fields, mainly the ones with lack of funds or expert practitioners to carry on the investigations.

A Feature Sensitivity analysis aims at finding a set with the principal features of the problem taking into account both the current context (e.g., sampled data quality and its relevance for a purpose) and the feature interactions [23]), which differs from usual feature selection approaches. *FS-OPA* can perform the feature sensitivity analysis for any type of data and grade of interactions. The results are arranged in a model that highlights feature relationships and relevance (based on phylogram(s)). The main values of variables relevant for the success are represented by joint probabilities, as in *OPA*, conditional probabilities, a regression method, classification approach, etc. *FS-OPA* was hybridized to a Cox approach to enable survival analysis. Then, the output of *FS-OPA* is a probabilistic model. The synthesis of a system in such model enables practitioners to realize mechanisms of relatively complex systems, as well as to make decisions with higher level of confidence. Next Section introduces *FS-OPA*.

Feature sensitivity through criterion-based resampling form OPA

FS-OPA is described from by the modifications of it highlighted in (Fig 3). They SS (Salient Samples through a criterion) and SM (Sampling from a phylogram Model). Next Sections present both SS and SM.

SS in FS-OPA

Tournament operator is a procedure used by *EDAs* to highlight promising regions in the search space. *OPA* performs it through selecting points (solutions/samples) in such regions, according to a purpose or optimization criterion. It picks up randomly a small number of samples (as individuals that compete in a tournament) and saves the best of them according to the criterion (also called objective function or fitness) in a set. This process repeats until the set has enough samples for modeling. In a certain way, the tournament is a resampling method with

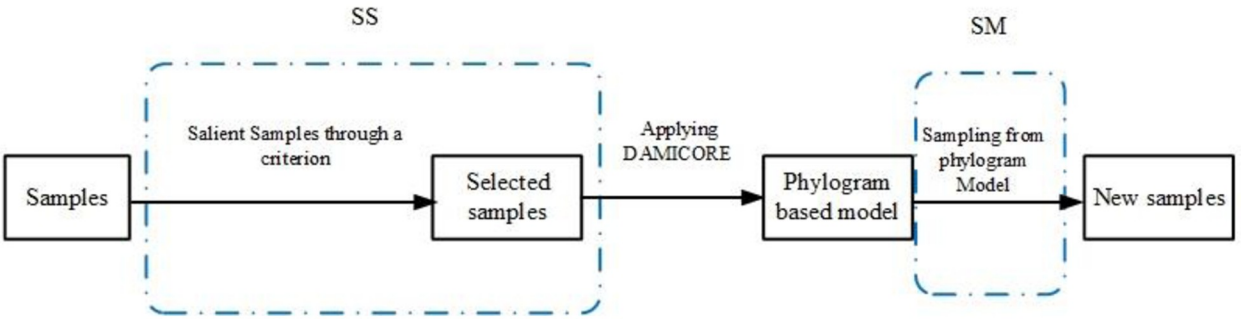


Fig 3. SS and SM rectangles in blue dashed lines highlight the OPA steps modified for the construction of FS-OPA.

https://doi.org/10.1371/journal.pone.0235147.g003

similar benefits, as it reduces the data unbalance bias, among other aspects that benefit the modeling process. The selection pressure (usually two, denoted as $s = 2$) is the number of individuals that compete in each tournament and it is a parameter to setup. Another relevant operator used by *EDAs* to salient promising regions of the search space is called ranking. After sorting all samples (according to a criterion) into a vector, the first best segment [8] of the vector composes the set of selected samples. The size of the top is adjustable by the selection pressure imposed (size of the top is equal to equal to $\frac{n}{s}$ where n is the total amount of number of samples), managed as a parameter.

FS-opa bypasses such a parameter by developing a selection that works with three levels of selection pressure. It uses ranking selection for the sake of simplicity in the code implementation and validation since it is a deterministic procedure. Besides, the stochasticity added by the use of the three levels can contribute to the reduction of bias. First, *FS-opa* ranks all the dataset according to the used criterion (e.g., minimizing LOS), generating an ordered sequence, a rank R. The three levels of selection pressures result in three levels of categorizations of R. First, FS-opa ranks all the dataset according to the used criterion (e.g., minimizing LOS), generating an ordered sequence, a rank R. The three levels of selection pressures result in three levels of categorizations of R. Differently from *OPA*, FS-opa saves both the samples at the top (head of the rank) and the bottom (tail of the rank) of each categorization:1- two-level categorization ($2C$, $s = 2$), 2-four-level categorization ($4C$, $s = 4$), and 3- eight level categorization ($8C$, $s = 8$).

A set called B_{2C} saves the best samples (according to the criterion) from $2C$, and another, W_{2C} stores the worst samples from it. Similarly, B_{4C} (B_{8C}) is the head with best samples after splitting R into four (eight) categories, and W_{4C} (W_{8C}) is the tail with the worst samples. (Fig 4) summarizes the SS procedure. Note that highest pressure can increase the number of categorizes subsets, however, results based on *OPA* show that those three values are usually enough for variety of complex problems [3].

SM in FS-opa

The execution of the model-building strategy of *OPA* for each of the six subsets generated by SS produces six phylogram-based models. (Fig 5) synthesizes the SM procedure. The phylograms represent the main relationships among variables (also called features or factors, depending on the domain). Based on them, we determine a set of principal features, as described in the sequel.

The algorithm used by SM for phylogram constructions is *DAMICORE* [3]. This method can work with data from any type and structure (integer, real and complex numbers, categorical data, images, sound, movies, etc.), as well as mixed data types, without any pre-processing

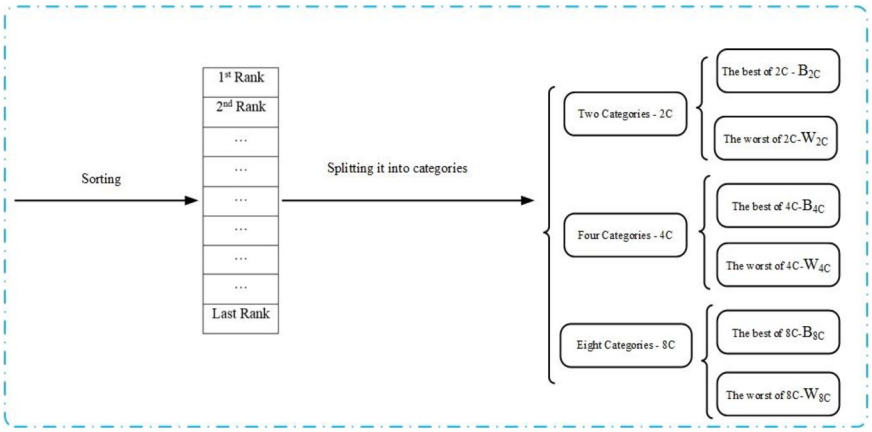


Fig 4. Components of SS in FS-opa that find relevant sample subsets B_{2C}, W_{2C}, B_{4C}, B_{8C}, W_{4C}, and W_{8C}.

https://doi.org/10.1371/journal.pone.0235147.g004

(filtering, outlier detection, feature extraction, among others). *DAMICORE* associates to each phylogram possible clusters of the strongly correlated objects in the analysis, where each object corresponds to a vector with the values of a variable manipulated by *FS-opa*. Moreover, it requires no parameter set up to run, although another code compressor (gzip is the default choice) improving compression can benefit the resulting quality.

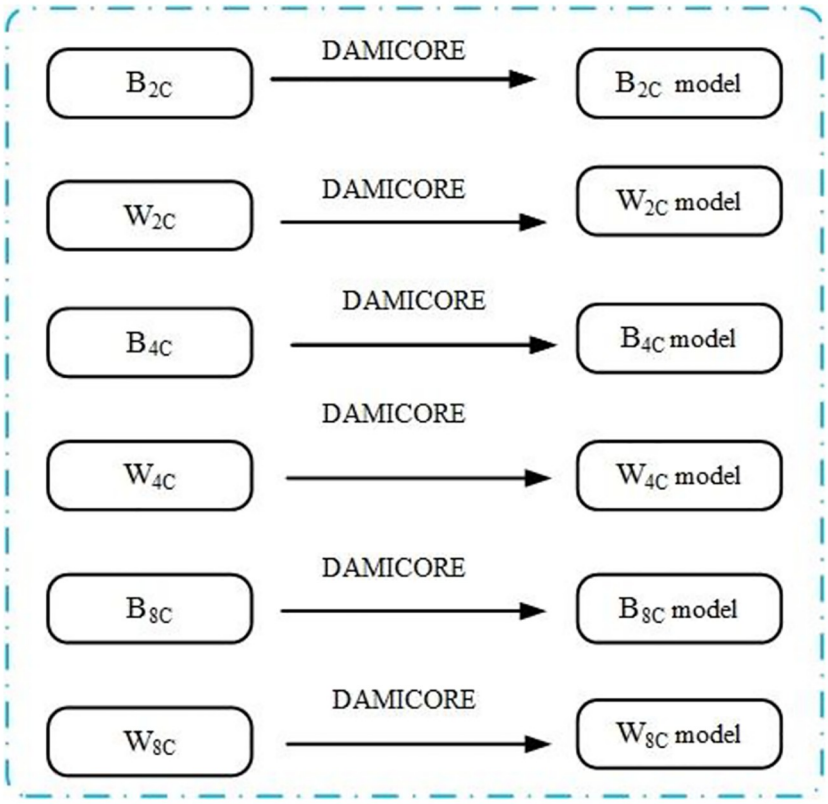


Fig 5. SM constructs six phylogram-based models.

https://doi.org/10.1371/journal.pone.0235147.g005

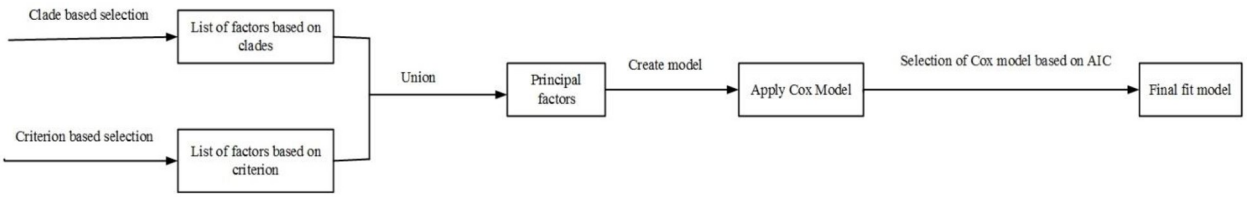


Fig 6. Feature sensitivity from models and their use in the Cox approach.

https://doi.org/10.1371/journal.pone.0235147.g006

SM applies DAMICORE to the six subsets (B_{2C}, W_{2C}, B_{4C}, W_{4C}, B_{8C}, W_{8C} samples (Fig 6), but the objects for *DAMICORE* manipulation are values associated to each variable. If data has a spreadsheet structure, each entire column corresponding to a variable (x_i) composes an object (also called x_i). In practice, a file called $x_i^{B_{2C}}$ ($x_i^{W_{2C}}$, $x_i^{B_{4C}}$, $x_i^{W_{4C}}$, $x_i^{B_{8C}}$, $x_i^{W_{8C}}$), saved in a directory/folder called B_{2C} (W_{2C}, B_{4C}, W_{4C}, B_{8C}, W_{8C}), stores the data associated with variable x_i from samples in B_{2C} (W_{2C}, B_{4C}, W_{4C}, B_{8C}, W_{8C}). It results in six directories, each one with l variables, where l is the total amount of variables of the problem (number of columns of a spreadsheet). The size of files $x_i^{B_{2C}}$, and $x_i^{W_{8C}}$, are $\frac{n}{2}$ in B_{2C} and W_{2C}, respectively. Similarly, the sizes of files $x_i^{B_{4C}}$ and $x_i^{W_{4C}}$ ($x_i^{B_{8C}}$ and $x_i^{W_{8C}}$) are $\frac{n}{4}$ $\left(\frac{n}{8}\right)$. Note that the size of files (number of samples used) decreases as the selection pressure increases. Finally, *DAMICORE* runs for each directory generating the six phylograms (preliminary models) with clusters associated (also called clades in the domain of evolution theory that used here by convenience).

(Fig 6) illustrates the analysis of the feature sensitivity of SM, organized into two techniques: one strictly based on clades found for each of the six categorized subsets (B_{2C}, W_{2C}, B_{4C}, W_{4C}, B_{8C}, W_{8C}) and another based on the sizes of the graph paths from a reference variable (leaf node) to the other variables (also leaf nodes) calculated for each phylogram. The former process is called SM1 and the latter, SM2. Note that finding the principal and less correlated factors is a way to find a parsimonious inference model. They can benefit further analysis by experts since they can focus on the more relevant aspects of a relatively complex system. Moreover, fewer features may reduce collinearity and improve the accuracy of prediction models until a certain point.

First, SM1 splits phylograms into two clades: C_1 and C_2. Supposedly, the majority of the objects in each clade have low-level common information. Thus, the remotion of the branch (graph edge) in the middle of the largest (graph) path of the phylogram generates such clades. Note that other strategies to estimate the largest uncorrelated clades from a phylogram are possible according to *OPA* principles, but the middle-path-based criterion is relatively simple and enough to produce relevant results. Moreover, if the largest path has an even number of branches, both remotions are tested and the one with higher congruence in the procedure described in the next paragraph is chosen.

SM1 applied to the six preliminary models produces twelve clades, denoted: C_1B_{2C}, C_2B_{2C}, C_1W_{2C}, C_2W_{2C}, C_1B_{4C}, C_2B_{4C}, C_1W_{4C}, C_2W_{4C}, C_1B_{8C}, C_2B_{8C}, C_1W_{8C}, C_2W_{8C}. Then, SM1 compares nodes from clade C_1B_{2C} (C_1B_{4C}, C_1B_{8C}) with both the nodes from C_1W_{2C} (C_1W_{4C}, C_1w_{8C}) and C_2W_{2C} (C_2W_{4C}, C_2W_{8C}) to find the most congruent clade pair (those with the most nodes in common, suppose it is (C_1B_{iC}, C_1W_{iC})). This pair and the remaining clades in each category compose respectively pairs p_1^i and p_2^i, for each category i (i in {2C, 4C, 8C}). In other words, $p_1^i = (C_1B_{iC}, C_1W_{iC})$ and $p_2^i = (C_2B_{iC}, C_2W_{iC})$ assuming the most congruent pair is (C_1B_{iC}, C_1W_{iC}); otherwise, $p_1^i = (C_1B_{iC}, C_2W_{iC})$ and $p_2^i = (C_2B_{iC}, C_1W_{iC})$ is the most congruent pair.

Suppose p_1^i is (C_1B_{iC}, C_1W_{iC}). SM1 selects the leaf nodes whose siblings changed in the corresponding phylograms when comparing any pair (x, y) of leaf nodes from p_1^i, x in C_1B_{iC} and y in C_1W_{iC}. Then, SM1 stores the leaf nodes (selected features) from p_1^i into s_1^i. SM1 runs also for p_2^i producing s_2^i. The output of SM1 is called a clade-based list, s_{clade}, corresponding to the union of s_1^i and s_2^i for all i.

Note that features (leaf nodes) whose siblings (relationships) changed from the model based on the best samples to the model constructed from the worst samples should have enough information to identify each of these two types of samples. Thus, those features are expected to be meaningful for problem modeling and solving.

(Fig 7) synthesizes the clade-based sensitivity by SM1 from two phylogram models, highlighting clades C_1 and C_2, respectively, from the best and the worst models. p_1^i is pair (C_1B_{iC}, C_1W_{iC}), where C_1B_{iC} and C_1W_{iC} are the clades surrounded by the green dashed lines respectively in the best and worst models. Similarly, p_2^i is pair (C_2B_{iC}, C_2W_{iC}) with clades highlighted by dashed blue lines. The siblings changed are in $s_1^i = \{A, B\}$ (the leaf nodes in red color) and $s_2^i = \{\}$, thus, $s_{clade} = \{A, B\}$.

SM2 uses the same four clades found by SM1 (C_1B_{iC}, C_2B_{iC}, C_1W_{iC}, C_2W_{iC}) for each category i. It also uses a unique pair q_1^i composed by the clades from the best and the worst models that have the target feature (leaf mode).

(Fig 8) highlights a target, feature A (yellow), and the clades of q_1^i (red-dashed lines) with it. The target is a feature with strong relation to a goal according to a criterion or a purpose of the analysis from the problem model. Such feature is called the target criterion.

Fig 8 also shows the two clades of pair q_1^i. The clade at left has two leaf nodes, A and B, in the subtree whose root is the sibling of the target feature, A. The clade at right has the sibling of A corresponding to leaf node B. Thus, $s^i_{criterion}$ results in {A, B, C}.

Note that SM2 requires some previous knowledge from the problem domain. Since it can indicate relevant features, SM2 runs based on them to compose $s_{criterion}$. This property of SM2 seems useful for adaptive approaches with human intervention when, for example,

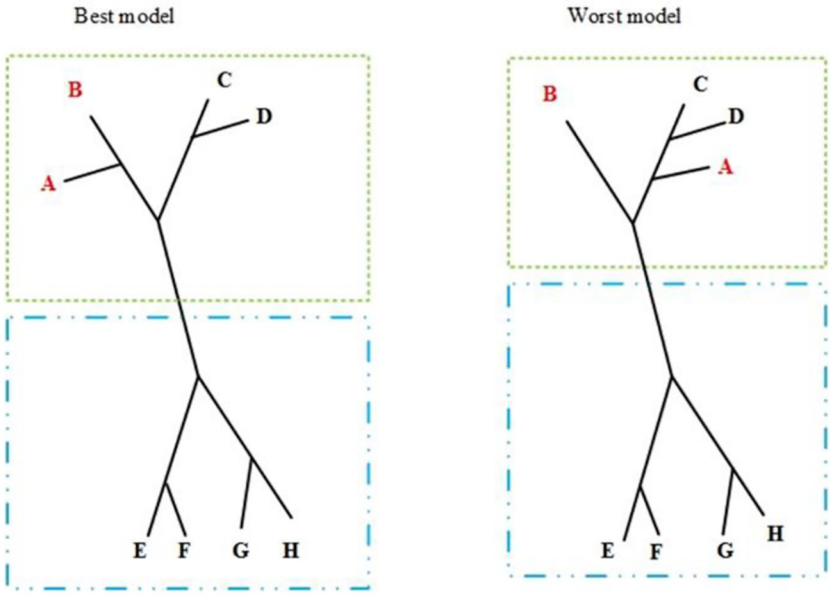


Fig 7. Clade-based list: p_1^i (p_2^i) with clades rounded by green (blue) dashed lines. $s_1^i = \{A, B\}$ and $s_2^i = \{\}$, therefore, $s_{clade} = \{A, B\}$.

https://doi.org/10.1371/journal.pone.0235147.g007

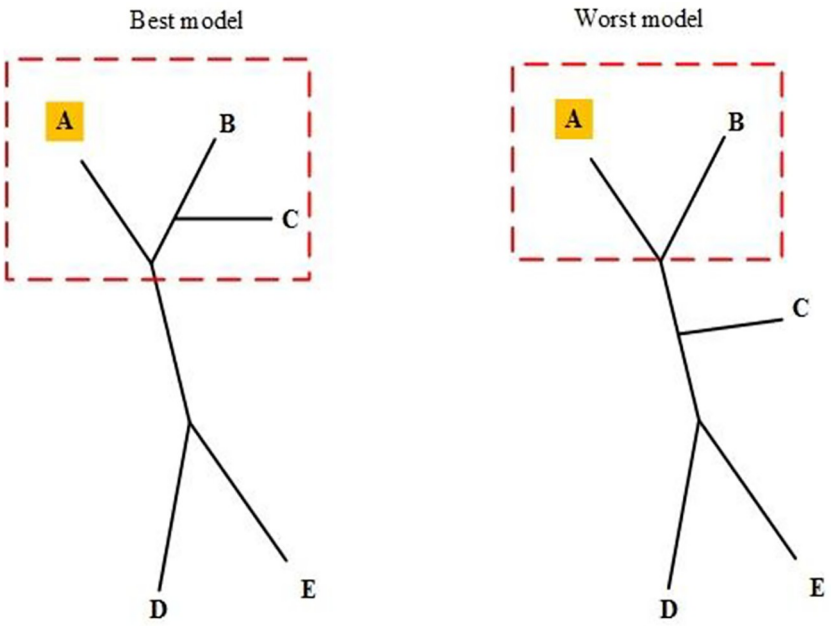


Fig 8. Criterion-based lists, Red dashed lines highlight subtrees found by SM2 using node A as a target feature. $s_{criterion}$ = {A, B, C}.

https://doi.org/10.1371/journal.pone.0235147.g008

practitioners motivated by some partial results manipulate the dataset or the found models according to their experience. In this way, *FS-opa* can deal with several levels of knowledge from the problem domain, benefiting from it, when possible. For example, in our dataset, other features have some relation to LOS. Each of those features can also be treated as a target to improve problem modeling. In such a case, SM2 should run for each of the new targets generating new criterion-based lists that must be combined through the union operator.

Finally, the lists produced by SM1 and SM2 are combined in a same set *r*, by applying the union operator to s_{clade} and $s_{criterion}$. Based on *r*, regression methods, machine learning, among other approaches, can construct complete models for prediction, diagnosis, etc.

Cox approach in FS-opa

Our purpose of investigation in this project concerns on time length of treatment or staying in hospitals, then survival analysis is a relevant strategy to construct a stochastic problem model. Cox approach is used to complete a stochastic phylogram-based modeling through *FS-opa*. (Fig 9) illustrates the integration of SM and Cox approach in *FS-opa*.

Cox regression model is a semi-parametric regression model, which is the most popular in common in medical research to analysis data with time to discharge [24–26]. For each patient

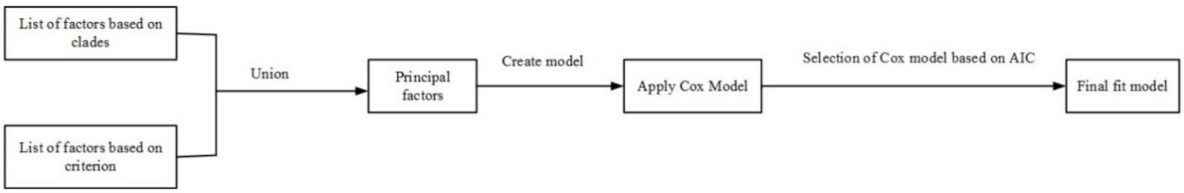


Fig 9. Feature sensitivity from models and their use as the main variables in the Cox approach.

https://doi.org/10.1371/journal.pone.0235147.g009

i in the dataset, we have the times of hospitalization m_i, m_i is the number of observed events for patient i. Thus, the LOS for the i^{th} patient and for j^{th} stay is given by $t_{ij} - t_{ij-1}$ where $T_{i,0} = 0$, for all $i = 1, \ldots, n$ and $j = 1, \ldots, m_i$, where n is the total number of patients. The hazard function for the j^{th} event of i^{th} subject at time t, $i = 1, \ldots, n$ and $j = 1, \ldots, m_i$, is given by Eq (1) [26]. The Cox model was applied to determine the significant covariate, which are associates with LOS The model was fitted using survival package in R [27,28].

$$h(t_i, X_i) = h_0(t_i)\exp(\sum_{j=1}^{p} X_{ij}\beta_j) \quad (1)$$

$h_0(t_i)$is the baseline function.

$X_i(X_{1j}, X_{2j}, \ldots, X_{pj})$ is the set of feature n vectors.

Experiments

The experiments of FS-opa with the mental dataset are displayed in this section. In *Fs-opa* with the raw-full dataset, part presents the main results of *FS-opa* from the raw-full dataset. In addition, in the section of column-constrained datasets illustrate how column-constraint datasets are obtainable by revisiting the dataset armed with simple hypotheses to verifying the consistency of the found results.

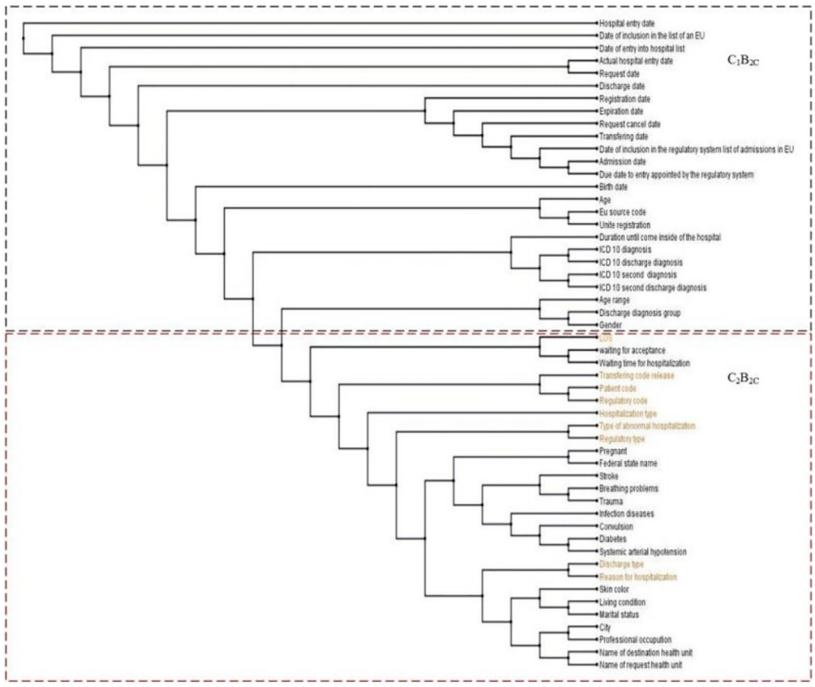
Fs-opa with raw-full dataset. *FS-opa* is first applied to the dataset as found, without any pre-processing or intervention. The idea is to verify what is possible to model with no previous knowledge of the problem domain. The dataset was obtained from the mental health care information system responsible for coordination of hospitalizations in mental health specialized hospitals in the region of Ribeirao Preto, Brazil and it was composed by 8,755 samples (rows) with 52 features (columns), labeled as shown S1 Appendix. A simple inspection of the dataset shows that the majority of them are male, single and living alone, with white skin color. Patients have aged from 5 to 89 years, among other aspects. Moreover, missing values and some inconsistencies can also be found in the raw dataset. Such information is not used in the modeling through *FS-opa* in the study based on the raw-full dataset, as presented in the sequel.

It is worth to remark that *FS-opa* first ranks samples according to a criterion and splits the ranks into some categories (2*C*, 4*C* and 8*C*), generating six subsets (B_{2C}, W_{2C}, B_{4C}, W_{4C}, B_{8C}, W_{8C}). Then data associated to each feature (column) is saved into the same file, with the same name of the corresponding column, in a directory with the same name of the subset. Then, FS-opa runs for each directory producing six phylogram-based models. It is worth to remark that *FS-opa* first ranks samples according to a criterion and splits the ranks into some Instead of presenting only set r with all the features selected by *FS-opa*. For instance, Figs 1 and 11 shows the partial results found by procedures SM1 and SM2 of the 2C category to illustrate how the practitioner can manipulate FS-opa to deal with a real scenario.

(Fig 10a) shows the phylogram and the selected variables (orange) obtained by SM1 applied to B_{2C}. Similarly, (Fig 10b) shows the phylogram and the selected variables (orange) produced by SM1 executed for W_{2C} (Fig 11a and 11b) synthesizes the models and other results generated by SM2 run for B_{2C} and W_{2C}, respectively.

Fs-opa with raw-full dataset. Based on the results from the 52-feature model, some improvement can be achieved by revisiting the dataset armed with simple assumptions. The process of revision is also a way of verifying the consistency of results, as when it is obtained by a resampling procedure. Six datasets, with less than 52 features, were derived from the 52-Feature Dataset (52-FD) and *FS-opa* applied to them. As described in the sequel, the resampling doesn't require any previous knowledge from the domain of mental health.

a: B_{2c}



b: W_{2c}

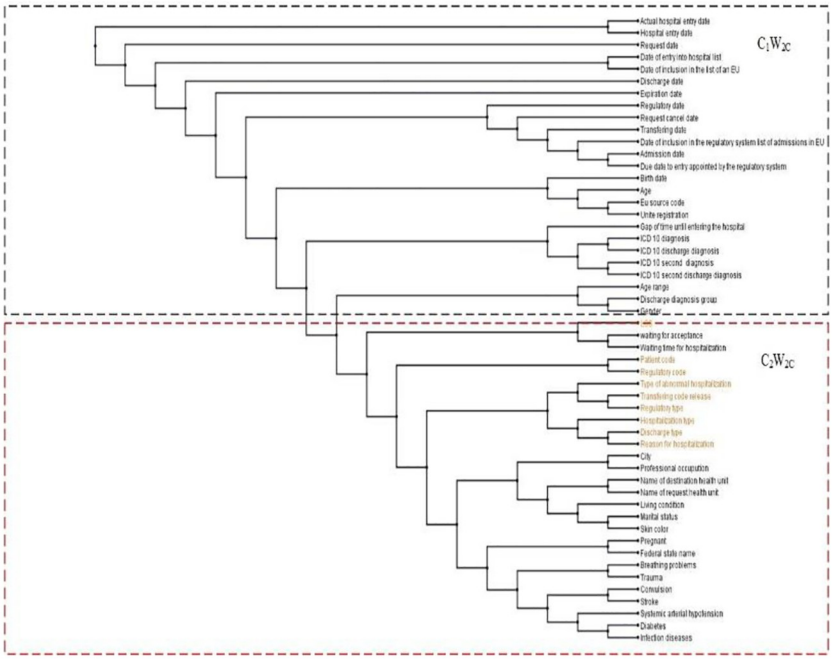
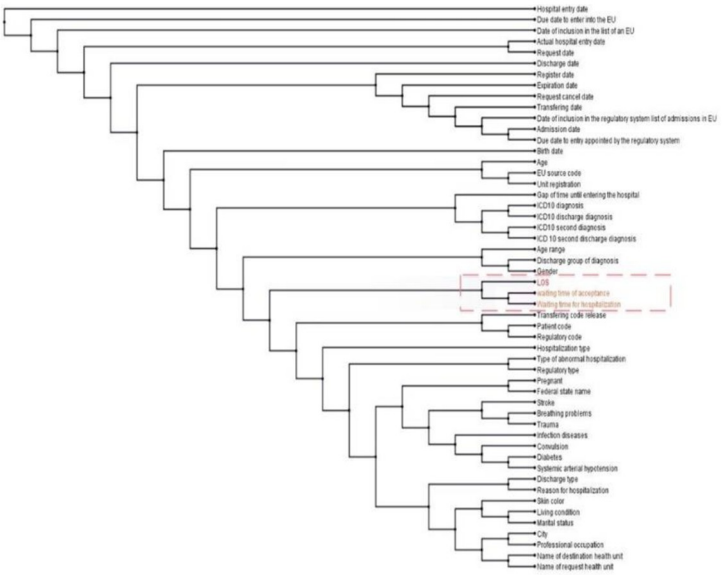


Fig 10. SM1 of B_{2c} and W_{2c}. a: B_{2c}, b: W_{2c}.

https://doi.org/10.1371/journal.pone.0235147.g010

a: B_{2c}



b: W_{2c}

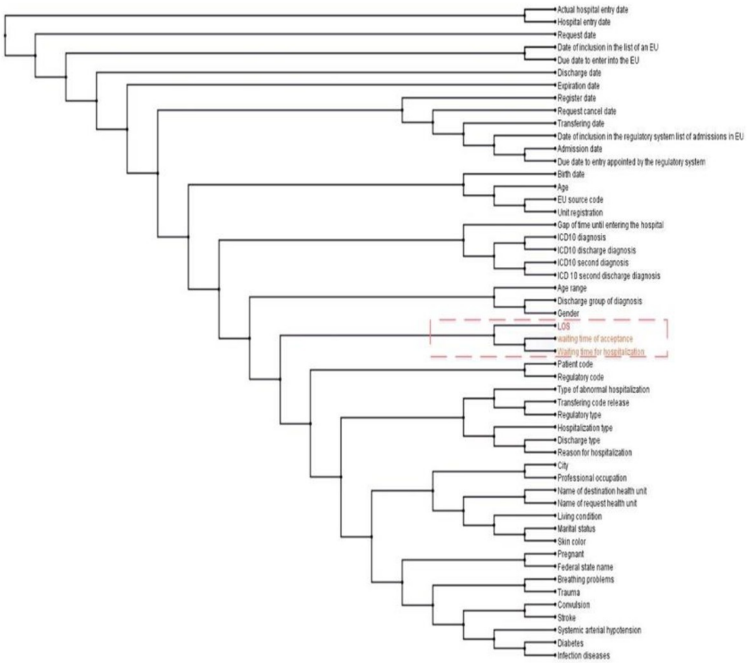


Fig 11. SM2 of B_{2c} and W_{2c}. a: B_{2c}, b: W_{2c}.

https://doi.org/10.1371/journal.pone.0235147.g011

A first resampling (RS1), the number of columns is constrained based on simple inspection of the amount of missing values for each variable. Columns with more than 85% with empty cells were removed transfer code release, admission date of emergency section, cancel request date, date of entry into emergency list, date of reserve into emergency section and transfer date generating the 46-Feature Dataset (46-FD).

Another resampling (RS2) produced the 37-feature dataset (37-FD) by removing columns that seem to contain mostly repeated information. The first features related to "date" were removed. For example, admission date and register date are usually both related to instant of time that a patient reached the hospital for the first time. Birth date and age is a similar case since age depends on the patient's birthday. Note that most of the meaningful information related to time is retained by LOS. Features storing codes (patient code, regulation code, and unit source code) were also removed. Data of type "code" can be seen as redundant labels or IDs. However, they can benefit human or computer-aided management; they are relatively less relevant for modeling than other variables.

Differently from raw-full dataset, RS1 and RS2, the resampling called RS3 is based on knowledge extracted from the results of *FS-opa*, but not on knowledge from experts on the area. Phylogram from 52-FD, 46-FS and 37-FD have some clades preserved in all of them. It means that the features associated by them didn't reveal any relevant information that enabled *FS-opa* to distinguish "good" samples from "bad" samples. Those features, diabetes, stroke, systemic arterial hypotension, trauma, ICD10 diagnosis, ICD10 discharge diagnosis, ICD10 second discharge diagnosis, infection disease, code patient, code regulation, unit code source, convulsion, date entrance, unite registration and breathing problems less relevant for problem modeling, thus, they were removed from RS2 in order to compose RS3, generating the 21-FD that phylogram trees.

Comparisons of sets r (unions of s_{clade} and $s_{criterion}$) obtained from RS1, RS2 and RS3 (columns three, four and five of in Tables 1 and 2) to the set r obtained from 52-FD (column two of Tables 1 and 2) can generate two new lists: *i*) with common features (RS4), that were selected by FS-opa in at least two of the three resamplings (RS1, RS2 and RS3) and they were also selected from 52-FD; *ii*) with novelty features (RS5), that selected by *FS-opa* in at least two of the three resamplings but they were not selected from 52-FD.

Finally, the list with common and novelty features are then combined into RS6. Common features are one estimate of the most robust subset among all the sets *r* that were found. On the other hand, novelty features are features that require resampling to become salient, but with potential possess relevant information for modeling.

A certain concerning is that novelty features can be just noise. Once again, resampling by the bootstrap technique [29] is a way to verify it, but in this thesis, we evaluate it by checking if AIC of the associated Cox model is improved. Note that other information criterion to determine a model representativeness of data can be investigated as, for example, Bayesian Information Criterion (BIC) [30], likelihood ratio test [31], Bayes Factor [32], and minimum description Length [33] as a measure of parsimony.

The lists of features from s_{clade} and $s_{criterion}$ obtained by SM1 and SM2 of *FS-opa* from all the seven datasets investigated (52-FD / Full, 46-FS / RS1, 37-FD / RS2, 21-FD / RS3, 11-FS / RS4, 6-FD / RS5, and 15-FD / RS6 are synthesized Tables 1 and 2, respectively.

Results and discussions

The final step of *FS-opa* (Fig 9) is the construction of Cox model based on the selected lists in order to find the best set.

Table 1. Clade-based selected lists, s_{clade}, obtained by SM1 from 52, 46, 37, and 21-FDs, respectively, Full, RS1, RS2 and RS3 resamplings; as well as, common (9-FD) and uncommon features (6-FD) and union of both sets (15-FD) corresponding to RS4, RS5 and R.

	raw-Full dataset and basic resamplings				Common features in relation to Full	Novelty features in relation to Full	Common and novelty features together
#	52-FD	46-FD (RS1)	37-FD (RS2)	21-FD (RS3)	11-FD (RS4)	6-FD (RS5)	15-FDRS6
1		Age	Age	Age		Age	Age
2	Age range						
3	Code patient	Code patient					
4	Code Regulatory	Code Regulatory					
5	Date birth (*)						
6	Discharge diagnosis group	Discharge diagnosis group	Discharge diagnosis group		Discharge diagnosis group		Discharge diagnosis group
7	Type of discharge	Type of discharge	Type of discharge	Type of discharge	Type of discharge		Type of discharge
8	Hospitalization Type	Hospitalization Type	Hospitalization Type	Hospitalization Type	Hospitalization Type		Hospitalization Type
9	Living condition	Living condition		Living condition	Living condition		Living condition
10	LOS	LOS	LOS	LOS	LOS		LOS
11	Marital status						
12		Pregnancy	Pregnancy	Pregnancy		Pregnancy	Pregnancy
13	Gender	Gender	Gender	Gender	Gender		Gender
14	Skin color						
15	Type of abnormal hospitalization	Type of abnormal hospitalization	Type of abnormal hospitalization	Type of abnormal hospitalization	Type of abnormal hospitalization		Type of abnormal hospitalization
16	Regulatory type	Regulatory type	Regulatory type	Regulatory type	Regulatory type		Regulatory type
17	Reason for hospitalization	Reason for hospitalization		Reason for hospitalization	Reason for hospitalization		Reason for hospitalization
18				Federal State name		Federal State name	Federal State name
19				Waiting time for hospitalization		Waiting time for hospitalization	Waiting time for hospitalization
20				Waiting time of acceptance		Waiting time of acceptance	Waiting time of acceptance

(*) Date birth wasn't used in Cox model since it has date format.

https://doi.org/10.1371/journal.pone.0235147.t001

Table 2. Criterion-based selected lists, $s_{criterion}$, obtained by SM2 from 52, 46, 37, and 21-FDs, respectively, Full, RS1, RS2 and RS3 resamplings; as well as, common (9-FD) and uncommon features (6-FD) and union of both sets (15-FD) corresponding to RS4, RS.

	raw-Full dataset and basic resamplings				Common features in relation to Full	Novelty features in relation to Full	Common and novelty features together
#	52-FD	46-FD (RS1)	37-FD (RS2)	21-FD (RS3)	11-FD (RS4)	6-FD (RS5)	15-FD RS6
1				Age		Age	Age
2	Discharge diagnosis group	Discharge diagnosis group	Discharge diagnosis group		Discharge diagnosis group		Discharge diagnosis group
3	LOS	LOS	LOS	LOS	LOS		LOS
4	Gender	Gender	Gender	Gender	Gender		Gender
5	Regulatory type	Regulatory type	Regulatory type		Regulatory type		Regulatory type
6	Waiting time for hospitalization	Waiting time for hospitalization	Waiting time for hospitalization	Waiting time for hospitalization	Waiting time for hospitalization		Waiting time for hospitalization
7	Waiting time of acceptance	Waiting time of acceptance	Waiting time of acceptance	Waiting time of acceptance	Waiting time of acceptance		Waiting time of acceptance
8				Gap of time until entering the hospital		Gap of time until entering the hospital	Gap of time until entering the hospital

https://doi.org/10.1371/journal.pone.0235147.t002

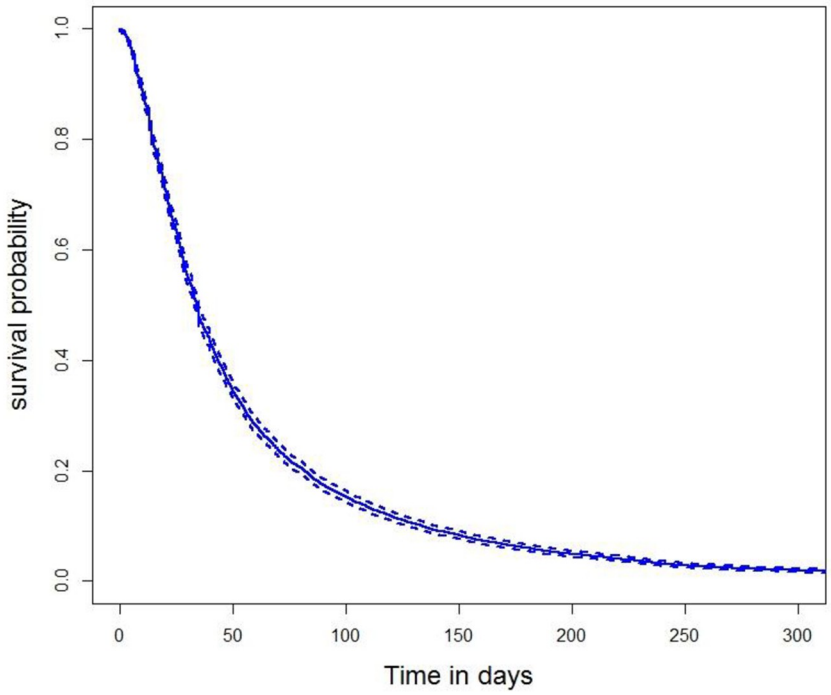


Fig 12. Survival probability according to LOS obtained by the model with best AIC (RS6).

https://doi.org/10.1371/journal.pone.0235147.g012

First Cox regression is applied to each of the seven lists of features investigated. The covariates of Cox model are shown in Tables 1 and 2. The best-fitted model, as well as the significant covariates, can be chosen by using the Akaike Information Criterion (AIC) [34] since we don't have nested models (two models are nested if one model contains all the terms of the other, and at least one additional term) [35] When we simultaneously analyze the significance of the covariates in the modelling. In order to choose the significant covariates, we proposed to use the forward approach [36] combined with the AIC. Using this method, we keep the covariate in the model if it decreases the AIC value, otherwise it is assumed as non-significant for explaining the target variable, in our case, the LOS in psychiatric hospitals.

The AIC values related to Cox models obtained are depicted in (Fig 12). The lowest valued for AIC is 68010.27 for the 15-FD based model. It corresponds to RS6, the union of features from RS4 (common features, robust ones) and RS5 (novelty features), thus, such combination seems to be capable of extracting information from data better than other resamplings for modeling through Cox approach.

(Fig 13) presents an evaluation of Cox models according to a multicriteria decision making technique based on non-dominated sets [37] as used by *OPA*. First solutions are plotted in the bi-objective space with the two dimensions Number of Features and Normalized AIC values (normalized with high AIC value = 106868.6). A "rectangle" is associated in the quadrant at the "northeast" of each point (reference). Note that any other point in such a rectangle has equal or highest AIC as well as equal or higher number of features than the reference. We say that a reference point is non-dominated by the points inside its rectangle, in the same way that the points of the rectangle are dominated by the reference point. Thus, the multi-criteria decision making can be applied to the Cox models generated using the seven lists of features in order to compare them.

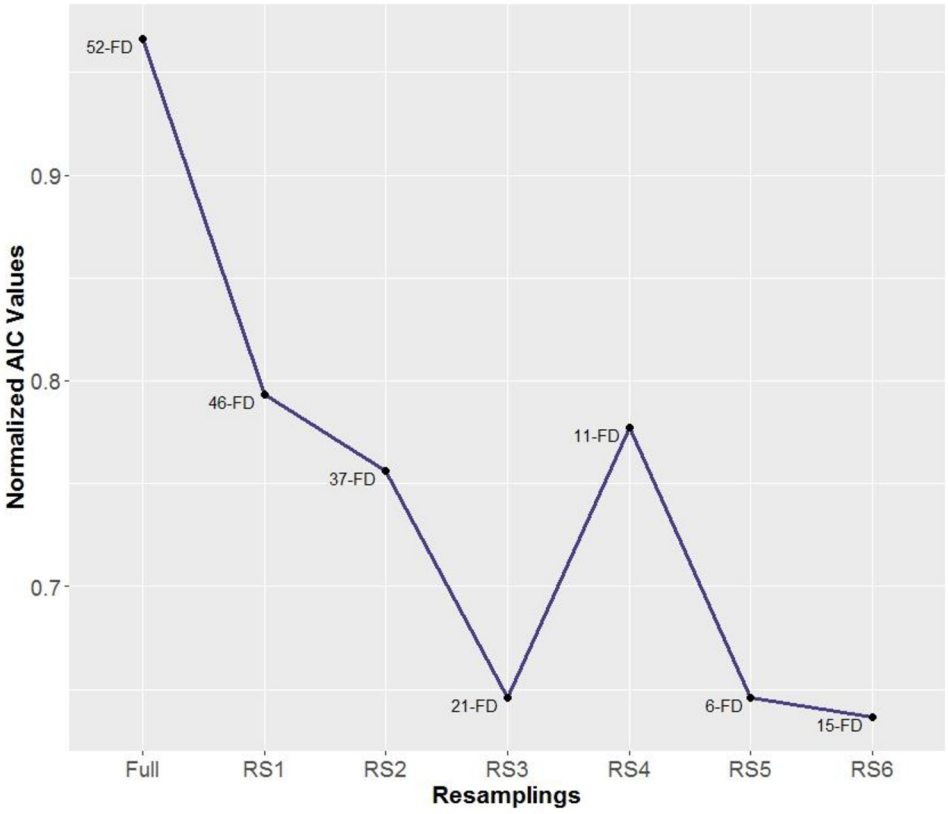


Fig 13. Comparison of the seven Cox models generated according to AIC.

https://doi.org/10.1371/journal.pone.0235147.g013

(Fig 14) shows that RS2 point dominates RS1 point, RS3 point dominates RS2, RS4 dominates RS1, RS6 dominates RS4, RS3, RS2 and RS1, and RS5 dominates RS4, RS2 and RS1. However, no point dominates RS5 and RS6, then both points are non-dominated and compose the best set of models according to multicriteria decision making based on AIC and number of feature objectives. The trade-off between RS5 and RS6 models are clear since RS6 has the lowest AIC (69020.64, while RS6 has AIC equal to 68010.27) but RS5 uses 6 features while RS6 requires 15 features.

Next, we compare each of the performance of Cox models for each column-constrained dataset due to the feature lists obtained by each of the procedures SM1 and SM2, as well as those obtained by the complete *FS-opa* (SM1+SM2). (Fig 15) enables the evaluation of the relative contribution of each subset in order to improve Cox models. Clearly, SM1 sensitively provide the largest improvements in AIC, while SM2 increments such performance. Note that SM1 requires no previous knowledge from data, it is data agnostic. On the other hand, SM2 requires at least a target variable, assumed highlighted related to the scope or purpose of the model.

In a certain way, the results show that much of the modeling improvements modeling based on the proposed feature sensitivity analysis doesn't depend on any previous domain knowledge. Such a result is coherent with usual *OPA* achievements since it works with black-box problems (the term related to no knowledge from the problem in the optimization field) and it has found optimal solutions for large-scale multimodal optimization problems.

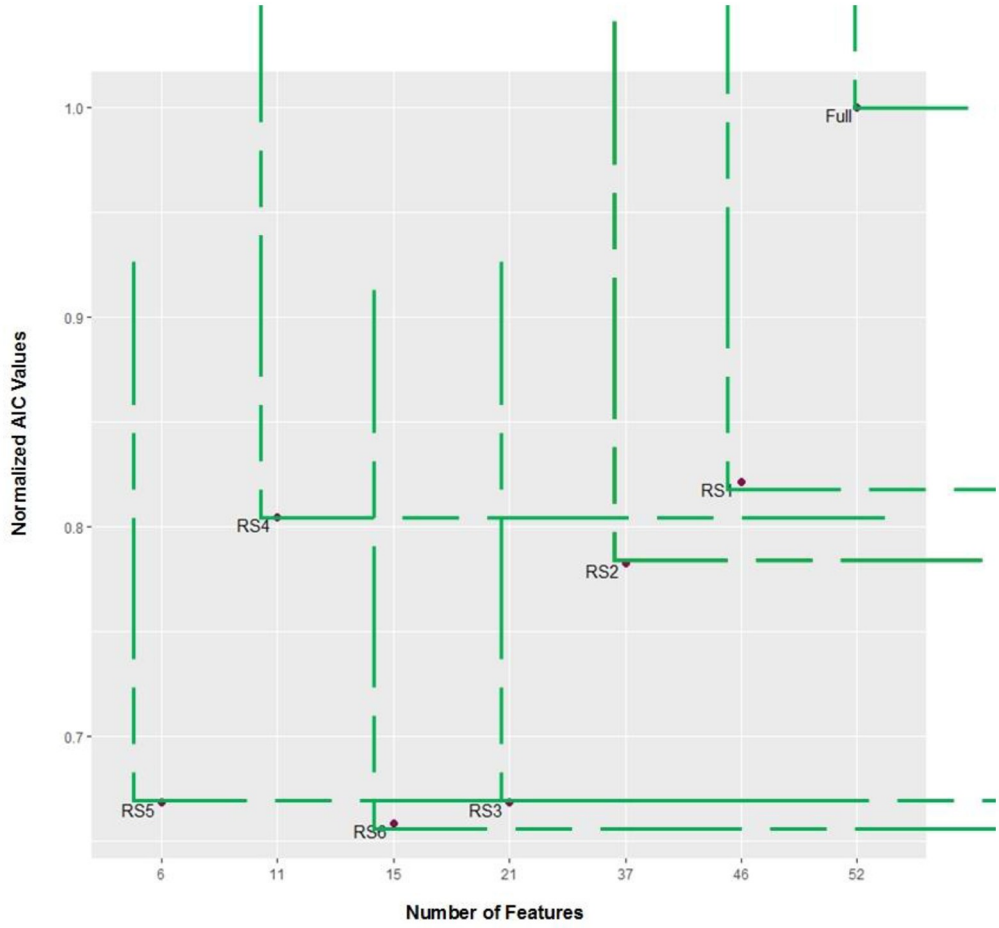


Fig 14. Evaluation of the seven models by a multicriteria decision-making technique. The colored quadrant (rectangle corners) highlights regions dominated by each point (model). RS5 and RS6 are the non-dominated models.

https://doi.org/10.1371/journal.pone.0235147.g014

(Fig 16) presents other relationships among the quality of the Cox models generated according to a bicriteria evaluation, synthesized by the non-dominated sets. A curve or line connecting them in the biobjective space is called a non-dominated front. Beside AIC value, the number of features of each set is another criterion for minimizing, since it is a measure of parsimony. Analysis of trade-offs produces extreme points in the front, corresponding to an adequate value for one criterion and poor value for the other. The Cox model based on 6-FD-SM2 uses the lowest number of variables, while its AIC (larger than 0.95) is near to the largest found, related to 52-FS-SM1 (1.00). On the other hand, the Cox model from 15-FD (SM1+SM2) in the front has 15 variables and the found lowest AIC.

There are several techniques to choose a solution (a set) from a front, but a relatively simple and useful indicates the solution in the elbows of the front. (Fig 16) shows a unique elbow composed by sets 6-FS-SM21 and 6-FD. Both are near to the ideal point (that corresponds to zero AIC and one variable used). According to the number of variables both models are similar, but the 6-FD based model has the second-best found AIC, highlighting it in some way.

Such result was the best *FS-opa* that provided without additional knowledge for the mental health dataset. The evaluation of the relevance of the 6-FD set according to experts on the mental disorder as well as the LOS estimates from the corresponding Cox model are both

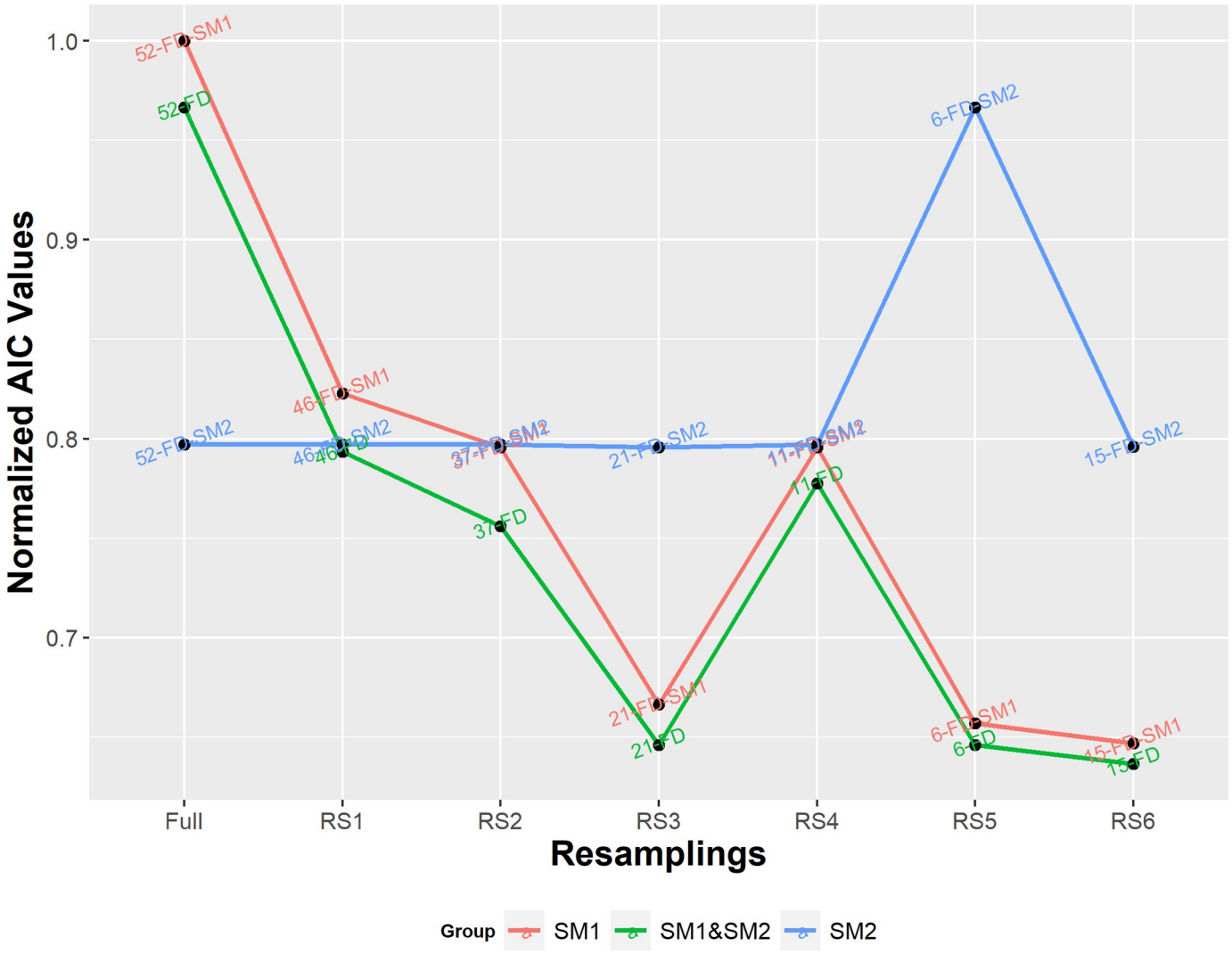


Fig 15. Individual and relative contribution of each dataset to the performance of Cox models.

https://doi.org/10.1371/journal.pone.0235147.g015

investigations that should be performed. Although their relevance, they are proposed as future work since it demands a relatively long-time cooperation with experts. In a certain way, such perspective conflicts with the main purpose of the *FS-opa*, to extract as much as possible information from a dataset without previous knowledge from the problem domain. It is important to remark that *FS-opa* results also should contribute for further investigations since: i) solutions in front are the best found ones, thus, by using them, experts will only work with consistent sets, with relatively low level of redundancy; ii) *FS-opa* found high quality models using sets with low number of variables, then any new hypothesis (involving new combinations of those variables or parameter setups of Cox Model) could be tested; and iii) other hypotheses or models based on knowledge from experts can be compared to models in the front, as reference (with no bias and agnostic to data) for evaluating and/or normalizing the improvement of them.

Finally, (Fig 17) presents another comparison based on bicriteria evaluation of the Cox models generated, but now in relation to the literature results from the literature for mental health disorders. A review of papers in this field found a consensus of the principal features

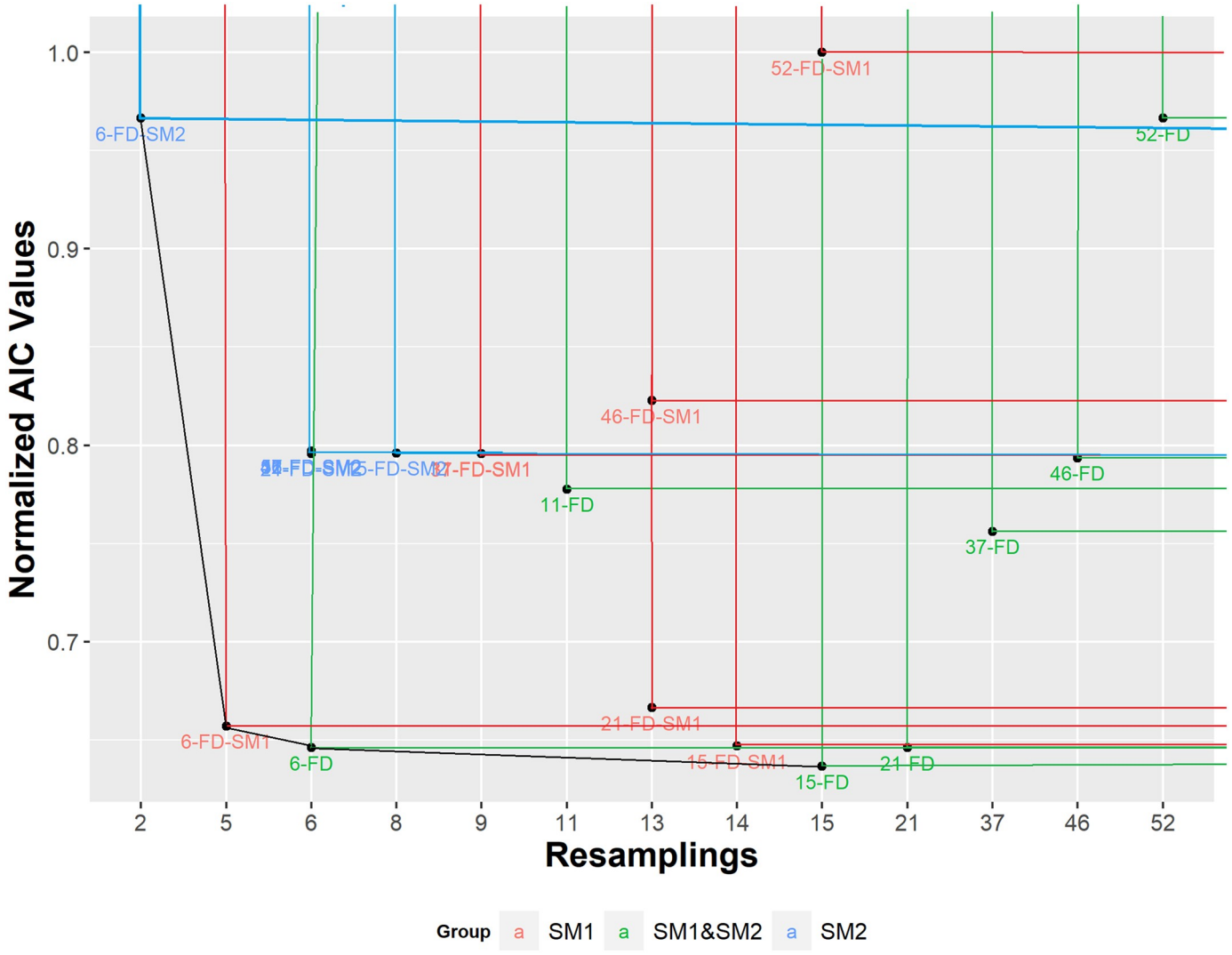


Fig 16. Relative contribution of each set to the Cox models in a trade-off analysis based on non-dominance, according to two criteria: Normalized AIC and number of features used.

https://doi.org/10.1371/journal.pone.0235147.g016

that have been used as indicators for LOS, such as age, gender, living condition, skin color, marital status, and professional occupation. Such a consensus set corresponds to the best model that *FS-opa* generated without additional knowledge.

Conclusions

The essence of data mining techniques is the possibility of discovering valuable data; they have recently become a predominant field of research with broad applications, specifically in medical healthcare. This study focused on developing a mining approach based on the optimization method called *OPA*, since can work with relatively complex problems without previous knowledge. This method constructs probabilistic models of correlated variables based on hierarchical clustering techniques. *DAMICORE* is one of the main clustering methods employed by *OPA* that was also used in this project. Based on a combination of three other methods (normalized compression distance, Neighbor-Joining and Fast Newman [14] it can deal with different

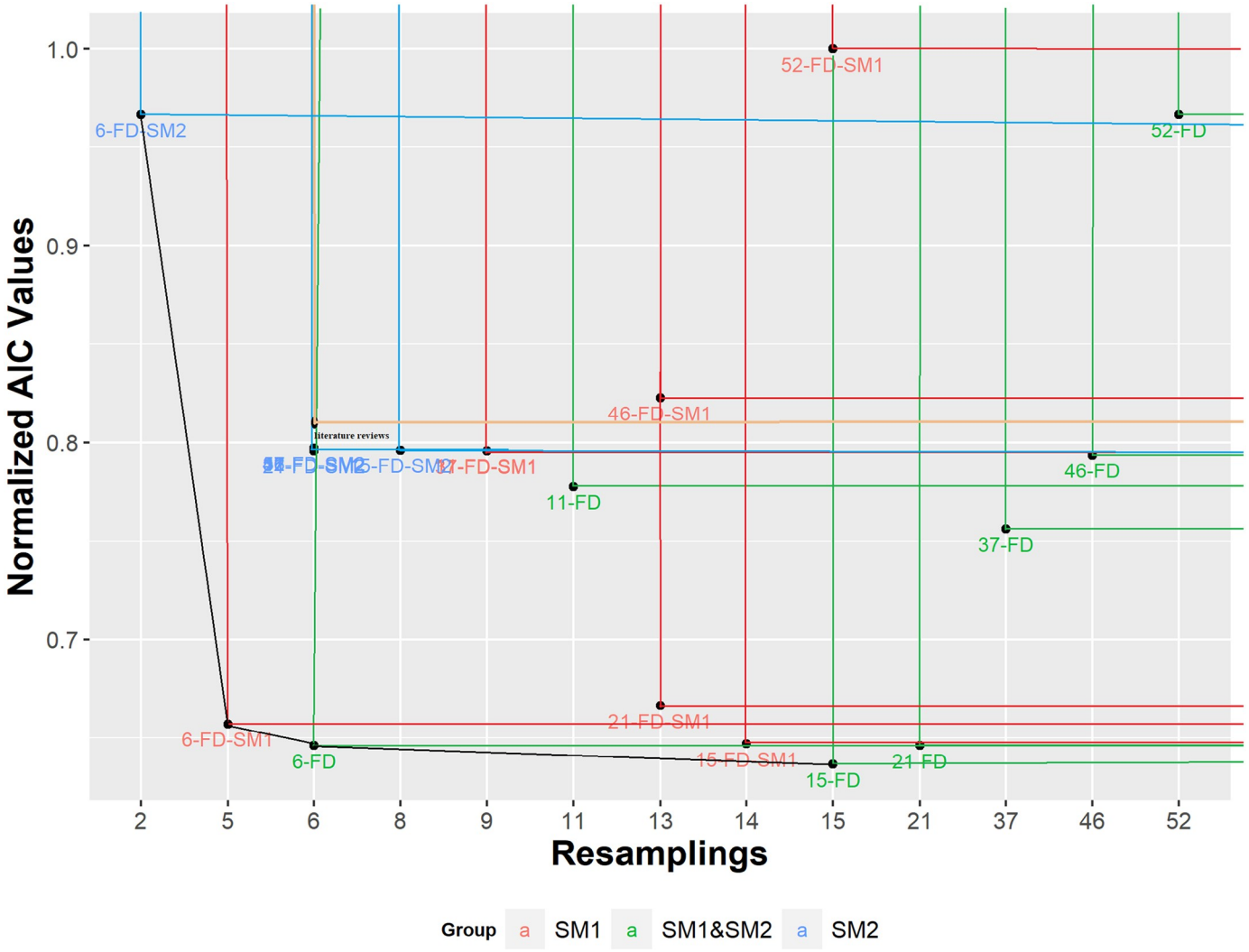


Fig 17. Relative contribution of each set to the Cox models in a trade-off analysis based on non-dominance, according to two criteria: Normalized AIC and number of features used in comparisons with literature reviews.

https://doi.org/10.1371/journal.pone.0235147.g017

types of dataset together without the previous transformations that may require knowledge from the problem domain.

The main contribution of this study is a new approach for Feature Sensitivity analysis from *OPA*, called *FS-opa*. It also involves the combination of other statistical methods in order to enable practical results from such analysis. For the study case of mental health disorders, the Cox approach for survival analysis was employed for predicting LOS.

Need to mention that the Current feature selection method will require at least a data normalization or rescaling (for example, transform data to real numbers or Binary values) and some treatment-related to the imputation of missing data. That requires some prior knowledge from the domain, and the obtained result is not raw data. Thus, the required comparison is only possible with a violation of our purpose (work directly with raw data without any prior knowledge from the domain). The only way to proceed with the comparison would be in case the raw data has no issues to treat (requiring no expertise form the problem domain). Still, this case is trivial and is not motivated for a more robust data mining method.

Moreover, *FS-opa* can work with a raw dataset, showing that it requires no knowledge from the problem domain to obtain preliminary prediction models. Moreover, the improvement of them from the raw dataset is viable through a series of simple hypothesis among data quality, i.e. that usually doesn't require knowledge from experts and are checkable by the *FS-opa* together with a multicriteria decision-making approach based on non-dominated sets. In fact, the improvement of data quality is arranged as a resampling procedure, where each new resampling set is determined according to the hypothesis. Results show that the consensus of variables selected from each resampling as well as the novel found features in each resampling are relevant for modeling.

In the experiments performed, the multicriteria decision-making strategy of *FS-opa* to find the best Cox models generated from the resampling procedure found that RS5- and RS6-based models are the non-dominated Cox models with an adequate trade-off. These results also emphasize that it is possible to construct relevant models from a relatively complex raw dataset without prior knowledge from the problem domain.

It also worth to note that our results are consistent with several studies [38–41] Some previous research showed that living with family and not married affected the long duration of psychiatric hospitalization [42,43].

This study demonstrates that the factors related to the time and type of hospitalization in our sample do significantly influence on the LOS. Furthermore, many studies find an association between LOS and diagnosis [41,44,45] We found the diagnosis, treatment type (involuntary, voluntary and compulsory treatment) have a positive impact on the LOS.

Naturally, knowledge from experts must be used when they are available. In this way, FS-opa can be oriented according to it. For example, by using the relative importance of features through the SM2 procedure of *FS-opa*, new criterion-based selected lists can be generated. Then, they can be analyzed together with the other lists obtained by *FS-opa* as in the resampling procedure combined with the multicriteria decision-making approach. Moreover, these findings have important implications for efficient economic management and reduce LOS of psychiatric patients in the health care system. Indeed, the proposed approach enables researchers with little knowledge of the evolutionary computation field to apply *FS-opa* for their dataset.

Our findings improve in some way the psychiatric service and the socioeconomic status of the psychiatry department. Moreover, it can benefit the directions of future studies crucially needed in this area. Due to the increasingly more effective and efficient data collection and storage mechanisms in a variety of medical fields coupled with the enormity of ever more complex problems, *FS-opa* seems a method that can contribute to deal with such complexity. Finally, other approaches based on *FS-opa* principles may enable the improvement of analysis in the areas of healthcare.

Supporting information

S1 Appendix. List of all variables form 52-FD, their meaning in English, the corresponding data types and value ranges.
(DOCX)

Acknowledgments

We would like to thank Centro de Informaçao e Informatica em Saude (CIIS) at Ribeirao Preto Medical School—University of São Paulo for providing us with psychiatric dataset on admission and discharge dates.

Author Contributions

Conceptualization: Fatemeh Gholi Zadeh Kharrat, Alexandre Cláudio Botazzo Delbem.

Formal analysis: Fatemeh Gholi Zadeh Kharrat, Newton Shydeo Brandão Miyoshi.

Funding acquisition: Fatemeh Gholi Zadeh Kharrat.

Methodology: Fatemeh Gholi Zadeh Kharrat.

Project administration: Fatemeh Gholi Zadeh Kharrat, Alexandre Cláudio Botazzo Delbem.

Resources: João Mazzoncini De Azevedo-Marques, Paulo Mazzoncini de Azevedo-Marques, Alexandre Cláudio Botazzo Delbem.

Supervision: João Mazzoncini De Azevedo-Marques, Paulo Mazzoncini de Azevedo-Marques, Alexandre Cláudio Botazzo Delbem.

Visualization: Juliana Cobre.

Writing – original draft: Fatemeh Gholi Zadeh Kharrat.

Writing – review & editing: Juliana Cobre, João Mazzoncini De Azevedo-Marques, Paulo Mazzoncini de Azevedo-Marques, Alexandre Cláudio Botazzo Delbem.

References

1. Gandomi A, Haider M. Beyond the hype: Big data concepts, methods, and analytics. Int J Inf Manage. 2015;
2. Hauschild M, Pelikan M. An introduction and survey of estimation of distribution algorithms. Swarm Evol Comput. 2011;
3. Soares A, Râbelo R, Delbem A. Optimization based on phylogram analysis. Expert Syst Appl. 2017;
4. Isken MW, Rajagopalan B. Data mining to support simulation modeling of patient flow in hospitals. J Med Syst. 2002;
5. B R.E., M J.M., S J.L., Z A.W., D-B C.M. Impact of length of stay for first psychiatric admissions on the ratio of readmissions in subsequent years in a large Brazilian catchment area. Soc Psychiatry Psychiatr Epidemiol. 2016;
6. Miyoshi NSB, De Azevedo-Marques JM, Alves D, De Azevedo-Marques PM. An eHealth Platform for the Support of a Brazilian Regional Network of Mental Health Care (eHealth-Interop): Development of an Interoperability Platform for Mental Care Integration. JMIR Ment Heal [Internet]. 2018 Dec 7 [cited 2020 Mar 22]; 5(4):e10129. Available from: http://mental.jmir.org/2018/4/e10129/
7. Douzenis A, Seretis D, Nika S, Nikolaidou P, Papadopoulou A, Rizos EN, et al. Factors affecting hospital stay in psychiatric patients: The role of active comorbidity. BMC Health Serv Res. 2012;
8. Sockalingam S, Alzahrani A, Meaney C, Styra R, Tan A, Hawa R, et al. Time to Consultation-Liaison Psychiatry Service Referral as a Predictor of Length of Stay. Psychosomatics. 2016;
9. Masters GA, Baldessarini RJ, Öngür D, Centorrino F. Factors associated with length of psychiatric hospitalization. Compr Psychiatry. 2014;
10. Tulloch AD, Fearon P, David AS. Length of stay of general psychiatric inpatients in the United States: Systematic review. Adm Policy Ment Heal Ment Heal Serv Res. 2011;
11. L C.-L., L P.-H., C L.-W., L S.-J., M N.-H., L S.-F., et al. Model-based Prediction of Length of Stay for Rehabilitating Stroke Patients. J Formos Med Assoc. 2009;
12. Stevens A, Hammer K, Buchkremer G. A statistical model for length of psychiatric in-patient treatment and an analysis of contributing factors. Acta Psychiatr Scand. 2001;
13. Fekadu A, Desta M, Alem A, Prince M. A descriptive analysis of admissions to Amanuel Psychiatric Hospital in Ethiopia. Ethiop J Heal Dev. 2007;
14. Zhang J, Harvey C, Andrew C. Factors associated with length of stay and the risk of readmission in an acute psychiatric inpatient facility: A retrospective study. Aust N Z J Psychiatry. 2011;
15. Addisu F, Wondafrash M, Chemali Z, Dejene T, Tesfaye M. Length of stay of psychiatric admissions in a general hospital in Ethiopia: A retrospective study. Int J Ment Health Syst. 2015;

16. Newman L, Harris V, Evans LJ, Beck A. Factors Associated with Length of Stay in Psychiatric Inpatient Services in London, UK. Psychiatr Q. 2018;

17. Barekatain M, Maracy MR, Hassannejad R, Hosseini R. Factors Associated with Readmission of Patients at a University Hospital Psychiatric Ward in Iran. Psychiatry J. 2013;

18. B F.L.C., DR N.S., F M.P. Predictors of length of stay in an acute psychiatric inpatient facility in a general hospital: A prospective study. Rev Bras Psiquiatr. 2018;

19. T A.D., K M.R., F P., D A.S. Associations of homelessness and residential mobility with length of stay after acute psychiatric admission. BMC Psychiatry. 2012.

20. Mansour MR, Delbem ACB, Alberto LFC, Ramos RA. Integrating hierarchical clustering and pareto-efficiency to preventive controls selection in voltage stability assessment. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2015.

21. Yoshiura VT, Azevedo-Marques JM, Rzewuska M, Vinci ALT, Sasso AM, Miyoshi NSB, et al. A web-based information system for a regional public mental healthcare service network in Brazil. Int J Ment Health Syst. 2017;

22. Jensen F V. Bayesian networks basics. AISB Q. 1996;

23. Ahmadi-Javid A, Jalali Z, Klassen KJ. Outpatient appointment systems in healthcare: A review of optimization studies. European Journal of Operational Research. 2017.

24. Wei LJ, Glidden DV. An overview of statistical methods for multiple failure time data in clinical trials. Stat Med. 1997;

25. Y H.P., M A. Performance of joint modelling of time-to-event data with time-dependent predictors: An assessment based on transition to psychosis data. PeerJ. 2016; https://doi.org/10.7717/peerj.2582 PMID: 27781169

26. Ihwah A. The Use of Cox Regression Model to Analyze the Factors that Influence Consumer Purchase Decision on a Product. Agric Agric Sci Procedia. 2015;

27. Therneau TM. A Package for Survival Analysis in S. Version 2.38. CRAN website—http//cran.r-project.org/package=survival. 2015;

28. The R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing Vienna Austria. 2012.

29. Archer GEB, Saltelli A, Sobol IM. Sensitivity measures, anova-like techniques and the use of bootstrap. J Stat Comput Simul. 1997;

30. Vrieze SI. Model selection and psychological theory: A discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). Psychol Methods. 2012;

31. Anisimova M, Gascuel O. Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. Syst Biol. 2006;

32. Berger JO, Pericchi LR. The intrinsic bayes factor for model selection and prediction. J Am Stat Assoc. 1996;

33. Rissanen J. A Universal Prior for Integers and Estimation by Minimum Description Length. Ann Stat. 1983;

34. Akaike H. Statistical predictor identification. Ann Inst Stat Math. 1970;

35. MacCallum R. Specification Searches in Covariance Structure Modeling. Psychol Bull. 1986;

36. Toutenburg H. Draper, N., and H. Smith: Applied regression analysis. John Wiley & Sons, New York 1966. 407 S., 43 Abb., 2 Tab., 180 Literaturangaben, Preis: s 90. Biom Z. 1969;

37. Abakarov A, Sushkov Y, Mascheroni RH. A multi-criteria optimization and decision-making approach for improvement of food engineering processes. Int J Food Stud. 2013;

38. J R.E., L R.M., M M., D A. Observed-predicted length of stay for an acute psychiatric department, as an indicator of inpatient care inefficiencies. Retrospective case-series study. BMC Health Serv Res. 2004;

39. Barnow S, Linden M, Schaub RT. The impact of psychosocial and clinical variables on duration of inpatient treatment for depression. Soc Psychiatry Psychiatr Epidemiol. 1997;

40. Huntley DA, Cho DW, Christman J, Csernansky JG. Predicting length of stay in an acute psychiatric hospital. Psychiatr Serv. 1998;

41. Øiesvold T, Saarento O, Sytema S, Christiansen L, Göstas G, Lönnerberg O, et al. The Nordic Comparative Study on Sectorized Psychiatry—Length of in- patient stay. Acta Psychiatr Scand. 1999;

42. Hachesu PR, Ahmadi M, Alizadeh S, Sadoughi F. Use of data mining techniques to determine and predict length of stay of cardiac patients. Healthc Inform Res. 2013;

43. Pauselli L, Verdolini N, Bernardini F, Compton MT, Quartesan R. Predictors of Length of Stay in an Inpatient Psychiatric Unit of a General Hospital in Perugia, Italy. Psychiatr Q. 2017;

44. D J., G P., P G., M M. Classifying psychiatric inpatients: seeking better measures. Med Care. 1999;

45. Scheytt D, Kaiser P, Priebe S. [Duration of treatment and case cost in different inpatient psychiatric facilities in Berlin]. Psychiatr Prax. 1996