Universidade de Aveiro
Ano 2022

MIGUEL ALEXANDRE GARCIA CABRAL

DESENVOLVIMENTO E ANÁLISE DE UM SISTEMA DE RECOMENDAÇÃO PARA SUGESTÃO DE ARTIGOS MÉDICOS NO INTERNAMENTO DO HOSPITAL DA LUZ LISBOA

DEVELOPMENT AND EVALUATION OF A RECOMMENDATION SYSTEM TO SUGGEST MEDICAL ITEMS FOR INPATIENTS IN HOSPITAL DA LUZ LISBOA

Universidade de Aveiro
Ano 2022

MIGUEL ALEXANDRE GARCIA CABRAL

DESENVOLVIMENTO E ANÁLISE DE UM SISTEMA DE RECOMENDAÇÃO PARA SUGESTÃO DE ARTIGOS MÉDICOS NO INTERNAMENTO DO HOSPITAL DA LUZ LISBOA

DEVELOPMENT AND EVALUATION OF A RECOMMENDATION SYSTEM TO SUGGEST MEDICAL ITEMS FOR INPATIENTS IN HOSPITAL DA LUZ LISBOA

Relatório de estágio apresentado à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Estatística Médica, realizada sob a orientação científica do Professor Doutor Luís Miguel Almeida da Silva, Professor Auxiliar do Departamento de Matemática da Universidade de Aveiro, e sob supervisão do Doutor Nuno André da Silva, membro da equipa da Hospital da Luz Learning Health.

o júri

Presidente	Professor Doutor Pedro Miguel Ferreira de Sá Couto Professor Auxiliar da Universidade de Aveiro
Arguente	Professora Doutora Pétia Georgieva Professora Associada da Universidade de Aveiro
Orientador	Professor Doutor Luís Miguel Almeida da Silva Professor Auxiliar da Universidade de Aveiro

agradecimentos

Quero agradecer a todas as pessoas envolvidas direta e indiretamente no meu percurso ao longo deste ano. Sem elas esta experiência não teria sido a mesma.
Ao Professor Luís por ter aceitado o desafio de me orientar neste projeto e pela sua disponibilidade.
Aos Doutores Nuno Silva e José Moreira, da equipa da Learning Health do Hospital da Luz Lisboa, e ao Doutor Miguel Fiel pelo constante acompanhamento e contribuição para a minha evolução durante o estágio curricular.
À Professora Vera pela sua incansável dedicação e preocupação pelos seus alunos e por me ter ajudado a encontrar este estágio espetacular que me permitiu seguir um percurso mais feliz.
Aos meus colegas e amigos do mestrado, em especial à Leonor e à Sofia, pela sua paciência, desabafos, partilha de ideias e por todos os jantares.
À Àna, Inês e Rita por perpetuarem os “Felices los 4” e fazerem de Aveiro a minha casa.
Aos amigos de longa data de Vila Real que embora a distância, estão sempre lá.
Ao João por todo o apoio, atenção e preocupação que muitas vezes me fizeram ultrapassar as dificuldades que surgiram da melhor maneira.
Aos meus primos, tios e avós, em especial à minha Avó Isabel, por toda a companhia e amparo durante o período em que estive presencialmente no Hospital da Luz Lisboa. Tornaram esta aventura muito mais fácil.
À minha irmã, Ana, que apesar de ser chata, não pensa duas vezes em me ajudar independentemente do tipo situação.

E, por fim, um agradecimento muito muito grande aos meus pais que apesar de todas as adversidades, fazem de tudo com muito amor e carinho para eu conseguir atingir as minhas ambições. Quero deixar-vos orgulhosos e que saibam que não há palavras para descrever o quão importante é o vosso apoio.

palavras-chave

Sistemas de recomendação, filtragem colaborativa, domínio da saúde, artigos médicos, internamentos.

resumo

Os avanços da internet têm aumentado a quantidade de informação disponível, pelo que o seu excesso tende a dificultar a capacidade de gestão e filtragem da mesma. Este fenómeno pode ser observado no domínio da saúde onde a digitalização dos serviços médicos levou a um aumento substancial dos dados registados nos hospitais. Com o intuito de ajudar os profissionais de saúde a integrar toda a informação e, assim, realizarem decisões eficientes e efetivas, vários sistemas de recomendação (SR) foram desenvolvidos. Neste projeto, propõe-se um SR preliminar baseado em filtragem colaborativa para reduzir o tempo despendido pelos profissionais de saúde do Hospital da Luz Lisboa no registo de artigos médicos consumidos durante o período de internamento de doentes. Para isso, o SR foi desenvolvido de modo a formular recomendações relativamente aos artigos médicos e respetivas quantidades necessárias para o primeiro dia de internamento de um doente. A construção do SR teve por base os diagnósticos, procedimentos cirúrgicos e registos de consumos de artigos médicos associados a propostas cirúrgicas de doentes que foram internados no período de um ano no Hospital da Luz Lisboa. O conjunto de dados foi filtrado, reestruturado e analisado (N = 5088 propostas cirúrgicas), para posteriormente ser dividido em conjuntos de treino e de teste (75-25%). Foi aplicada uma *4-fold cross-validation* sobre o conjunto de treino para a afinação dos hiperparâmetros do algoritmo, sendo o SR foi testado e avaliado relativamente às suas recomendações a nível global e em cada especialidade médica do hospital em termos de *accuracy*, *classification performance* e *coverage.* Foi igualmente avaliado o grau de confiança no SR por parte dos profissionais de saúde do hospital. O SR apresentou uma performance global razoável (precisão = 0.608, sensibilidade = 0.729, F1 = 0.663, RMSE = 6.901) e demonstrou diferentes níveis de qualidade de recomendações dependendo da especialidade médica. Os melhores valores de precisão, sensibilidade e F1 foram observados nas previsões dos artigos médicos mais frequentemente registados, que correspondem a cerca de 85% dos consumos feitos no primeiro dia de internamento dos doentes do conjunto de teste. O algoritmo nunca sugeriu aproximadamente 80% dos artigos médicos utilizados no conjunto de teste, no entanto, estes apenas correspondiam 5.57% dos consumos totais. Por fim, e embora do ponto de vista dos enfermeiros do hospital haja alguma confiança nos resultados do SR, foram dadas sugestões para futuros ajustes do algoritmo. Não obstante as limitações do SR, os resultados obtidos representam um ponto de partida para o desenvolvimento de uma ferramenta de apoio aos profissionais de saúde do Hospital da Luz nos registos dos artigos médicos necessários durante o internamento de doentes.

keywords Recommendation systems, collaborative filtering, health domain, medical items, inpatients.

abstract

The internet advances have led to an increase of data and information availability. This overload of information tends to compromise the capacity to manage and filter the available data. In the health domain, the increasing digitalization in healthcare led to a substantial rise of the recorded data. Various recommendation systems (RS) have been developed to help healthcare professionals integrate all information and make efficient and effective decisions. Here, a preliminary RS based on collaborative filtering is proposed to reduce the time that healthcare professionals spend in registering medical items consumed during patients' hospitalization. For that purpose, the RS was built to perform suggestions of the medical items and respective quantities needed in the first day of hospitalization of a patient. Data regarding the diagnostics, surgical procedures and medical item records associated to surgeries of inpatients during a period of one year in Hospital da Luz Lisboa was filtered, restructured, and analysed (N = 5088 surgeries) for the construction of the RS. A 75-25% split of the data was considered with a 4-fold cross-validation procedure applied on the train set to tune the hyperparameters settings for the algorithm. The RS was then tested and evaluated regarding its overall performance in terms of accuracy, classification performance, and coverage. The same measures were applied to assess the quality of the recommendations for each medical specialty of the hospital. Furthermore, the trust of healthcare professionals in the RS was also assessed. A moderate overall performance was achieved (precision = 0.608, recall = 0.729, F1-Measure = 0.663, RMSE = 6.901) and the quality of the algorithm's recommendations varied between medical specialties. Additionally, the algorithm presented higher values of precision, recall and F1-Measure in the predictions of the most frequently registered medical items in the test set, which corresponded to approximately 85% of the consumptions in the first day of hospitalization. Regarding the coverage of the RS, approximately 80% of the medical items used in the test set were never recommended by the algorithm, corresponding to only 5.57% of the consumptions. Lastly, although in the point of view of hospital's nurses there is some trust in the RS results, several suggestions were given for further improvements of the algorithm. Despite the limitations of the RS, the observed results represent a starting point for the development of a tool that can support healthcare professionals of Hospital da Luz Lisboa in registering medical items needed during inpatients' hospitalization.

Index

List of Figures

List of Tables

1. Introduction

1.1. Internship Contextualization and Hosting Entity

The present report arises from the internship carried out within the scope of the program of the master's degree in Medical Statistics from the University of Aveiro. The internship was prosecuted at Hospital da Luz Learning Health between 1st January 2022 and 31st July 2022.

Luz Saúde is a large private health care group founded in Portugal in 2000, whose mission is to achieve the best healthcare outcomes for their patients by executing effective and efficient diagnoses and treatments [1]. It comprises 29 different units (14 private hospitals, 13 private outpatient clinics and 2 senior residences) located between the north and central-south regions of Continental Portugal and in the Autonomous Region of Madeira [1].

Hospital da Luz Learning Health is the company within Luz Saúde Group that aims to help in the mission of reaching medicine of excellence and innovate in different areas of health care [2]. It is based in the medical simulation center at Hospital Luz Lisboa and works in three main fields [2–4]:

- **Training** – Planning, development and evaluation of the training and simulation programs for healthcare professionals.
- **Research** – Development, promotion, dissemination, and support on the research carried out in several areas such as computational medicine, imaging, human factors engineering, and clinical research.
- **Innovation –** Provision of resources for the development of new products and services focusing the improvement of the health care of Luz Saúde clients.

The internship was carried out in the Data Science research team. This area is committed in developing models for disease characterization, treatment personalization, and healthcare delivery improvement, using data mining and machine learning methods [4].

1.2. Basics of Recommendation Systems

1.2.1. Basic Concepts

The growing advances of the internet have led to an increase of data and information availability [5–8]. Even though this allows the users to have a wider freedom of choice, the overload of information tends to overwhelm them. Consequently, user's capacity of managing and filtering information can be compromised, leading them to make non optimal decisions [5,6,8]. Recently, there has been a great investment to solve this problem in the Artificial Intelligence (AI) community [5] by developing algorithms, such as recommendation systems, that perform personalized suggestions according to the user's preferences and/or necessities aiming to imrpove decisions [5,6,8]. More specifically, recommendation systems are tools and methodologies that elaborate lists of ranked items according to a given information, subsequently used as the best suggestions for a specific user [5–8].

Recommendation systems are used in several domains, such as e-commerce [6,7], entertainment [6,7], healthcare [9], etc. There are three basic concepts that are transversal to all recommendation systems: domain, users, and items [6–9]. The *domain* corresponds to the environment in which users and items interact, i.e., the set of factors that will induce a recommendation of a certain item to a specific user. *Users* are the individuals to whom the suggestions will be made. Usually, each user is characterized by a "user's profile", i.e., a group of features that are important to the recommendation process. Finally, *items* are the elements that will be suggested to a user [6–9]. A recommendation system usually makes recommendations of a specific type of item, which defines the design and methodology of the algorithm so that the suggestions are performed effectively [8]. In brief, a recommendation system generates suggestions of potentially relevant items considering the information about users, items, and their interactions.

The first recommendation system emerged in the mid-1990s [10], and the popularity has been increasing since then [8,11]. For instance, several well-known internet sites use these methodologies as customized services for their subscribers/customers (e.g., Netflix, Amazon.com, YouTube, etc.) [7,12–14]. Moreover, conferences and workshops regarding this research area have been arranged in the past years [8]. For example, the premier international

forum for the discussion of recommendation systems research - *ACM Recommendation systems Conference* - takes place annually since 2007 for the presentation of new results, systems, and techniques [8,15]. There has also been a growing interest of recommendation systems as an academic field given the number of journals that published special issues covering this matter [8]. Among them are *AI Communications (2008)* [16], *IEEE Intelligent Systems (2007)* [17] and *Intelligent Decision Technologies (2015)* [18].

1.2.2. Recommendation Techniques

A recommendation system is a tool that makes use of data to formulate recommendations for its users. The data's nature defines which recommendation technique is more adequate to explore the information and, consequently, compute relevant recommendations [6–8]. In most cases, data is mainly about the group of items that are available to be recommended, however, it can vary regarding its complexity [6–8]. Throughout this topic, the traditional concepts of three commonly implemented recommendation methodologies will be explained (collaborative filtering, content-based filtering and knowledge-based recommendations).

1.2.2.1. Collaborative Filtering

The collaborative filtering (CF) approach is considered the most popular and commonly used method to build recommendation systems [6–8,19], since it does not require domain knowledge [6]. The CF technique is based on the simple assumption that "If users shared the same interests in the past, then they would have similar tastes" (Figure 1) [9]. Thus, in most CF-based recommendation systems (i.e., in the e-commerce or entertainment domains), the prediction of a rating given by a user u to an item i, which will determine if the item is going to be recommended or not, is based on the ratings given by other users v to the same item [6,8,19,20]. The higher the similarity between v and u's ratings, the more likely they are to give similar ratings to item i[6,8,19,20]. Here, the term "rating" can be interpreted as the interaction between the user and the item [19], as this measure is domain specific.

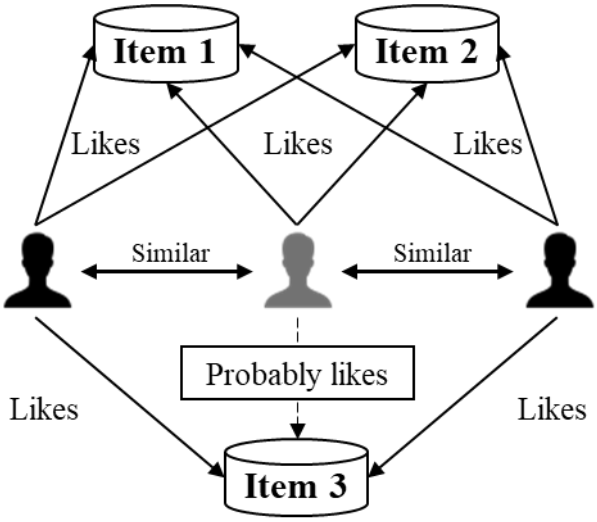


Figure 1: *Schematic representation of collaborative filtering methods.*

CF can be grouped in two different methodologies depending on the way the ratings are used to execute the predictions and recommendations: neighborhood-based CF or model-based CF. The first one uses the existing ratings directly in the estimation of ratings for a new item [21,22], while the latter aims to train predictive models using the user-item interactions [23,24].

Neighborhood-based Collaborative Filtering

The neighborhood-based CF is a generalization of the nearest neighbors' classifiers [22] and has two distinct approaches (user-based and item-based) to predict the rating that a user u would give to an item i [19,21,24,25]. Both require the computation of a neighborhood which corresponds either to a group of users that have similar patterns of rating as u and have rated item i[20–23,25,26], $V(u)$, or a set of items that were already rated by u and were consistently given similar ratings as i by other users v [19,21,22,25,27,28], $J(u)$. Independently of the used method, the neighborhood-based CF follows three essential steps [20–22,26]:

1. Provide the user's profile (i.e., ratings) to the recommendation system;
2. Select the neighborhood either for user u or item i by computing similarity metrics, which quantify the association between users or items (please see "Recommendation Algorithm" in the Methods section);

3. Estimate u's rating to i, using the neighborhood's ratings and similarity scores, which will determine if i should be recommended to u.

Step 1 can be represented by a user-item matrix where users are represented in the rows, items are represented in the columns and each entry of the matrix consists in the rating r_{ui} that a user u gave to an item i (see Table 1) [21,22,29]. The rating matrix is usually composed by sparse data (i.e., with missing entries or zeros spread in the matrix), which is what allows to compute the neighborhood considering either the rows or columns similarity, in contrast to the nearest neighbors' classifiers that only consider the similarity between rows [22].

Table 1: *Ratings given by three users to five items. Users and items are represented in the rows and columns, respectively. The scale includes values between 1 and 5.*

	Item 1	**Item 2**	**Item 3**	**Item 4**	**Item 5**
User 1	5	5	1	4	3
User 2	r_{ui}	4	3	5	3
User 3	3	2	5	3	1

The r_{ui}, in a user-based recommendation, is estimated using the ratings from users $v \in V(u)$ [19–22,25,26,29], taking into account that such users may have different levels of similarity with u [19,21,22]. Thus, similarity scores not only must be used to define the u's neighborhood, but also incorporated in the rating prediction in order to weight the contribution of users $v \in V(u)$ [19,21,22]. Similarity scores between user u and other users $v \in V(u)$ are computed row wise and r_{ui} can be estimated by the following expression [19,21,22,30]:

$$\hat{r}_{ui} = \frac{\sum_{v \in V(u)} W_{uv} \times r_{vi}}{\sum_{v \in V(u)} |W_{uv}|}$$

where W_{uv} represents the similarity score between user u and user v and r_{vi} is the rating given by user v to item i. The sum of the absolute values of the similarity scores in the denominator guarantees that the predicted values do not surpass the stipulated rating scale [21,22].

On the other hand, the item-based approach predicts r_{ui} by using the ratings that u gave to items $j \in J(u)$ [19,21,22,25,27,28], being the similarity scores now computed column wise

[21,22]. Here, W_{ij} is the similarity score between item i and item j while r_{uj} represents the rating given by user u to item j.

$$\hat{r}_{ui} = \frac{\sum_{j \in J(u)} W_{ij} \times r_{uj}}{\sum_{j \in J(u)} |W_{ij}|}$$

Several variations of the aforementioned expression (Table 2) were developed to adapt the method to different scenarios and, consequently, achieve the best recommendations [21,22,31].

***Table 2**: Variations of the neighborhood-based CF* [21,22,31].

	Expression	Observation
Mean Centering	$\hat{r}_{ui} = \bar{r}_u + \frac{\sum_{v \in V(u)} W_{uv} \times (r_{vi} - \bar{r}_v)}{\sum_{v \in V(u)} \|W_{uv}\|}$	$\bar{r}_u$ and $\bar{r}_v$ correspond to the average ratings of user u and v, respectively;
Z-score Normalization	$\hat{r}_{ui} = \bar{r}_u + \sigma_u \frac{\sum_{v \in V(u)} W_{uv} \times (r_{vi} - \bar{r}_v)/\sigma_v}{\sum_{v \in V(u)} \|W_{uv}\|}$	σ_u and σ_v correspond to the ratings' standard deviation of user u and v, respectively;
Baseline	$\hat{r}_{ui} = \bar{r}_u + \frac{\sum_{v \in V(u)} W_{uv} \times (r_{vi} - b_{vi})}{\sum_{v \in V(u)} \|W_{uv}\|}$	b_{vi} (bias) is a factor that is independent from the user-item interaction and has an effect on the ratings (e.g. "systematic tendencies for some users to give higher ratings than others").

Note: These expressions are related to the user-based method, however, they can also be adapted for the item-based approach.

Advantages and Limitations

Neighborhood-based methods present three main advantages when compared to other recommendation techniques: "simplicity", since their implementation is relatively

intuitive; “justifiability”, due to the easy and concise explanations that can be given to justify a certain recommendation; and “efficiency”, because of their low computational cost [21].

On the other hand, neighborhood-based recommendation systems also have some limitations. When it comes to dealing with substantial amounts of users and items, memory problems may arise [6,21,22,29,32]. Consequently, a reduction of the used information is often needed so that the neighborhood-based methods become usable. This can be achieved by filtering users’ neighborhoods to restrict the number of similarity scores considered in the recommendations, while also avoiding the inclusion of noise from data that comes from weak relations between users [21]. Top-*k* and threshold filtering are two ways of reducing users’ neighborhoods. The first one consists in choosing only the *k* higher similarity scores [21,30,32], while the latter excludes cases in which similarity scores are below a certain threshold [21,30]. Both *k* and the similarity threshold should be carefully defined since they also influence the algorithm’s performance.

Additionally, neighborhood-based methodologies are sensitive to sparse data [20,21,33]. Usually, users only interact with a small part of the items that are available [6,21,33], leaving up to 99% (on average) of the user-item matrix empty [6]. Hence, in some cases the probability of two users or items being on each other’s neighborhood is low due to the unlikeliness of having common interactions[20,21]. Accordingly, not only the neighborhoods sizes are limited, but also the similarity scores are calculated with a restricted number of ratings, possibly leading to biased recommendations [21]. Dimensionality reduction approaches [34–36] were developed to solve the problems of data sparsity, consisting essentially in “projecting users and items into a reduced latent space that captures their most salient features”, which decreases the recommendations’ vulnerability to sparsity [21,34].

Lastly, item popularity strongly influences the neighborhood-based recommendation systems’ ability to recommend all available items [21,22,37–39]. In fact, in several cases, most user-item interactions are associated with a small fraction of popular items, being the data regarding the remaining interactions sparse [21,22,37–39]. Therefore, user-item interactions distribution (see Figure 2) often divides items into two groups: the popular items (short-head of the distribution) and the unpopular items (long-tail of the distribution) [21,22,37–39]. Since long-tail items are involved in a low percentage of interactions, recommendation systems do

not have access to enough data, being harder to recommend them – *Long Tail Problem* – in contrast to short-head items [21,22,37–39].

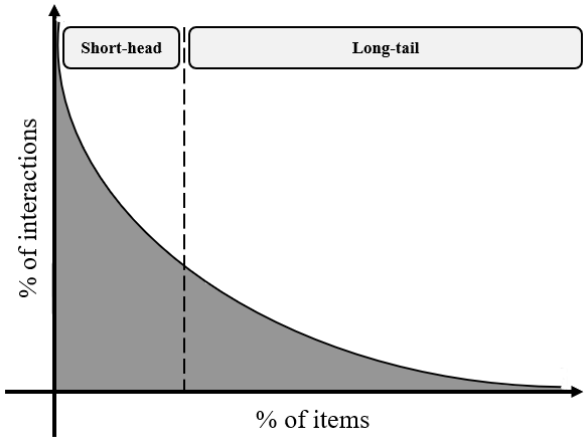


Figure 2: *Distribution of the user-item interactions.*

Model-based Collaborative Filtering

Model-based methods rely on data mining processes to extract latent factors (i.e., unobserved variables that can explain the dependence between other variables [40]) that characterize the user-item interaction patterns, which are used as parameters in the predictive models [23,24,33,41]. Machine learning methods are commonly applied to build recommendation algorithms and fit them to the available data, in order to predict the missing user-item interactions [23,24,33,41]. Several approaches were developed such as Matrix Factorization Models (MFM) [42], Co-clustering [43], Markov decision process [44,45], etc. To exemplify, a brief description of the MFM method is disclosed below.

Matrix Factorization Models [23,29,42,46]

MFM aim to explain user-item interactions by identifying latent factors in the user-item interaction matrix. Thus, the resulting latent space tries to characterize item and users' factors (inferred from users' feedback).

In practice, an item i is associated with a vector q_i whose elements quantify the association (negative or positive) between item i and the aforementioned factors. On the other hand, a user u is associated with a vector p_u whose elements quantify the interest of

user u in items that are highly associated with the corresponding factors. The user-item interaction ("overall interest of the user in characteristics of the item") is obtained by the dot product between both vectors:

$$q_i^T p_u = \sum_{k=1} q_{ik} \times p_{uk}$$

where k represents the index of the element of each vector.

Baseline predictors, also called bias (i.e., factors that are independent from the user-item interaction and influence the ratings) of users (b_u) and items (b_i) can be added to this expression in order to estimate the rating given by a user to an item:

$$\hat{r}_{ui} = \bar{r}_u + b_i + b_u + q_i^T p_u$$

where $\bar{r}_u$ is the average ratings of user u.

Finally, it is necessary to minimize a regularized squared error to learn the model parameters, which is usually performed by applying methodologies such as alternating least squares or stochastic gradient descent.

Advantages and Limitations

Model-based approaches emerged to solve some of the limitations inherent to neighborhood-based methodologies [29]. Specifically, they tackle memory problems and data sparsity by reducing the user-item interaction matrix to a latent space [23,29]. However, the complex training processes may entail a heavy computational cost [23,24,33]. Additionally, these methods are more limited regarding their ability to cover a wide range of users compared to the neighborhood-based approaches [33].

1.2.2.2. Content-based Filtering

Content-based filtering (CBF) is a recommendation method that predicts user-item interactions by exploring the items' characteristics (attributes) and assessing their relevance to the user [41,48–52]. For that purpose, two sources of data are needed: a description of each item defining its set of attributes and the users' profile of interactions [50–52]. Therefore, the

user-item interactions are predicted by comparing an item's set of attributes with the characteristics of other items that the user has already interacted with (Figure 3) [48–52].

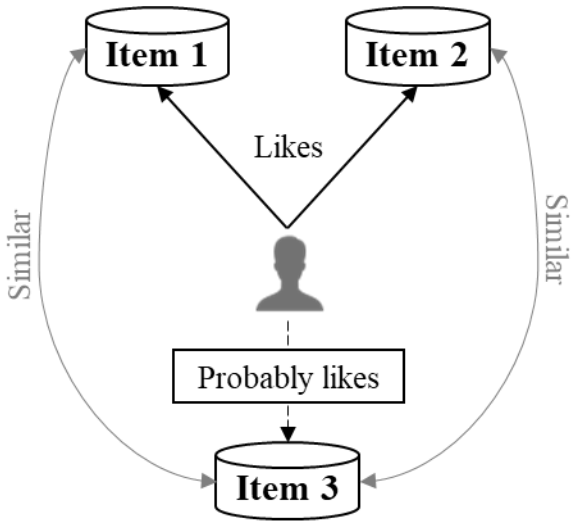


Figure 3: *Schematic representation of content-based filtering methods.*

Items' attributes can be represented in a structured way where each attribute takes a value from a known set of options (e.g. a movie can be described by the genre, year, cast, etc.) or in textual formats in which the attributes do not have well-defined values (e.g., synopses of a movie) [50–53]. In the first case, machine learning methods can easily use the data to learn a user's profile and make recommendations [50]. On the other hand, unstructured data must be carefully prepared since semantic problems such as polysemy (i.e., "presence of multiple meaning for one word") or synonymy (i.e., "multiple words with the same meaning") interfere with the quality of the recommendations [50,52,53].

Different approaches can be applied to compute the similarities between items' attributes and therefore make recommendations. For instance, machine learning methods (e.g., Näive Bayes Classifiers, Decision Trees, etc. [41]) can learn the relationship between items and, together with the user-item interactions data, predict if a new item should be recommended or not to a certain user [41,50–52]. Nonetheless, most CBF recommendation systems resort in Vector Space Models (VSM), such as the Term Frequency-Inverse Document Frequency ($TF\text{-}IDF$). Here, to find item similarities, item attributes are represented in a vectorial space with as many dimensions as the words used to characterize an item [41,48–51]. This approach is particularly useful since most recommendation systems deal

with unstructured data. Thus, to build a CBF algorithm using a VSM, three essential steps have to be executed [50–52]:

1. Textual data (item's attributes) preprocessing;
2. Information extraction and conversion to a vectorial space;
3. Item similarity estimation.

Textual data preprocessing

The preprocessing step is crucial to mitigate and avoid eventual polysemy and synonymy problems [50,51]. Among other preprocessing procedures there are: removal of words that are very common in a language but are not related to the item (e.g. pronouns and determinants) [50,51], standardization of the variations of the same word/concept [50–52], and detection of groups of words that appear together in the text several times, whose meaning may be distinct from when they are considered individually [51].

Information extraction and conversion to a vectorial space

Item descriptions are extracted and converted to a vectorial space using $TF\text{-}IDF$ [41,48–52,54]. A vector is composed by the weight of each word representing the association with the description in which they are included [48–52,54] considering the following rationale: "terms that occur frequently in one document (TF = term-frequency), but rarely in the rest of the descriptions (IDF = inverse-document-frequency), are more likely to be relevant to the description" [50]. In this sense, *TF-IDF* measures the importance of a word in an item description by

$$TF\text{-}IDF = TF \times IDF$$

where TF and IDF are given by

$$TF = \frac{N_{wd}}{N_d}, \qquad IDF = log\frac{D}{D_w}$$

where N_{wd} is the frequency in which a word w appears in the description d of an item, N_d consists in the total number of words in description d, D corresponds to the total number of descriptions available and D_w is the number of descriptions with the word w.

Therefore, the importance of a word in an item's description increases with the frequency in which it appears in the description but decreases if it is present in an increased number of different descriptions [49,54].

Item similarity estimation

Similarity measures are needed to estimate if two items are related to each other, so that the algorithm determines whether a recommendation should be formulated or not. In this case, both the new item and the items that the user already interacted with are represented as vectors. Thus, cosine similarity metric (please see "Recommendation Algorithm" in the Methods section) is widely used when $TF\text{-}IDF$ is applied [6,49,50].

Advantages and Limitations

CBF techniques allow to perform recommendations of items depending only on the items that the user in question has interacted with [6,41,50,51]. Thus, since recommendations do not require the computation of large similarity matrices, CBF systems can manage many users [6,51]. Additionally, these systems have the ability to comfortably recommend novel items that do not present interactions with any user [41,50,51]. Lastly, suggesting an item can be easily explained by highlighting the attributes that led to the recommendation [41,50,51].

On the other hand, besides the fact that CBF approaches usually need domain knowledge, they have an intrinsic limit of content data that can be associated to an item [50]. Consequently, if a recommendation algorithm does not include enough information to distinguish if items are interesting/necessary to the user or not, then it cannot provide reliable suggestions [6,41,50]. Over-specialty (i.e., only recommendations of items similar to what the user already knows are performed), which restricts the systems' degree of novelty, is also a limitation of CBF systems [6,41,50]. Moreover, recommendations are bounded to the previous user-item interactions, being difficult to propose items to new users to the system [50,51].

1.2.2.3. Knowledge-based Recommendations

From the three recommendation techniques here discussed, only the knowledge-based (KB) methodologies are strictly dependent of domain knowledge [6,56–59]. Explicit domain data regarding the users and items features are used to recognize the user's conditions and find items that can respond to the user's needs [6,56–59].

There are two main approaches to perform KB recommendations: case-based [58,60] and constraint-based [56,57,59]. Although both methods are equivalent in terms of knowledge sources usage (e.g., collection of the users' requirements, suggestion of adjustments in case that a solution is not found, etc.), the way each one computes the recommendations differs from the other [56–60]. For instance, the former makes suggestions relying on similarity measures, while the latter depends on a set of rules that have to be fulfilled to find a solution [56–60].

Case-based Recommendations

Case-based recommendations emerged after the case-based reasoning (CBR) which aimed to solve users' requirements/problems[58,60]. CBR formulates suggestions by comparing a user's situation with other users' problems that were previously solved [58,60]. Each case is divided into two major steps which are the *specification*, where the problem is described, and the *solution,* where the recommendation used to solve the problem is described [58]. In fact, CBR algorithms find the solution by sequentially receiving the user's problem, comparing it to previous cases and selecting the most similar ones, extracting the solutions used in those cases and adapting them to compute a new solution [58,60]. Here, the item knowledge is crucial to the adaptation of the set of possible solutions in order to match the user's problem specification [58].

Although, nowadays, case-based recommendation systems use the core principles of the CBR, they work in a way that resembles CBF systems. Nonetheless, case-based approaches aim to find items that have features similar to what the user is searching for, instead of trying to find items that are similar to items that the user already interacted with [58,60]. Furthermore, both methods differ in the way items are represented and in the similarity assessment between them [58,60]. In fact, in case-based methods data is mostly structured, in

contrast to CBF, which allows to resort to more complex similarity metrics that are based on domain knowledge. For instance, in CBF similarities are obtained by assessing the proximity of the terms used in item descriptions, while in case-based methods an item is considered similar to the users' requirements if the knowledge about that item's features are close to what the user pretends [58].

Constraint-based Recommendations

In this approach, the knowledge about users requirements/features and items characteristics are linked together by building rules on how to relate both [6,9,57,59]. Accordingly, recommendations obtained by the constraint-based methodology require five essential elements (Figure 4): two groups of variables (i.e., user properties and item properties) and three groups of constraints (user properties' constrains, item properties' constraints and filter constraints) [56,57,59].

- User properties: includes all variables that may represent user features/requirements;
- Items properties: set of characteristics that describe the items;
- User properties' constraints: set of constraints that define the compatibility or incompatibility between user properties;
- Item properties' constraints: set of constraints that define which alternatives that an item's characteristic may adopt;
- Filter constraints: rules that determine the relationship between users' features/requirements and items' characteristics.

The solution for a problem of this nature (recommendation) is obtained by ensuring that the user and item properties fulfill all constraints, while the recommended item is able to satisfy the user's requirements [56,57].

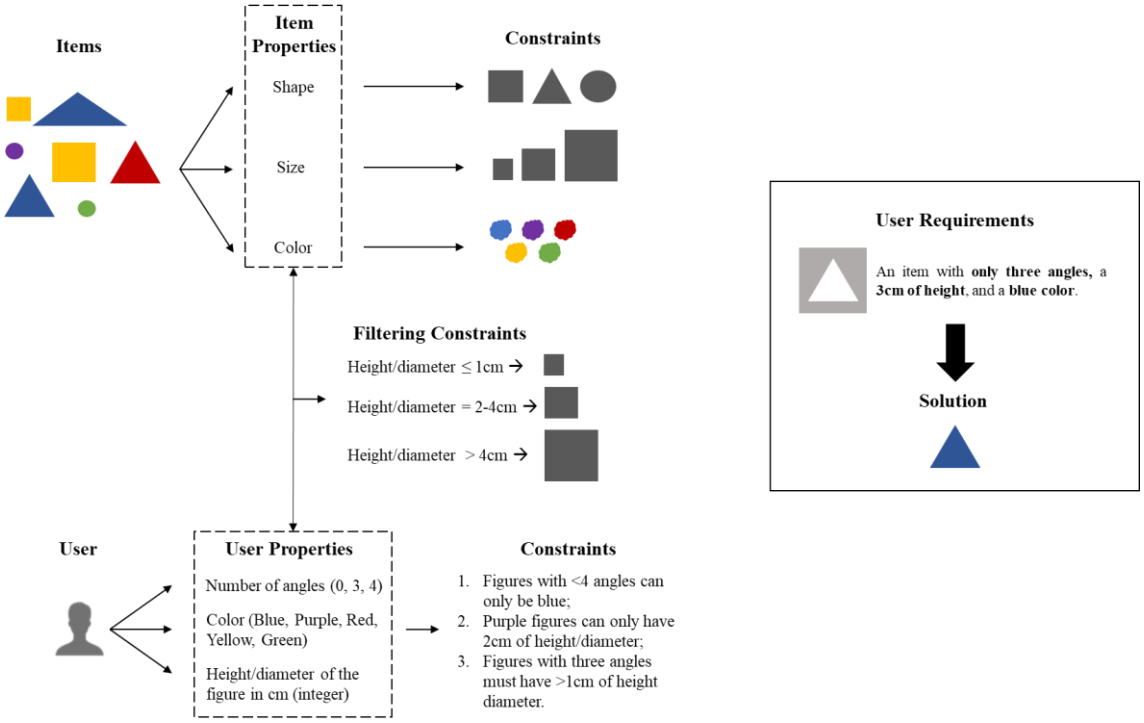


Figure 4: *Schematic representation of the essential elements of the constraint-based methods (item properties, user properties, item properties' constraints, user properties' constraints and filtering constraints) and their relationship.*

Advantages and Limitations

KB approaches are advantageous when compared to CF and CBF in terms of data usage. In contrast to interaction-dependent methods, KB algorithms solely require knowledge sources to formulate recommendations [6,57,59]. Therefore, the systems do not need previous users' data to increase the recommendations' quality since the users' requirements are provided directly to a single recommendation iteration [6,57,59]. In return, these methodologies experience the knowledge acquisition (i.e., "process of constructing rules and requirements needed for a knowledge-based system" [6]) bottleneck problem due to the lack of domain experts that have also the engineering "know-how" to convert their expertise into simple and usable representations [6,57,59]. Usually, knowledge engineers and domain experts have to work together, in a much harder way, to build functional knowledge bases [57].

1.3. Recommendation Systems in the Health Domain

The assumption underlying the recommendations formulated by the CF methodology in the health domain must be adapted to the scenario and nature of the data. Thus, health recommendation systems (HRS) perform predictions considering that "If patients share similar disease profiles/health conditions, then they would have similar treatment/health services" [9]. Taking that into consideration, the CF method in HRSs do not use the user-item interaction data to compute the association between users or find latent factors, since in the health domain the recommendation of items (e.g., treatments, diets, medication, etc.) depends on factors other than the interaction that the user had with previous items. In fact, most items in this domain are provided because of the users' features, which may be health conditions, diseases, body measurements, etc. Nonetheless, the user-item interaction data is also needed in the recommendation process. To illustrate, Stark, B. *et al.* [47] developed a recommendation system that uses the individual features (e.g., age, allergies, blood pressure, etc.) from a new migraine patient (user) to find similar migraine patients (neighborhood) whose data about their medication (items) is extracted to recommend a drug that fits best to the user's condition.

HRSs based on CBF methods recommend "healthcare services that fit patient's health conditions/disease situation and are similar to those assigned to him/her in the past" [9]. In the health domain, CBF is mostly implemented together with CF to build hybrid recommendation systems. For example, Aberg, J. [55] created a recommendation algorithm that proposes suitable meals to the elderly, resorting to both methodologies. The system not only relied on the similarity between users regarding their features (e.g., tastes, dietary restrictions, nutritional needs, etc) but also used the content data from the meals in the database (e.g., cost, preparation difficulty, nutritional properties, etc.) to execute adequate recommendations of food recipes.

Lastly, several recommendation systems were developed using health domain knowledge to link the user needs (requirements) and the corresponding healthcare services (e.g., prescriptions, treatments, diets, etc.). These tools mainly aimed to help healthcare professionals in the clinical decision-making [61–66].

Literature Review

With the increasing digitalization in healthcare, recorded health information availability has grown substantially, allowing patients and healthcare professionals to access a vast amount of data for patient-oriented decision-making processes [9,11,67]. This may have a positive impact on the desire of patients to inform themselves about their health status [68,69] and, consequently, improve the patient-doctor relationship leading to more dynamic and informed dialogues [11,70].

Additionally, electronic health records (EHRs) have been progressively adopted worldwide [71]. EHRs comprise varied health data about patients [11,71], which can be useful for the development of a personalized healthcare [67]. Nevertheless, due to time limitations and the overload of clinical data, healthcare professionals may find increasingly difficult to integrate all the information to make efficient and effective decisions [9,72,73].

HRSs are recommendation tools specialized in making suggestions of items from the health domain [8,9]. HRSs aim to solve the aforementioned problems by elaborating personalized and detailed recommendations that, consequently, will allow patients/users to understand better their medical condition and will help healthcare professionals in the clinical decision-making processes [9,11]. In this case, the user's profile corresponds to the medical history of a patient [9,11], which will be integrated in the HRS and provided with "medical facts" in order to recommend relevant items regarding the patient's health status [11]. However, the end-user of the system does not have necessarily to be a patient. It can be either a patient, a healthy person or a healthcare professional, which defines the type of item that will be recommended and the complexity of the recommendation [9,11].

Usually, HRSs that are focused on patients or healthy people as end-users tend to make suggestions of nutritional information [55,74,75], physical activities [63,75,76], healthcare services [77,78] or medication [75]. For example, Bankhele, *et al.* [75] created an android application that suggests diets, physical exercises, and medications to people with diabetes. Briefly, this HRS asks for a set of parameters from the end-user and matches that information with other users. Thus, the recommended items consist in diets, physical exercises, and medications of users similar to the end-user.

HRSs oriented to healthcare professionals use more complex health information to support in clinical decision-making processes [9,11] and, consequently, minimize related costs [9]. For instance, several HRSs were developed for early disease prediction [79–81], in response to the rise of chronic diseases worldwide [79]. Other types of diseases also promoted the construction of HRSs for diagnosis and disease management [66,82,83]. Depending on the nature of the suspected disease, different parameters are considered in building a HRS and make recommendations. Shen, Y *et al.* [66], for example, developed a decision support system that uses knowledge about infectious diseases, patient's symptoms, bacteria, syndromes, and drugs, to identify a potential infectious disease based on a patient's condition and afterwards recommend the most indicated antibiotic treatment.

Drug-related adverse events, as well as the possible hazard of unappropriated treatments have motivated the development of data-driven recommendation systems to assist healthcare professionals in making prescriptions. A considerable number of studies proposed HRSs for drug recommendation regarding specific conditions, such as infectious diseases [65,66], migraines [47], diabetes [64,82], cardiac diseases [83] and oncologic diseases [84,85]. In these cases, the HRSs generally require specific information about the patient's demographics and health condition, risk factors and drugs accepted for the treatment of a given disease. There are also recommendation systems like GalenOWL [61] and Panacea [62], developed by Doulaverakis, C., *et al.* in 2012 and 2014 respectively, that formulate drug suggestions without being restrictive to a type of disease. Both algorithms make recommendations taking into consideration a complex set of drug information (e.g. contraindications, recommended dosage, excipients, drug-drug reactions, etc.) in addition to the patient data (e.g. diseases, allergies, current medication, etc.) [61,62].

Medical orders are error prone since modern clinical practices suffer from undesirable variability [86,87], which may "compromise quality of care and cost efficiency" [87] and, consequently, represent a serious public health problem [88]. Thus, similarly to entertainment and e-commerce platforms, it is possible to predict and recommend clinical items in the health domain, in order to reduce the clinical practices' uncertainty [9] and workload [89]. Besides medication, medical orders may also include lab tests, radiologic imaging, nursing orders, etc. [87]. To illustrate, ClinicNet [86] and OrderRex [87] are clinical decision support tools that aim to predict the most adequate medical items for a particular condition. Although each HRS presents different underlying algorithms (i.e. neural networks

and näive bayes-based collaborative filtering, respectively), both make use of information from EHR to perform data-driven recommendations [86,87].

Lastly, there are studies that dedicated their efforts to reinforce clinical decisions regarding patient hospitalization and associated practices. As an example, Prasad, N. *et al*. [90] proposed a decision support tool that aids healthcare professionals to identify the best moment for mechanical ventilation weaning and corresponding sedation, considering patient data and different clinical parameters.

1.4. Objectives

The report of medical items that are consumed in healthcare facilities is a daily responsibility of healthcare professionals, a time-consuming task that increases their workload [89]. In Hospital da Luz Lisboa, the complexity of this task and the other demanding duties that healthcare professionals must perform, besides being overwhelming, may result in inconsistencies in the medical item registering process. For instance, errors and the absence of some medical item records may occur due to oversight or lack of time/resources. Concurrently, other departments of the hospital, such as the resources management, are also negatively affected by this problem since the data from the medical item records is directly used to calculate the hospital's resources availability.

Taking all into account, this work aims to be a starting point to mainly help healthcare professionals reducing the time they spend in the medical item registration process and to optimize the resource management in Hospital da Luz Lisboa. For that purpose, the following objectives were proposed: **1)** development of a baseline neighborhood-based recommendation algorithm, that exploits data from patients' records to predict medical items and respective quantities needed in the first day of hospitalization of patients that underwent an elective surgery; **2)** evaluation of the algorithm's recommendations and analysis of its limitations for future improvements; **3)** assessment of the healthcare professionals opinion and suggestions regarding the achievements in this phase of the project.

The subsequent chapters of this report will be organized as follows. Chapter 2 is devoted to describing the methodologies carried out during this work; Chapter 3 presents all

the results regarding the descriptive analysis of the data, as well as the evaluation of the proposed recommendation algorithm; Chapter 4 and Chapter 5 consist in the discussion and conclusions of the work, respectively; and Chapter 6 presents the final remarks.

2. Methods

2.1. Dataset

This work was conducted with an anonymized dataset comprising information from EHRs of 12057 patients that underwent at least one surgery at Hospital da Luz Lisboa during a period of 1 year. The dataset included 3 components: diagnostics, surgical procedures and medical items used in the inpatient stay. The first two components include the attributes (features) of the patients that underwent a given surgery. A single admission in the hospital (event) is associated with those components in a sequential relationship as depicted in Figure 5. All surgeries were considered independent, since distinct sets of surgical procedures lead to the consumption of different sets of items during the patients' hospitalization.

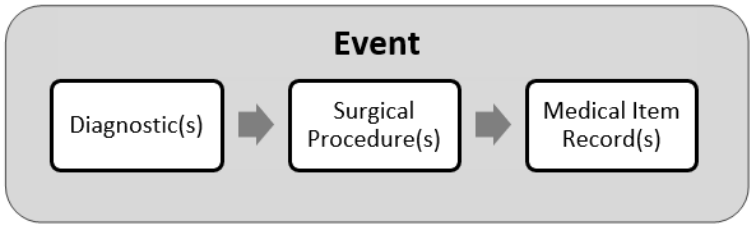


Figure 5: *Relationship between the three domains used in this project.*

Three initial databases (one for each component) were extracted. A brief description of those databases is shown in Table 1 from Appendix A.

The diagnostics database included the patient diagnostics that led to a surgery. Diagnoses were encoded according to the World Health Organization's (WHO) Ninth Revision of the International Classification of Diseases (ICD-9) [91], for a standardization of the medical conditions reporting.

The surgical procedures database contained information regarding the performed surgeries. Each surgery is identified with a unique code (*Número de Proposta*) and is allocated to one of thirteen medical specialties (*Especialidade*) (e.g., ophthalmology,

urology, etc.). Additionally, they are categorized depending on the urgency of the patient's intervention (*Grau de Prioridade*) as reported in the National Confidential Enquiry into Patient Outcome and Death (NCEPOD) classification [92] (Table 1, Appendix A). Several surgical procedures can be executed in one surgery. Hence, in such cases there is at least one main surgical procedure that may need other associated procedures. For example, a patient with lung cancer that is going to be subjected to a thoracoscopic lobectomy of lung (i.e., video-assisted removal of one lung's lobe) may also need a mediastinal biopsy (i.e., tissue extraction from the mediastinum) as a complementary intervention, in order to evaluate the spread of the tumor. ICD-9 [91] and *Código de Nomenclatura de Actos Médicos (OM)* [93] were used for the surgical procedure encoding. The first corresponds to the WHO's global classification, while the latter was developed by the *Ordem dos Médicos* and, thus, consisting in a Portuguese medical interventions' nomenclature.

Lastly, medical item records, that comprise the medical items (*Nome do Artigo*) and respective quantities (*Quantidades*) expended between the moment of hospitalization and discharge of a patient, are present in the medical items database. It is noteworthy that one medical item record refers to the registration of an item and the respective expended quantity after a surgery in the inpatient setting. On the other hand, the sum of the expended quantities in one or more surgeries will be referred as medical consumptions. Figure 6 summarizes the hierarchical relationship of the aforementioned data.

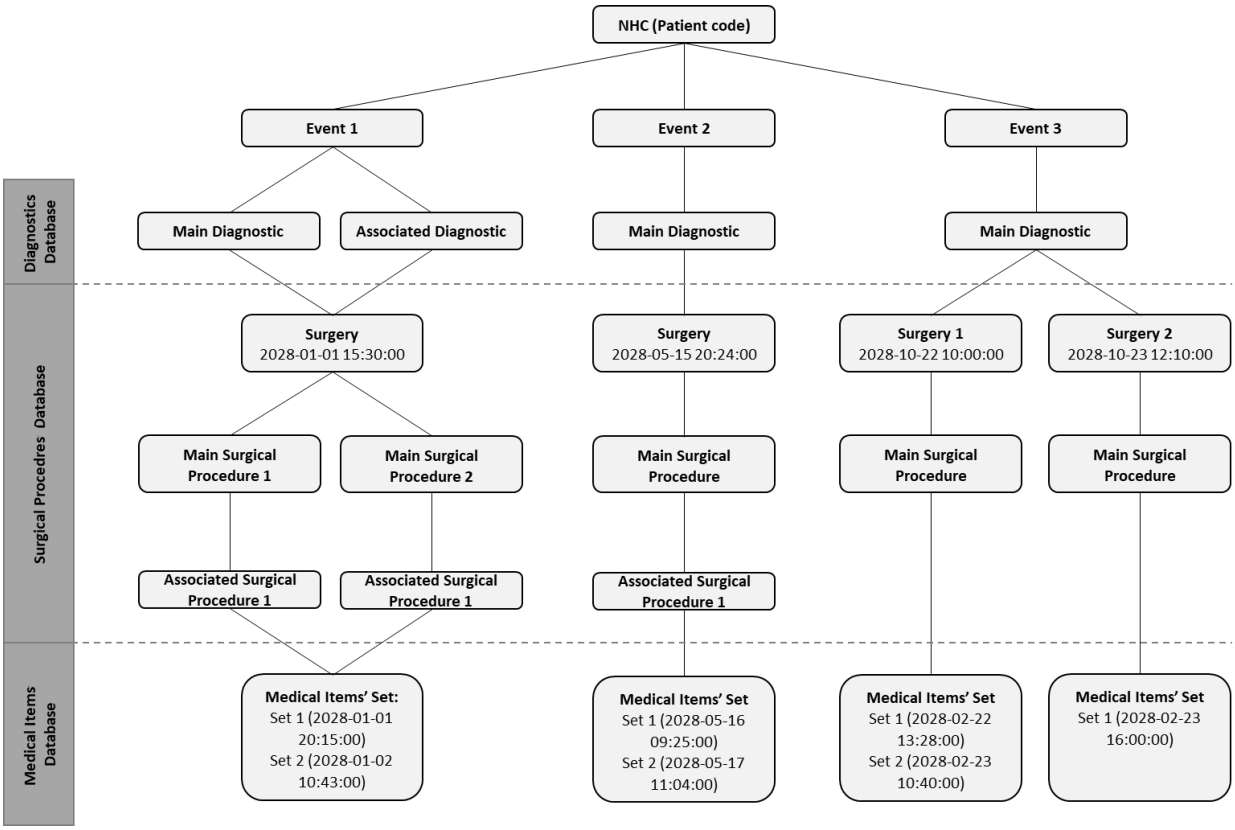


***Figure 6**: Schematic representation of the hierarchical relationship between the variables of the three databases. Each patient presents a unique code (NHC) in the hospital system and may have had more than one event. An event is associated with at least one diagnostic, which, in turn, can lead to one or more surgeries. A surgical procedure or a group of surgical procedures are performed in a surgery. There is at least one main surgical procedure in a surgery, which can be executed alone or together with other associated procedures. The consumption of medical items that were requested after a surgery is registered in specific timestamps, either in ambulatory cases or during post-operative hospitalizations.*

2.2. Data Preparation

The following steps were performed in Python v3.8, using its basic functions and the *Pandas* package.

2.2.1. Data Filtering

The complete dataset includes 12057 patients and, consequently, 13567 events, 13699 surgeries and 2327602 medical item records. Considering the proposed objectives for this project, data was filtered in line with the following inclusion criteria: **a)** patients who

had an event that led to a planned surgery procedure (elective intervention), **b)** surgeries where a post-operative hospitalization was needed (*Duração prevista internamento* ≠ 0), **c)** patients who underwent a surgery in which only one main surgical procedure was performed. Hence, a total of 5414 patients (44.9%), 6666 events (49.2%) and 6788 surgeries (49.6%) were excluded from further analysis, leading to a dataset that covered 6643 patients, 6901 events and 6911 surgeries. Figure 7 shows the detailed exclusion of data when each criterion was applied to the dataset. Additionally, 163 surgeries were not considered since there were no records of expended medical items for those cases. Therefore, the final dataset was composed by 6489 patients, 6738 events, 6748 surgeries and 676153 medical item records.

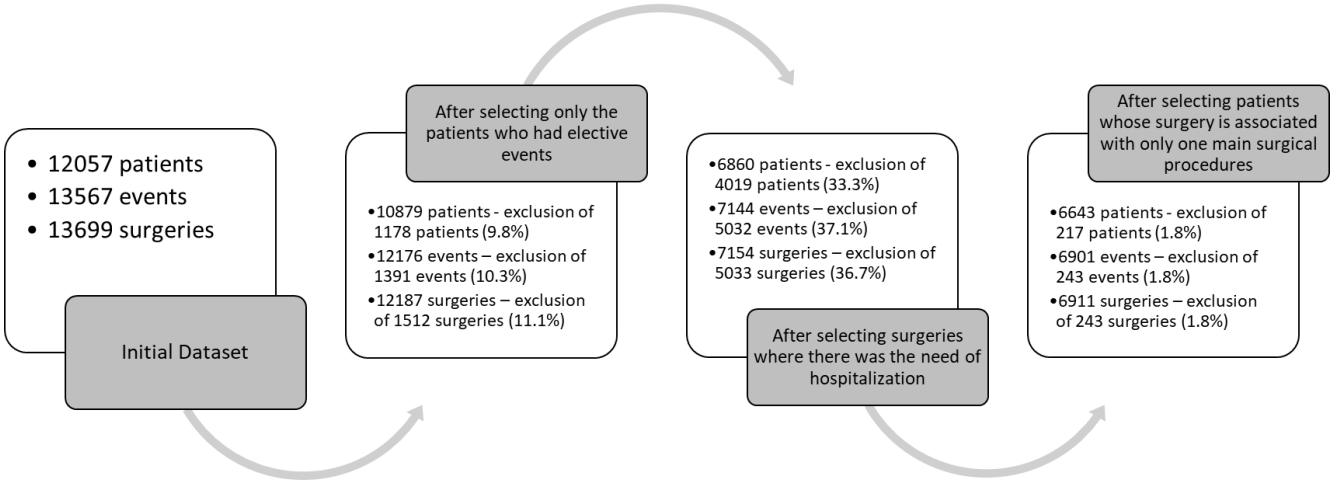


Figure 7*: Data filtering process. Frequency and percentage of patients, events and surgeries excluded from the dataset when an individual inclusion criterion was applied.*

To predict the medical item records in the first day of a post-operative hospitalization, only the first records after a surgery were considered in the development of the recommendation system. For that purpose, all timestamps were truncated to the day (e.g., 2028-01-01 20:30:00 would be abbreviated to 2028-01-01). Afterwards, all medical items registered with the closest timestamp after the respective surgery were selected (Figure 8). From those records, to identify the consumables registered in the same day as the surgery i.e., medical items that were registered until 11:59pm of the day of surgery, the differences between surgeries' and selected medical items records' timestamps were computed.

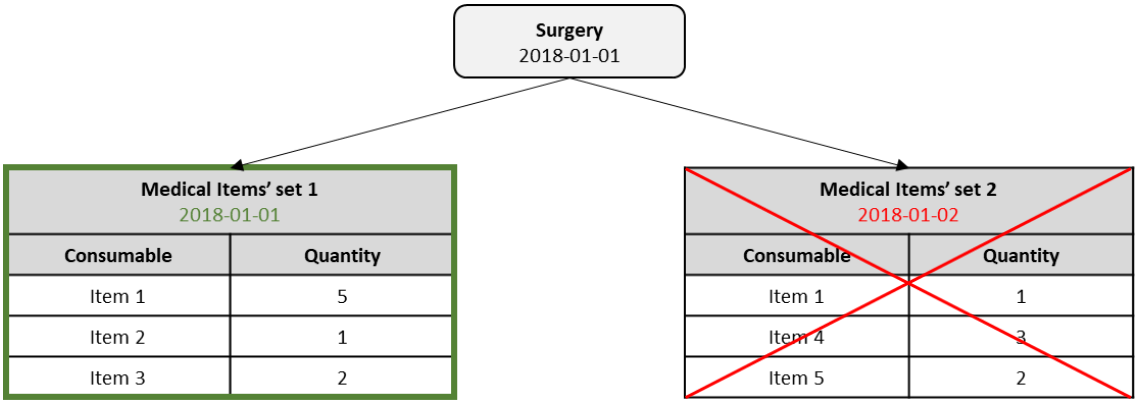


Figure 8: *Example of the process of medical item records selection. Only medical item records registered with the closest timestamp to the surgery were selected.*

2.2.2. Data Structure

To perform a descriptive analysis of the data, the surgical procedures and diagnostics databases were reorganized from the long format to the wide format so that each surgery was represented by one row. Moreover, each column represented a feature. Afterwards, to create a table with all patients' features, both databases were merged. On the other hand, the medical items database remained in the long format. Here we consider a surgery as the study unit instead of a patient, since a set of medical items are consumed during hospitalization in consequence of a surgical procedure (Figure 5).

The proposed recommendation algorithm was based on a neighborhood-based collaborative filtering which implied the computation of similarity measures between patients regarding a set of selected categorical features [19,21,22], i.e., diagnostics (ICD9), surgical procedures (ICD9 and OM) and estimated duration of hospitalization. For that purpose, those features were transformed using the One-hot Encoding. Hence, a new patients features table was created (Figure 9A), where surgeries (S) were represented in the rows, each column (in total 3506) corresponded to one value (f) of the selected features (F), and each cell Q_{sf} indicated the presence or absence of a given $f \in F$ associated to a surgery $s \in S$. Note that each surgery is associated to a patient, therefore, the term "features" refers to the patient's features that are linked to a given surgery ($s \in S$). Furthermore, the medical items database was reshaped into a cross table (Figure 9B) where S and the item set (I) were

represented in the rows and columns, respectively. Thus, a cell Q_{si} consisted in the quantity of an item i that was consumed in the first day of hospitalization after a surgery s.

A

Surgeries (S)	Features' values ($f \in F$)			
	0	0	1	1
	0	Q_{sf}	1	0
	1	1	0	0
	1	0	1	0

B

Surgeries (S)	Items (I)			
	1	0	0	0
	0	Q_{si}	3	0
	4	2	0	0
	0	0	1	0

Figure 9**: Adequate structure of the two tables included in the algorithm for the estimation of the medical item records needed for the first day of hospitalization of a patient that was subjected to an elective intervention.* ***A) *Patients' feature table where each row represents a surgery $s \in S$ and each column corresponds to one value f of the set of features F. A cell Q_{sf} indicates the absence ($Q_{sf} = 0$) or presence ($Q_{sf} = 1$) of a given f in surgery s;* ***B)*** *Medical items table where each row represents a surgery $s \in S$ and the columns correspond to the medical item $i \in I$. A cell Q_{si} indicates the quantity of a given i that was consumed in the first day of hospitalization after a given surgery s.*

2.3. Descriptive Analysis

A descriptive statistical analysis was performed to characterize the filtered dataset. Measures of location and dispersion (i.e., means and standard deviations, respectively) were computed to describe numerical variables, while frequencies and proportions were determined for categorical variables.

An analysis of the patients' data, including the input features (i.e., diagnostics, surgical procedures, and estimated duration of hospitalization), was executed in order to understand its complexity. Regarding the whole medical item records/consumptions data, the number of distinct expended medical items, the average number of records/consumptions for each patient, and the items' distribution considering their frequency of records/consumptions were described.

Posteriorly to the selection of the first medical item records registered after a given surgery, the absence of records in the first day of hospitalization was determined. For that purpose, the differences between surgeries and respective records' timestamps were calculated. Moreover, the proportion of surgeries without records in the first day of

hospitalization was computed for each medical specialty. Those surgeries were excluded from further analyses, being the remaining (N = 5088) employed to characterize the medical item records/consumptions in the first day of hospitalization, using the previously explained approach. The same analysis was carried out for each medical specialty.

All statistical analyses and graphical representations were performed in Python v3.8 and RStudio v4.0.2.

2.4. Recommendation Algorithm

To build the CF-based recommendation algorithm, the *Scikit-Learn*, *Scipy*, *Math*, *Numpy* and *Pandas* packages from Python v3.8 were used.

The aim of the proposed algorithm is to predict the medical item records needed in the first day of hospitalization of a new patient u, i.e., a patient that will be subjected to an elective intervention. These predictions are based on the data from the surgeries (S) of known patients, i.e., patients that were already subjected to an elective intervention and, consequently, expended medical items during their period of hospitalization. Data from the surgeries of known patients must be included in the neighborhood (N) of the new patient.

Figure 10 represents the workflow of the recommendation algorithm, which is explained in the following steps:

1) Two types of data from known patients (i.e., patients' features and medical item records) were incorporated in the algorithm before the data from a new patient u was added as an input. The patient's features allowed the creation of a dummy variable table, while the medical item records data led to a crosstable, as shown in Figure 9.

2) The patient u features are added to the patients' features table to compute his/her similarity scores (please see below) with the features data associated to $s \in S$ (which will be referred as W_{us}), and consequently determine N.

3) Predictions of the medical consumption of an item i needed for a patient u in his/her first day of hospitalization after an elective intervention, Q_{ui}, are obtained for all $i \in I$.

$$Q_{ui} = \frac{\sum_{s \in N} W_{us} \times Q_{si}}{\sum_{s \in N} |W_{us}|}$$

4) The medical item records that will be recommended to patient u are selected based on the Q_{ui}.

Besides the new patient's features, there are three hyperparameters (i.e., parameters which are part of the algorithm's structure that can be tuned [94]) that also need to be given as inputs to the algorithm:

- **Similarity measure (SM)** – Expressions that allow the estimation of the similarity scores between patients regarding their features. Since all selected features are categorical, similarities scores can be calculated using two distinct measures:
    - **Jaccard Index** [95–97] - Ratio of features shared between patient u and the known patient that was subjected to surgery s, considering the total number of features of both patients. This measure varies between 0 and 1, being that the closer to 1, the more similar the two patients.

$$J(u, s) = \frac{|F_u \cap F_s|}{|F_u \cup F_s|}$$

where F_u represents the features of patient u; F_s corresponds to the features of the known patient who was subjected to surgery s; $|F_u \cap F_s|$ is the cardinality of features that are coincident in both patients; and $|F_u \cup F_s|$ denotes the total number of features of both patients.

    - **Cosine Similarity** [96–98] - This measure represents the features of each patient as a vector in an inner product space, being the similarity between two vectors computed as the cosine of the angle between them:

$$Cosine(u,s) = \frac{\sum_{f \in F_{us}} r_{uf} . r_{sf}}{\sqrt{\sum_{f \in F_u} r_{uf}^2} \sqrt{\sum_{f \in F_s} r_{sf}^2}}$$

where F_u and F_s consist in the features of patient u and the known patient that was subjected to a surgery s, respectively; F_{us} corresponds to the set of features shared by both patients; r_{uf} and r_{sf} are the values that represent the presence of a class of a feature in each patient ($r_{uf} = 1$ and $r_{sf} = 1$).The closer to 1, the more similar the two patients.

- **Similarity Threshold (ST) –** Value that determines which known patients will contribute to the predictions of a new patient u (neighborhood). Known patients with similarity scores equal or higher than the ST are selected. Therefore, the ST can take on values between 0 and 1.

- **Recommendation Threshold (RT) –** Value that determines which medical items records will be suggested for the first day of hospitalization of a new patient u. Medical items whose Q_{ui} are equal or higher than the RT are recommended.

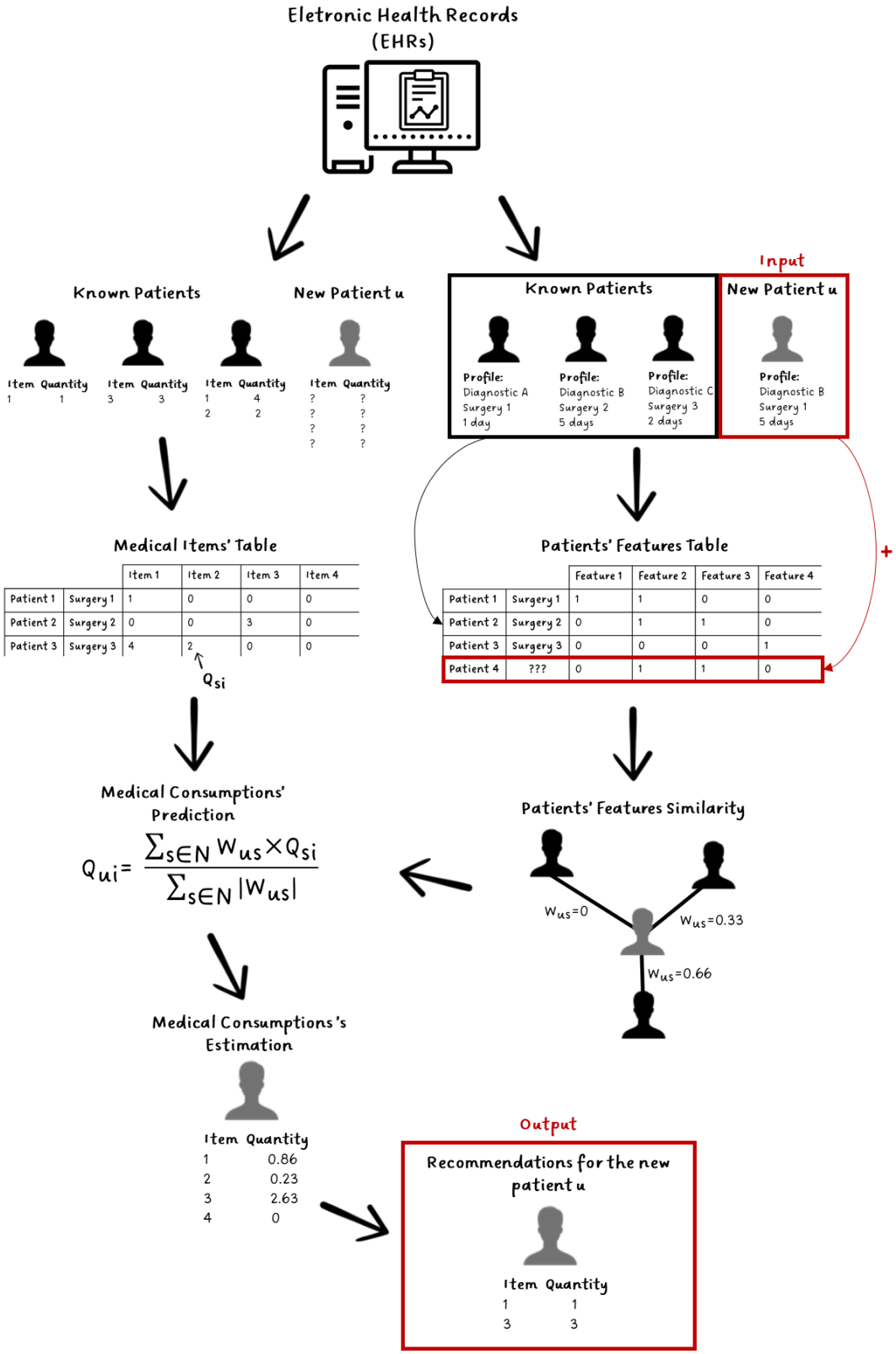


Figure 10: *Workflow of the recommendation algorithm. Two types of data are extracted from the EHRs (i.e., patients' features and medical item records) of patients who already had records of their first day of hospitalization after a given surgery s (known patients). This information is used to create two tables: one with the data from the medical item records of known patients, where each cell (Q_{si}) represents the medical consumption of an item i used during the first day of hospitalization of a known patient; and other with their features after applying One-hot Encoding. The input of the algorithm is the set of features of a new patient u to whom the recommendations will be performed, which will be added to the patients' features table. Afterwards, similarity scores W_{us} are computed between u and the known patients, in order to select the u's neighbors and predict the medical consumptions of each recommendable item $i \in I$ needed during his/her first day of hospitalization (Q_{ui}). Lastly, the medical item records are recommended depending on the Q_{ui}.*

2.5. Algorithm's Evaluation

The steps of the algorithm's evaluation were executed resorting to the *Scikit-Learn*, *Scipy*, *Math, Numpy* and *Pandas* packages from Python v3.8. Additionally, the package *ggplot2* from RStudio v4.0.2 was used for graphical representations.

To assess the performance of the recommendation algorithm, the total dataset was split into a training and test set. The training set consisted in 75% of the surgeries (N = 3814), which were used for the tuning of the algorithm's hyperparameters and to represent the group of known patients. On the other hand, the test set comprised 25% of the surgeries (N = 1272), being solely used for the algorithm testing as the new patients for whom medical item records were recommended. The split was performed in a stratified way to maintain the proportions of surgeries of each medical specialty in both groups. Thus, 'Gastrenterologia' was not included in the dataset for the algorithm's evaluation (N = 5086), due to the low number of surgeries (N = 2) which would impair the stratified split. Figure 11 outlines the process of evaluation of the algorithm.

2.5.1. Hyperparameters Tuning

To select the best settings of the hyperparameters to be applied in the testing step, a *k-fold cross-validation* method [99,100], namely, a *4-fold cross validation* (k = 4) was executed using the training set. This strategy consisted in splitting the training set in *4* equally sized and stratified folds so that each fold was considered in turn as the validation group (new patients), while the remaining 3 folds were used to formulate the predictions (known patients). The average and standard deviation of the performance measures obtained from each validation group were computed. This process is repeated for each setting of hyperparameters. Table 3 displays the settings selected for the hyperparameters tuning.

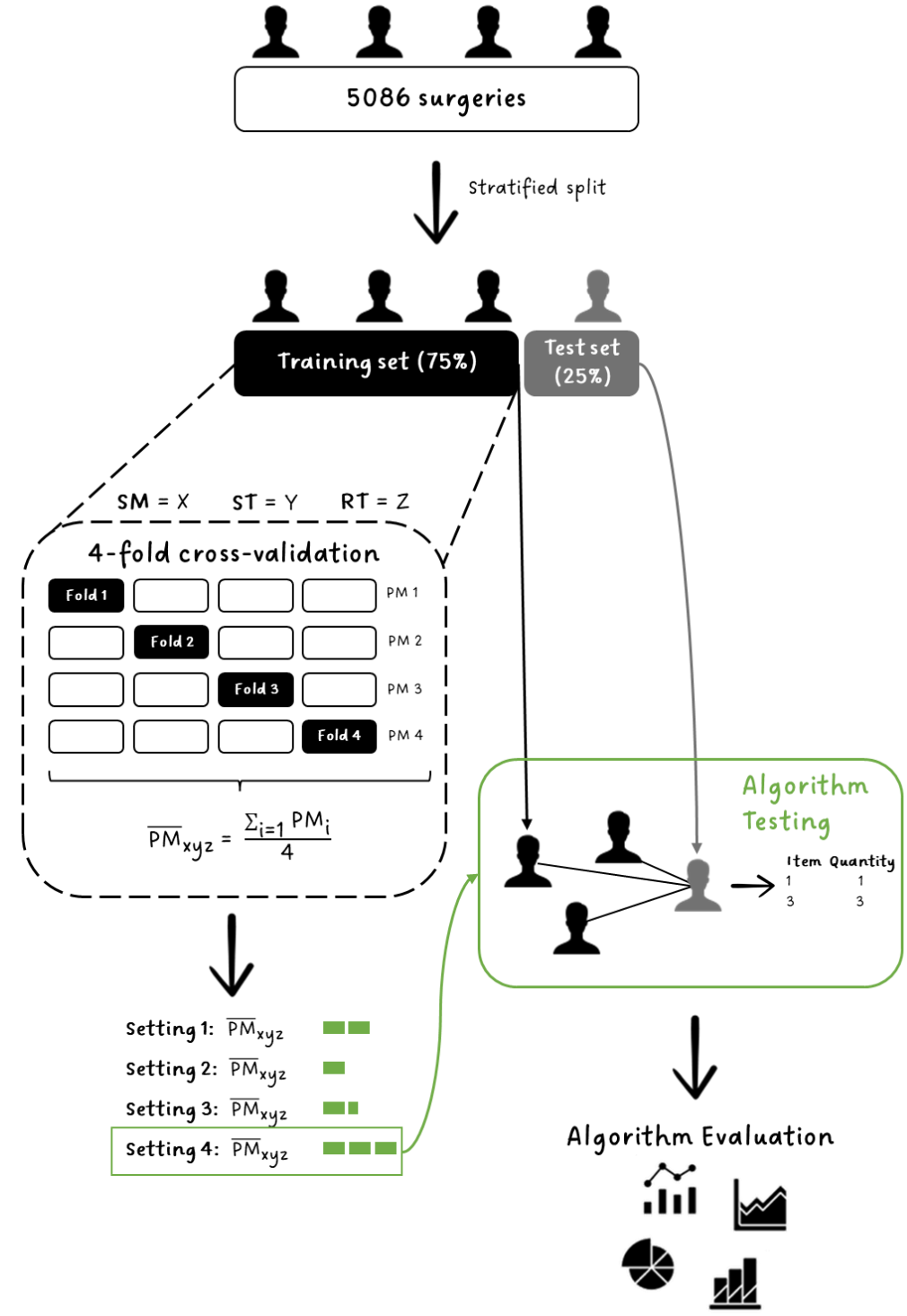


Figure 11: *Workflow of the algorithm testing. The total dataset was divided into a training set (75%) and test set (25%). A 4-fold cross-validation was applied on the training set to tune the hyperparameters (SM, ST and RT) settings. For each combination of hyperparameter settings (x, y, z), an average of the performance measures (PM) obtained in each fold of the cross-validation was computed, being the combination with the best performance selected. Posteriorly, a testing process was executed by predicting the medical item records needed for the test set and the results were evaluated.*

Table 3: *Hyperparameters settings used in the 4-fold cross-validation.*

Hyperparameter	Values
Similarity Measure	[Cosine Similarity, Jaccard Index]
Similarity Threshold	[0, 0.1, 0.2]
Recommendation Threshold	[0, 0.5, 1]

2.5.2. Performance Measures

2.5.2.1. Accuracy

Accuracy is an important measure to quantify the proximity of a recommendation system's estimates of the medical consumptions to the real data [101,102]. Thus, the Root Mean Square Error (RMSE) was computed to measure the prediction accuracy of the medical consumptions proposed by the algorithm.

$$RMSE = \sqrt{\frac{\sum_{(u,i)\in T} (\hat{y}_{ui} - y_{ui})^2}{T}}$$

where $\hat{y}_{ui}$ is the medical consumption predicted of an item i for a patient u; y_{ui} corresponds to the true medical consumption of an item i that was needed for a patient u; and T is the number of patient-item pairs (u, i) in which the medical items were either recommended by the algorithm or expended during the first day of hospitalization of the test set. The RMSE was calculated for all items, for each medical specialty and for each item individually to understand the global, specialty and individual performance.

2.5.2.2. Classification performance ("Usage Prediction")

To evaluate the *usage prediction* (i.e., whether the algorithm predicts properly the items that a patient would need during his/her first day of hospitalization), a comparison between the recommendations and the real data is needed [6,101,102]. For that purpose, confusion

matrices (Figure 12) were built to assess the algorithm's precision, recall, F1 measure, false positive rate (FPR) and area under the receiving operating characteristic curve (AUC) regarding its recommendations.

	Recommended	Not Recommended
Used	TP	FN
Not Used	FP	TN

Figure 12: *Confusion matrix. TP – True positives; FN – False negatives; FP –False positives; TN –True negatives.*

In the field of recommendation systems, true positives (TP) correspond to the number of correct recommendations; false positives (FP) correspond to the number recommendations of items that were not used; false negatives (FN) correspond to the number of cases where a given item was used but was not recommended by the system; and true negatives (TN) correspond to the number of cases where a given item was neither recommended nor used..

- **Precision** – Indicates the ratio of TP considering all the algorithm's suggestions [6,101,102].

$$Precision = \frac{TP}{TP + FP}$$

- **Recall –** Indicates the ratio of TP considering all used medical items [6,101,102].

$$Recall = \frac{TP}{TP + FN}$$

- **F1 Measure –** Represents the harmonic mean of the precision and recall of the algorithm [6,101,103]. This measure can take values between 0 and 1, being values closer to 1 indicators of a good performance [6,103].

$$F1\ Measure = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

- **FPR** – Indicates the ratio of incorrect recommendations (FP) considering all medical items that were not used [101,102].

$$FPR = \frac{FP}{FP + TN}$$

- **AUC –** The receiving operating characteristic (ROC) curve allows the evaluation of the capacity of classifiers or other statistical models in discriminating known classes (e.g., recommended item and not recommended item) [6,101,102,104–107]. In particular, it compares the recall (y-axis) and FPR (x-axis) of the model's predictions when different cutoffs, that define which class will be given to an observation, are taken into consideration [6,101,102,104,105]. The calculation of the AUC is a method to quantify the model's classification ability. This measure varies between 0 and 1, being values closer to 1 indicators of a good performance, while values near 0.5 mean that the model cannot separate the known classes [104–107].

Moreover, the algorithm was evaluated regarding its *usage prediction* of medical items individually, medical items expended by the whole test set, and medical items that were used in each medical specialty. The variation of the algorithm's performance by gradually including more medical items (ordered by their frequency of records) in the medical item's table was also assessed.

2.5.2.3. Coverage

The *coverage* of a recommendation system is defined as "the proportion of items that the recommendation system can recommend" [101,102], which is strongly affected by the long

tail problem [37,38,102]. To measure the *coverage* of the proposed recommendation algorithm, a comparison between the set of recommended medical items and the set of registered medical items for the test set was performed. The same approach was executed for each medical specialty. Additionally, the *coverage* of the algorithm when different percentages of items were gradually added (ordered by their frequency of records) to the medical item's table was explored.

To assess the overlap between recommendations, the sets of medical items recommended for surgeries that performed one of seven distinct surgical procedures, which were suggested by nurses of the hospital, (i.e., 'Tratamento de cataratas', 'Artroscopia do ombro', 'Tratamento de varizes', 'Tratamento de hernia inguinal', 'Queratoplastia', 'Colecistectomia', and 'Rinoplastia') were compared. For that purpose, the Jaccard index was computed between pairs of item sets.

$$J(I_1, I_2) = \frac{|I_1 \cap I_2|}{|I_1 \cup I_2|}$$

Where $|I_1 \cap I_2|$ is the cardinality of items that are coincident in both recommended item sets (I_1 and I_2); and $|I_1 \cup I_2|$ denotes the total number of distinct items present in both item sets.

2.5.2.4. Trust

Trust consists in the level of confidence that the users (i.e., healthcare professionals) have in the algorithm's recommendations [101,102,108]. In this work, a questionnaire (see Appendix A) was given to 2 nurses from Luz Lisboa Hospital, in order to measure their trust in the algorithm. For that purpose, the nurses were provided with 15 recommendations which were randomly selected from a set of patients who experienced one of the seven surgical procedures whose predictions were considered relevant for the hospital. Additionally, they were asked to formulate a list of medical item records based on their experience for each case to compare with algorithm's results and the real medical item records. A descriptive analysis of the results was performed.

2.6. Statistical Analysis

2.6.1. Correlation tests

To assess the relationship between the RMSE of the predicted medical consumptions of each item and the respective standard deviation of their registered medical consumptions, a correlation test was computed. Since the assumption of a normal distribution of the values of both variables was not validated, a non-parametric approach, i.e., Spearman's rank correlation test, was performed ($\alpha = 0.05$).

The Spearman's correlation coefficient is a non-parametric alternative to the Pearson's correlation coefficient, where the corresponding ranks of the values of both variables (x_i, y_i) are used instead of the real values [109–111]. Thus, it measures the monotonic relationship between two numerical variables, being $\hat{\rho} = 0$ the absence of an association, $\hat{\rho} = 1$ a perfect positive correlation, and $\hat{\rho} = -1$ a perfect negative correlation [109–111].

$$\hat{\rho} = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n^3 - n}$$

where d_i^2 is the difference between the corresponding ranks of each pair of observations (x_i, y_i) and n corresponds to sample size.

The statistical hypotheses and the test statistic (T) are as follows:

$$H_0: \rho = 0 \text{ (null hypothesis)}$$

$$H_1: \rho \neq 0 \text{ (alternative hypothesis)}$$

$$T = \hat{\rho}\sqrt{n-1}$$

where $\hat{\rho}$ is the sample Spearman's correlation coefficient and n is the sample size. When $n > 10$, T follows an asymptotic standard normal distribution.

The Pearson rank correlation test was also used to test the relationship between four performance measures (RMSE, precision, recall and F1 measure) regarding the predictions for each medical specialty and the respective number of surgeries.

2.6.2. Comparison of groups

Kruskal-Wallis tests [112] were executed to assess if there were significant differences between medical specialties regarding the algorithm's performance (precision, recall and F1 measure) in predicting the respective medical item records. The Kruskall-Wallis test is a non-parametric alternative to ANOVA and was computed since the ANOVA's residuals did not follow a normal distribution. Here, data from all groups is ranked together and the median ranks of the groups are the compared [112]. The statistical hypothesis are

$$H_0: med_1 = med_2 = \cdots = med_C = med$$

$$H_1: \exists_i \, med_i \neq med, \qquad i = 1,2, \dots, C$$

where med_i is the median of the ranks of group i and C is the number of groups being tested and the test statistic is given by [112]

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{C} \frac{R_i^2}{n_i} - 3(N+1)$$

where N is the total number of observations, n_i is the number of observations in the ith group, and R_i^2 is the sum of the ranks in the ith group.

2.6.2.1. Post-hocs

The Kruskal-Wallis tests that rejected the null hypothesis ($\alpha = 0.05$) were followed by non-parametric multiple comparisons (i.e., Mann-Whitney U tests [113]) to identify the pairs of medical specialties that were significantly different. The statistical hypotheses and the test statistic U correspond to [113,114]

$$H_0: med_1 = med_2$$

$$H_1: med_1 \neq med_2$$

where med_1 and med_2 are the median of the ranks in the group 1 and 2, respectively;

$$U = \min\{mn + \frac{\big(n(m+1)\big)}{2} - R_1; mn + \frac{\big(n(m+1)\big)}{2} - R_2\}$$

where m is the number of observations in group 1, n is the number of observations in group 2, R_1 is the sum of ranks in group 1, and R_2 corresponds to the sum of ranks in group 2.

To avoid the type I error, i.e., rejection of the null hypothesis when it is true, which results from the high number of statistical tests performed in one dataset [115], the obtained p-values were adjusted using the Bonferroni's correction. This method considers the α_m for each one of the m tests as [115]

$$\alpha_m = \frac{\alpha}{m}.$$

Therefore, the p-value must be lower than α_m to reject the null hypothesis [115]. To compare the p-value with the usual significance levels

$$p_{adj}(i) = \min\{m \times p_i, 1\}$$

where m corresponds to the number of statistical tests and p_i is the p-value obtained by the statistical test i.

3. Results

3.1. Descriptive Analysis - Complete Dataset

3.1.1. Patients' Features Analysis

Considering all patients who met the inclusion criteria (N = 6643), the majority had only one event in the hospital (N = 6402, 96.37%), while the remaining patients (N = 241, 3.63%) were admitted in the hospital between 2-4 times during the stipulated period. Moreover, only in 10 events (0.14%) two distinct surgeries were performed.

Regarding the selected features to compute similarities between patient features for the recommendation algorithm, 636 and 806 distinct main surgical procedures, labeled according to ICD-9 and OM codes respectively, were identified. Each main surgical procedure was performed together with an average of 1 (SD = 1.24) associated surgical procedure, although 3279 (~47.45%) main surgical procedures were executed alone. The number of associated surgical procedures ranged between 0-11, being the cases with a lower number of procedures more frequent. Furthermore, there were 882 different main diagnostics, of which 'Varizes das extremidades inferiores, assintomático' was the most reported (N = 287 patients and events, 4.32% and 4.16% respectively). The range of associated diagnostics was 0-4, which occurred in 6186 (89.65%) events, 626 (9.07%) events, 76 (1.10%) events, 10 (0.14%) events and 3 (0.04%) events, respectively. Additionally, patients had to stay hospitalized between 1 and 53 days after an elective intervention. Approximately 70.38% of the surgeries (N = 4749) led to a maximum of 5 days of hospitalization.

3.1.2. Medical Item Records Analysis

During the studied period, 676153 records of 1310 different medical items were registered in post-surgical hospitalizations. An average of ~104 (SD = 244.86) medical item

records were registered for each patient, ranging from 1 to 7636 records. Correspondingly, a total quantity of 2069141 consumables were expended and, consequently, patients used an average of ~319 (SD = 839.11, range = 1 - 23135) medical items during hospitalization. Table 4 shows the five most registered medical items, as well as the five predominantly used consumables (i.e., used in higher quantities).

Table 4: *Frequency, quantity, and percentage of the five most registered and consumed medical items.*

Medical Item Records (N = 676153)			**Medical Consumptions (N = 2069141)**		
Item	**Frequency**	**Percentage**	**Item**	**Quantity**	**Percentage**
Luva Nitrilo M	27554	4.08%	Luva Nitrilo M	382022	18.46%
Cloreto de Sodio 0,9% IV Amp 10 ml	23213	3.49%	Compressa N/Esteril Tnt 7,5x7,5cm 30gr	260030	12.57%
Sistema Administração Soro	22742	3.46%	Agulha Diluição 19gx1 ½ 1,10x40	141674	6.85%
Agulha Diluição 19gx1 ½ 1,10x40	22509	3.32%	Seringa 5ml (2 peças)	78816	3.81%
Seringa 5ml (2 peças)	20218	3.02%	Seringa 10ml (2 peças)	75731	3.66%

Regarding the records for each medical item, the distribution of their frequencies was assessed (Figure 13). Notoriously, most of the records were associated to a small subset of items, while the remaining records are sparsely distributed among the greater portion of the expended medical items. A handcrafted threshold of 0.1% of the total medical item records was defined to select the items that were registered equally or more than the obtained value ($N \approx 676$). Only 95 (~7.25%) items met the aforementioned condition, which covered ~91.30% of the records.

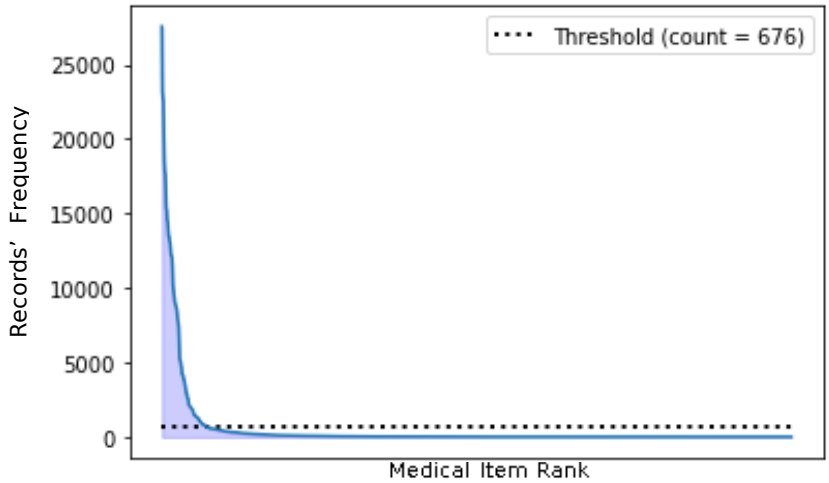


***Figure 13**: Medical item distribution regarding their frequency of records during the whole period of hospitalization. Dashed line – threshold of 676 medical item records.*

3.2. Descriptive Analysis – Dataset of the first day of hospitalization

3.2.1. Data from the first day of hospitalization after an elective surgery

Figure 14 presents the frequency of surgeries whose first medical item records were registered in the first day of hospitalization or after. In 24.60% of the cases (N = 1660), records were registered after the first day of hospitalization in the inpatient unit.

The remaining 75.40% of the surgeries (N = 5088) were considered as the ground-truth for the following analysis and development of the recommendation algorithm. There was a loss of information regarding 1654 (24.55%) events and 1697 (26.15%) patients. Nevertheless, the resulting dataset included 560 (88.05%) and 697 (86.48%) of the main surgical procedures categorized by ICD-9 and OM observed in the whole dataset, respectively, as well as 760 (86.17%) of the main diagnostics.

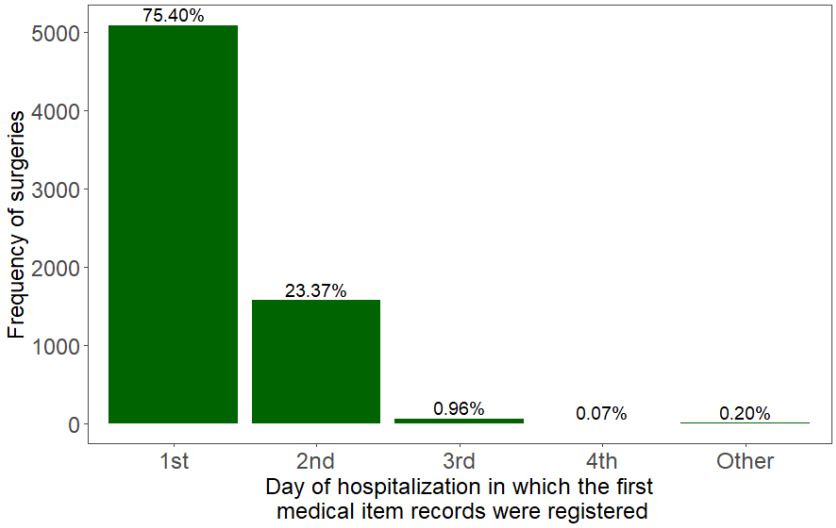


***Figure 14**: Frequency/percentage of surgeries in which the medical items expended in the first day of hospitalization were registered in the respective day or after.*

Thirteen medical specialties performed elective surgeries Hospital da Luz Lisboa, being the number of executed interventions different between them (range = 4 - 1742). A percentage of surgeries that did not present medical item records in the first day of hospitalization was observed in all medical specialties (Table 5). Concerning the 1660 surgeries without medical item records in the first day of hospitalization, 'Ortopedia' presented the highest portion of surgeries with missing records (N = 436, 26.27%), while in 'Gastrenterologia' only 2 (0.12%) surgeries evidenced that condition. On the other hand, the lack of medical item records in the first day of hospitalization were observed in 50% of the surgeries performed in 'Gastrenterologia'. 'Cirurgia Cardio-Torácica' had the lowest percentage of surgeries with missing records (N = 16, 7.48%).

***Table 5**: Frequency/percentage of surgeries that did not present medical item records in the first day of hospitalization in each medical specialty. N Surgeries – number of surgeries without medical item records; N Surgeries (Total) – Total number of surgeries performed in each medical specialty; % Total – percentage of the total number of surgeries that did not present medical item records; % Medical Specialty - percentage of surgeries in each medical specialty that did not present medical item records.*

Medical Specialty	N Surgeries	N Surgeries (Total)	% Total	% Medical Specialty
Cirurgia Cardio-Torácica	16	214	0.96%	7.48%
Cirurgia Vascular	82	484	4.94%	16.94%
Cirurgia Pediátrica	11	56	0.66%	19.64%
Cirurgia Geral	273	1291	16.44%	21.15%
Ginecologia-Obstetrícia	190	780	11.45%	24.36%
Neuro-Cirurgia	153	619	9.22%	24.72%
Ortopedia	436	1742	26.27%	25.03%
Cirurgia Plástica Reconstrutiva e Estética	56	190	3.37%	29.47%
Urologia	287	915	17.29%	31.37%
Otorrinolaringologia	111	339	6.69%	32.74%
Oftalmologia	29	79	1.75%	36.71%
Cirurgia Maxilo-Facial	14	35	0.84%	40.00%
Gastrenterologia	2	4	0.12%	50.00%

3.2.1.1. Medical Item Records Analysis

Overall Results

From the 1310 different medical items, 602 (45.95%) were expended during the first day of hospitalization. Furthermore, 17.38% (N = 117533) of the medical item records and 16.58% (N = 343118) of the medical consumptions were registered. Each surgery presented an average of 23.10 (SD = 15.90, range = 1 - 140) medical item records and 67.44 (SD = 80.13, range = 1 - 953) medical consumptions. Table 6 shows the five most registered and consumed medical items in the first day of hospitalization.

Table 6: *Frequency, quantity, and percentage of the five most registered and consumed medical items in the first day of hospitalization.*

Medical Item Records (N = 117533)			**Medical Consumption (N = 343118)**		
Item	**Frequency**	**Percentage**	**Item**	**Quantity**	**Percentage**
Luva Nitrilo M	5439	4.63%	Compressa N/Esteril Tnt 7,5x7,5cm 30gr	111882	32.61%
Cloreto de Sodio 0,9% IV Amp 10 ml	5311	4.52%	Luva Nitrilo M	53445	15.58%
Compressa N/Esteril Tnt 7,5x7,5cm 30gr	5263	4.48%	Agulha Diluição 19gx1 ½ 1,10x40	10819	4.15%
Seringa 5ml (2 peças)	5064	4.31%	Luva Vinyl M	10427	3.04%
Cobertura Descart Termómetro	5044	4.29%	Seringa 5ml (2 peças)	10075	2.94%

The item distribution regarding the frequency in which they were registered demonstrated an elevated density of records in a small portion of medical items (Figure 15). Considering a threshold of 0.1% of the total medical item records (~118), 91 (~15.10%) medical items were registered equally or above that value, covering approximately 95.40% of the records.

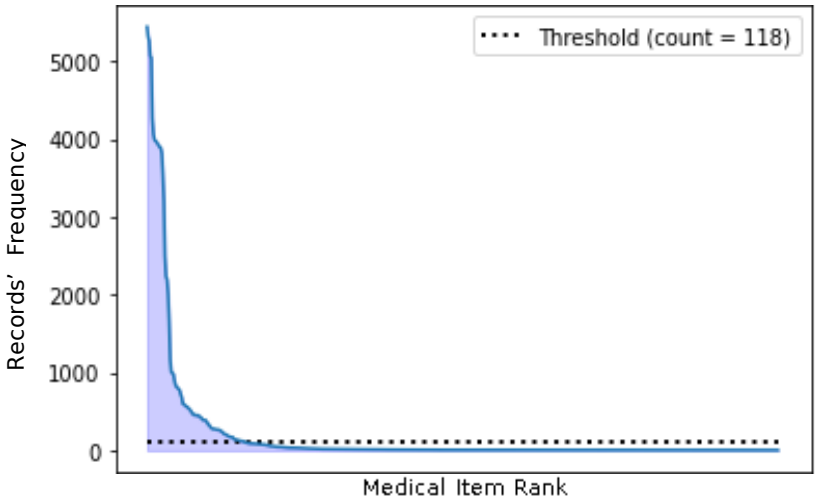


Figure 15: *Medical item distribution regarding their frequency of records during the first day of hospitalization. Dashed line – threshold of 118 medical item records.*

Results per Medical Specialty

Figure 16 depicts the total number of distinct medical items expended in the first day of hospitalization after the surgeries from each medical specialty. ‘Cirurgia-Geral’ used the largest variety of items (N = 283, 47.01%) closely followed by ‘Ortopedia’ (N = 279, 46.35%). In contrast, ‘Gastrenterologia’ only consumed 17 (2.8%) different medical items. An average of ~21 (SD = 12.28) distinct items were used after a surgery per medical specialty.

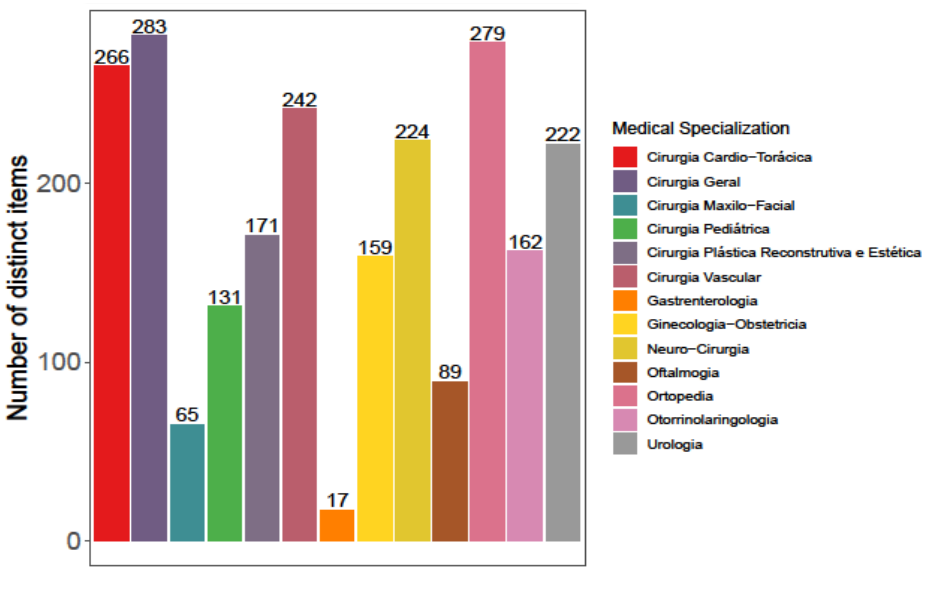


Figure 16: *Number of distinct medical items used in each medical specialty.*

Table 7 summarizes the medical item records and consumptions data of each medical specialty. The highest frequency of medical records and consumptions registered per surgery occurred in ‘Cirurgia Cardio-Torácica’, where there was an average of ~157 (SD = 179.38) records and ~40 (SD = 26.94) consumptions in the first day of hospitalization. Moreover, it also exhibited the highest variability regarding those two variables. ‘Ortopedia’ registered the most medical item records (N = 28928, 24.61%), and expended the highest amount of medical items (N = 75988, 22.15%).

The items registered in ≥60% of the surgeries in at least one medical specialty were considered as the most frequently used medical items (popular items). Twenty-one items

were identified, of which 18 were consumed in common in two or more medical specialties (Figure 17).

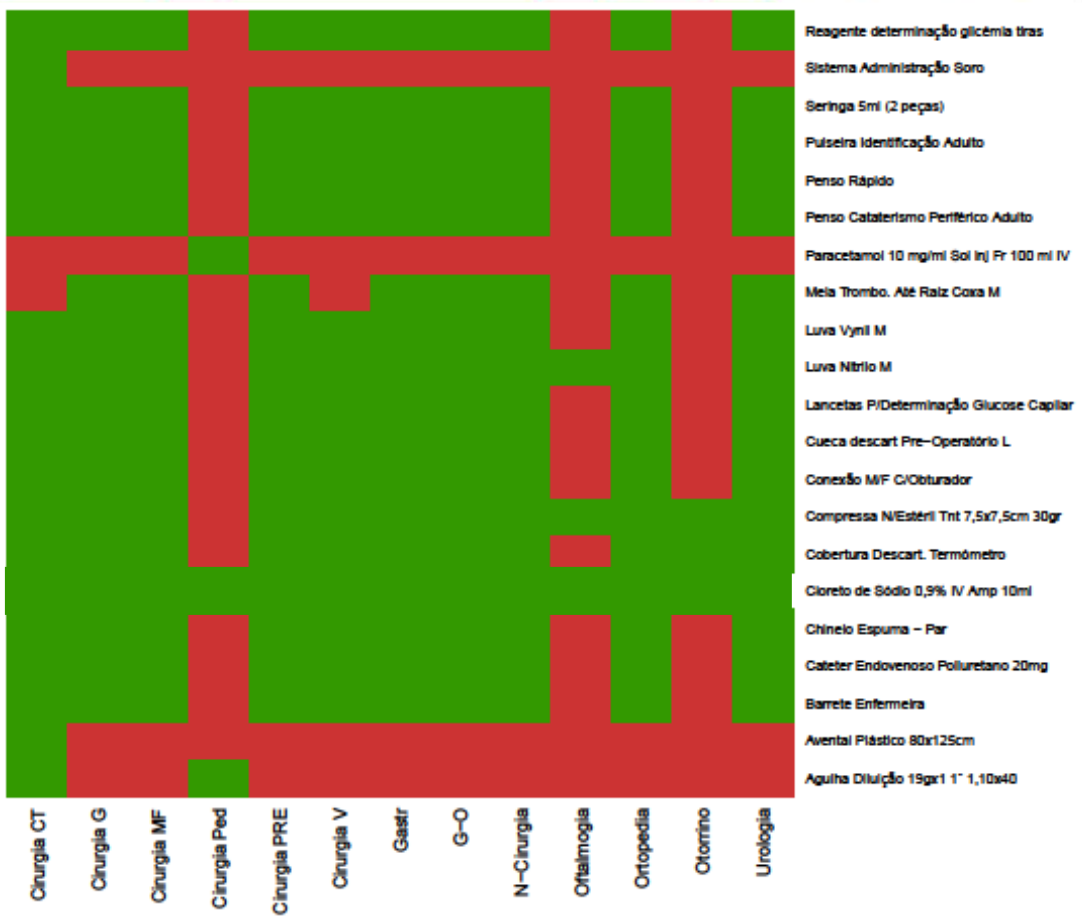


Figure 17*: Presence/absence of the twenty-one most frequently used medical items (i.e., registered in 60% or more surgeries in at least one medical specialty) in the set of most frequently registered items of each medical specialty. Green cell – presence of the medical item; red cell – absence of the medical item. Cirurgia CT – Cirurgia Cardio-Tóracica; Cirurgia G – Cirurgia Geral; Cirurgia MF – Cirurgia Maxilo-Facial; Cirurgia Ped – Cirurgia Pediátrica; Cirurgia PRE – Cirurgia Plástica Reconstrutiva e Estática; Cirurgia V – Cirurgia Vascular; Gastr – Gastrenterologia; G-O – Ginecologia-Obstetrícia; N-Cirurgia – Neuro-Cirurgia; Otorrino – Otorrinolaringologia.*

Table 7: *Summary of the frequency/quantity of medical item records and consumptions per surgery in each medical specialty. N Surgeries – Number of surgeries; SD – Standard deviation; Min – Minimum; 25% - First quartile; 50% - Median; 75% - Third quartile; Max – Maximum; Total – Total of medical item records/medical consumptions.*

		Medical item records per surgery in each medical specialty							**Medical consumptions per surgery in each medical specialty**						
Medical Specialty	**N Surgeries**	**Mean ± SD**	**Min**	**25%**	**50%**	**75%**	**Max**	**Total**	**Mean ± SD**	**Min**	**25%**	**50%**	**75%**	**Max**	**Total**
Cirurgia Cardio-Torácica	198	40.26 ± 26.94	2	18	32.5	56.5	140	7972	156.98 ± 179.38	2	33	62	233.5	953	31083
Cirurgia Geral	1018	24.17 ± 15.89	1	17	18	29	106	24608	71.42 ± 82.49	1	32	35	93	653	72710
Cirurgia Maxilo-Facial	21	19.57 ± 11.31	3	16	17	21	52	411	45.10 ± 33.39	3	32	33	55	133	947
Cirurgia Pediátrica	45	16.33 ± 10.62	1	8	13	25	43	735	33.87 ± 32.71	1	11	22	50	163	1524
Cirurgia Plástica Reconstrutiva e Estética	134	18.78 ± 13.99	1	17	17	18	118	2517	45.58 ± 58.69	1	32	32	33	532	6108
Cirurgia Vascular	402	26.55 ± 18.50	2	17	19.5	35	93	10672	91.00 ± 109.98	2	32	39	109.5	536	36584
Gastrenterologia	2	17.00 ± 0.00	17	17	17	17	17	34	32.00 ± 0.00	32	32	32	32	32	64
Ginecologia-Obstetrícia	590	21.38 ± 12.18	1	17	19	24	98	12612	51.99 ± 41.50	1	33	35	55.5	280	30674
Neuro-Cirurgia	466	23.95 ± 16.68	2	17	18	31	97	11159	75.09 ± 86.33	2	32	35	103	520	34991
Oftalmologia	50	17.24 ± 10.96	2	10	17	21	53	862	52.94 ± 43.81	2	25.5	39	91.5	201	2647
Ortopedia	1306	22.15 ± 13.71	1	17	19	27	93	28928	58.18 ± 49.36	1	32	34	88.5	368	75988
Otorrinolaringologia	228	17.98 ± 14.83	1	7	15.5	23.5	64	4100	53.15 ± 54.64	1	12.5	32	100.5	397	12118
Urologia	628	20.58 ± 13.67	1	12	19	23	123	12923	60.00 ± 72.16	1	32	34	66	783	37680

3.3. Performance of the recommendation algorithm

3.3.1. Parameters tuning

In the development of the recommendation algorithm three parameters were considered: similarity measure (SM), similarity threshold (ST) and recommendation threshold (RT). To assess the best combination of parameters (tuning) for the algorithm, a 4-fold cross-validation was executed using a training set of 75% of the surgeries. Each parameter combination is represented as *SM* | *ST* | *RT*.

Table 8 shows the average and standard deviation of five performance measures obtained by the different parameter combinations in the 4-fold cross-validation. Hyperparameter settings with the same RT presented similar values regarding precision, recall, FPR and F1 Measure. Nonetheless, the best precision (0.664) and recall (0.992) were obtained by the combinations of *Jaccard* | *0* | *1* and *Cosine* | *0* | *0*, respectively. Increasing RTs improved the algorithm's precision and FPR, while deteriorating recall. The highest F1 measure (0.660) was achieved by a combination of *Jaccard* | *0* | *0.5*. Moreover, hyperparameter settings with the same combination of SM | ST demonstrated the same values of RMSE, being the lowest prediction error (6.855) achieved by *Cosine* | 0.2 | RT combination. *Jaccard* | *0* | *0.5* produced the best overall performance, i.e., best synergy between performance measures. Hence, the algorithm built with those parameters was selected for further testing and analysis.

Table 8: *Average and standard deviation of 4-fold cross-validation results for the different parameters combinations of the CF-based recommendation algorithm. FPR – False positive rate (1 - specificity); RMSE – Root mean square error; SD – Standard deviation. Dark green cells indicate the best value of a performance measure. Light green row indicates the parameters combination with the best overall performance.*

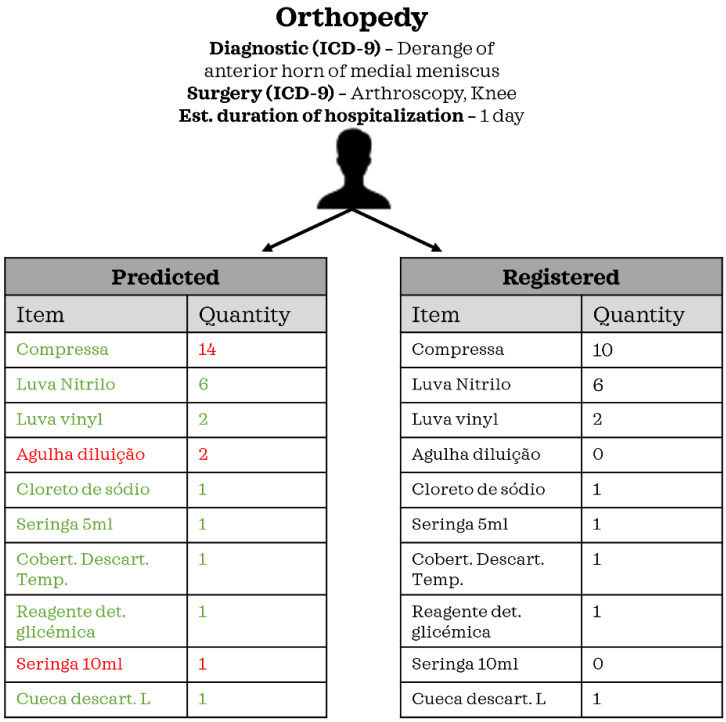
Settings			Precision	Recall	FPR	F1 Measure	RMSE
SM	**ST**	**RT**	**Mean ± SD**	**Mean ± SD**	**Mean ± SD**	**Mean ± SD**	**Mean ± SD**
Cosine	0	0	0.077 ± 0.001	0.992 ± 0.001	0.421 ± 0.009	0.144 ± 0.002	6.861 ± 0.199
Cosine	0	0.5	0.601 ± 0.010	0.719 ± 0.006	0.017 ± 0.000	0.654 ± 0.003	6.861 ± 0.199
Cosine	0	1	0.658 ± 0.006	0.251 ± 0.006	0.005 ± 0.000	0.363 ± 0.006	6.861 ± 0.199
Cosine	0.1	0	0.078 ± 0.001	0.991 ± 0.001	0.418 ± 0.009	0.145 ± 0.003	6.858 ± 0.199
Cosine	0.1	0.5	0.600 ± 0.010	0.719 ± 0.007	0.017 ± 0.000	0.654 ± 0.003	6.858 ± 0.199
Cosine	0.1	1	0.657 ± 0.006	0.251 ± 0.006	0.005 ± 0.000	0.363 ± 0.006	6.858 ± 0.199
Cosine	0.2	0	0.079 ± 0.001	0.991 ± 0.000	0.414 ± 0.009	0.146 ± 0.002	6.855 ± 0.199
Cosine	0.2	0.5	0.599 ± 0.010	0.719 ± 0.006	0.017 ± 0.000	0.654 ± 0.003	6.855 ± 0.199
Cosine	0.2	1	0.656 ± 0.006	0.251 ± 0.007	0.005 ± 0.000	0.363 ± 0.006	6.855 ± 0.199
Jaccard	0	0	0.111 ± 0.002	0.987 ± 0.000	0.281 ± 0.004	0.200 ± 0.003	6.871 ± 0.194
Jaccard	0	0.5	0.607 ± 0.011	0.723 ± 0.007	0.017 ± 0.001	0.660 ± 0.004	6.871 ± 0.194
Jaccard	0	1	0.664 ± 0.009	0.264 ± 0.007	0.005 ± 0.000	0.377 ± 0.006	6.871 ± 0.194

Note: Hyperparameter settings in which SM = 'Jaccard Index' and ST ≥ 0.1 are not represented, since there were patients from the test set that did not present Jaccard indexes ≥ 0.1 with any other patient, which impaired the cross-validation process.

3.3.2. Algorithm's overall performance

Table 9 depicts the model's results regarding six performance measures. The distribution of the predictions' residuals is depicted in Figure 1 from Appendix B. Figure 18 shows an example of a recommendation from the proposed algorithm.

Additionally, the algorithm's capacity of making recommendations for each medical specialty was assessed (Table 10). The best overall performance was observed in the predictions for 'Cirurgia Maxilo-Facial', which ranked first in all performance measures. On the other hand, recommendations for 'Cirurgia Pediátrica' presented the lowest quality. The algorithm achieved mostly values above 0.6 and 0.7 of precision and recall, respectively. Low values of FPR were obtained and F1 measure ranged between 0.236 and 0.898. Prediction errors varied between 1.638 and 8.432, being the highest RMSE observed in the recommendations for 'Cirurgia Vascular'.



Predicted	
Item	Quantity
Compressa	14
Luva Nitrilo	6
Luva vinyl	2
Agulha diluição	2
Cloreto de sódio	1
Seringa 5ml	1
Cobert. Descart. Temp.	1
Reagente det. glicémica	1
Seringa 10ml	1
Cueca descart. L	1

Registered	
Item	Quantity
Compressa	10
Luva Nitrilo	6
Luva vinyl	2
Agulha diluição	0
Cloreto de sódio	1
Seringa 5ml	1
Cobert. Descart. Temp.	1
Reagente det. glicémica	1
Seringa 10ml	0
Cueca descart. L	1

Figure 18: *Example of a recommendation, from the CF-based algorithm, for the first day of hospitalization of a patient diagnosed with a derange of the anterior horn of the medial meniscus, who was subjected to an arthroscopy in the knee, and to whom was estimated the need of one day of hospitalization. Green – item/quantity recommended correctly; Red - item/quantity recommended incorrectly.*

Table 9: *Overall results of the best recommendation algorithm. FPR – False positive rate (1 - specificity); AUC – Area under the receiving operating characteristic curve; RMSE – Root mean square error.*

Performance Measure	Value
Precision	0.608
Recall	0.729
FPR	0.017
F1-Measure	0.663
AUC	0.856
RMSE	6.901

Table 10: *Performance of the recommendation algorithm in predicting the needed medical items for patients of each medical specialty. FPR – False positive rate (1 - specificity); AUC – Area under the receiving operating characteristic curve; RMSE – Root mean square error.*

Medical Specialty	Precision	Recall	FPR	F1 Measure	AUC	RMSE
Otorrinolaringologia	0.448	0.637	0.021	0.526	0.808	6.989
Cirurgia Plástica Reconstrutiva e Estética	0.555	0.857	0.019	0.674	0.919	3.599
Cirurgia Vascular	0.591	0.697	0.021	0.640	0.838	8.432
Cirurgia Geral	0.619	0.760	0.017	0.682	0.872	7.823
Ginecologia-Obstetrícia	0.702	0.784	0.011	0.741	0.886	5.159
Cirurgia Pediátrica	0.177	0.357	0.036	0.236	0.661	5.234
Neuro-Cirurgia	0.636	0.720	0.016	0.676	0.852	7.031
Urologia	0.600	0.739	0.016	0.662	0.861	6.523
Ortopedia	0.636	0.708	0.014	0.670	0.847	6.199
Cirurgia Cardio-Torácica	0.515	0.690	0.037	0.590	0.827	7.586
Oftalmologia	0.441	0.736	0.022	0.552	0.857	7.308
Cirurgia Maxilo-Facial	0.822	0.989	0.007	0.898	0.991	1.638

3.3.3. Impact of the model

Figure 19 reveals the variation of four performance measures (AUC, F1 measure, precision, and recall) when adding medical items in the recommendation algorithm, ordered by their frequency of records (i.e., firstly the performance was assessed when only the most frequently registered 5% of the 602 items were added to the algorithm. Afterwards, the same analysis was performed when the most frequently registered 10% of the 602 items were considered, and so on.). Moreover, the respective proportions of medical item records and consumptions for each item set (i.e., 5%, 10%, etc.) are represented. It is noteworthy that the test set did not expend all medical items, wherefore the displayed records and consumptions only refer to the items of each item set that were used in the test set. Figure 2 from Appendix B shows the corresponding proportions of TP, TN, FP, and FN obtained by the algorithm predictions.

The algorithm exhibited a decreasing of the F1 measure, precision and recall when adding less frequently registered items, being the sharpest drop (F1 measure: -0.05; precision: -0.02; recall: -0.12) observed between 5% and 10%. Additionally, the performance plateaued after the addition of 20% of the items. The highest values were achieved when 5% of the medical items were considered (F1 measure: 0.75; precision:

0.64; recall: 0.92). In contrast, an increasing of AUC values occurred by adding more medical items in the algorithm.

These trends were also detected in the predictions of most medical specialty (Figure 3, Appendix B). However, the F1 measure, precision and recall of the recommendations for 'Cirurgia Maxilo-Facial', as well as the precision in 'Cirurgia Pediátrica', 'Ginecologia-Obstetrícia', 'Otorrinolaringologia', 'Oftalmologia', and 'Ortopedia' remained stable while adding more medical items to the algorithm. Furthermore, the best performance was obtained with 5% of the medical items in all medical specialties, except for 'Cirurgia Maxilo-Facial'.

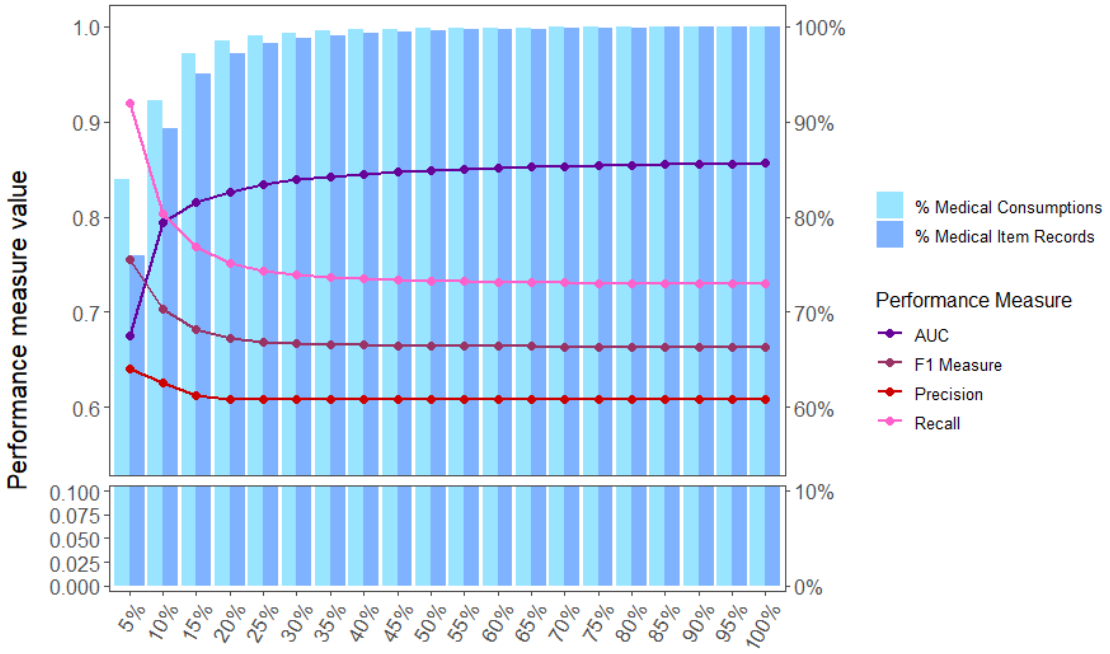


Figure 19: *Variation of four performance measures (AUC, F1 measure, precision, and recall) with the cumulative addition of medical items (ordered by their frequency of records). In each step, 5% of the 602 distinct medical items that could be recommended are added to the algorithm (x-axis). The percentage of medical item records and consumptions in the test set, covered by the items added in each step (right y-axis), are represented by the bars in the background. Note that the test set did not use all medical items, wherefore the medical records and consumptions only refer to items that were used in the test set. AUC - Area under the receiving operating characteristic curve.*

3.3.4. Long tail problem

To address the long tail problem, the set of medical items suggested by the algorithm was compared to the set of medical items that was registered. A total of 390 distinct items were expended during the first day of hospitalization of the test set. In contrast, 86 distinct items were suggested by the algorithm, of which 84 (97.67%) were present in the group of expended items. All twenty-one popular items were recommended at least once, while there was a recommendation of only 17.57% of the non-popular items. The same analysis was performed for each medical specialty individually, whose results are shown in Table 11.

Furthermore, there was an increase in the proportion of expended items that were never recommended with the cumulative inclusion of medical items (ordered by their frequency of records) in the recommendation algorithm, as it can be seen in Figure 20. The corresponding percentage regarding the total medical consumptions in the test set varied between 0.63% and 5.57%.

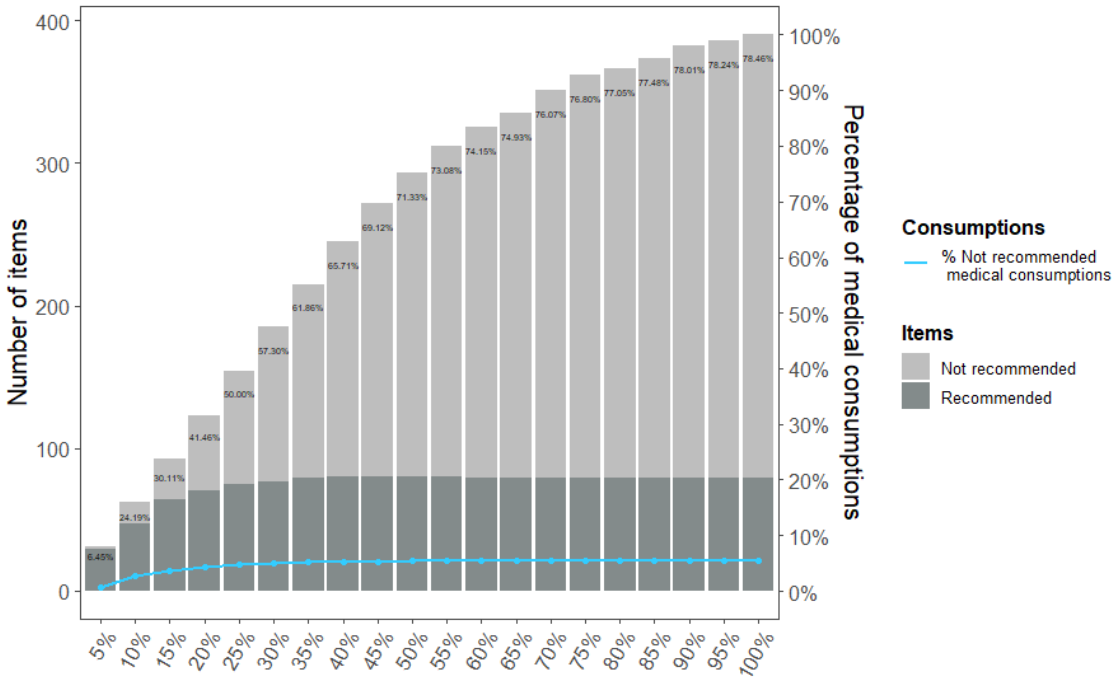


Figure 20: *Variation of the percentage of medical items that were not recommended with the cumulative addition of medical items (ordered by their frequency of records). In each step, 5% of the 602 medical items that could be recommended are added to the algorithm (x-axis). The percentage of medical consumptions in the test set covered by the items that were not recommended (right y-axis) are represented by the blue line. Note that the test set did not use all medical items, wherefore the medical records and consumptions only refer to items that were used in the test set.*

3.3.4.1. Recommendations overlap

Recommendations for surgeries that executed one of seven distinct surgical procedures, selected by the nurses of the hospital, were analyzed. To assess the overlap between recommendations for each surgical procedure, the sets of most frequently recommended medical items (i.e., items recommended for all surgeries in which the surgical procedure was performed) were identified and compared by computing the Jaccard index between pairs. The resulting scores varied from 0.87 to 1.00.

Table 11: *Summary of the results regarding the comparison between medical items that were consumed and recommended in the first day of hospitalization after a surgery of each medical specialty.*

Medical Specialty	N distinct items	N suggested items	N items suggested from the consumed items	% items suggested from the consumed items	% items not suggested	N popular items	% suggested popular items	% suggested non-popular items	N surgeries (test set)
Otorrinolaringologia	119	24	24	20.17	79.83	3.00	100.00	18.10	57
Cirurgia Plástica Reconstrutiva e Estética	80	61	60	75.00	25.00	17.00	100.00	69.84	33
Cirurgia Vascular	163	63	63	38.65	61.35	16.00	100.00	31.97	101
Cirurgia Geral	188	69	69	36.70	63.30	17.00	100.00	30.41	255
Ginecologia-Obstetrícia	102	41	41	40.20	59.80	17.00	100.00	28.24	148
Cirurgia Pediátrica	54	41	41	75.93	24.07	2.00	66.67	75.00	11
Neuro-Cirurgia	174	51	51	29.31	70.69	17.00	100.00	21.66	117
Urologia	149	49	49	32.89	67.11	17.00	100.00	24.24	157
Ortopedia	175	62	61	34.86	65.14	17.00	100.00	28.48	327
Cirurgia Cardio-Torácica	171	64	64	37.43	62.57	19.00	100.00	29.61	49
Oftalmologia	59	24	24	40.68	59.32	3.00	100.00	37.50	12
Cirurgia Maxilo-Facial	19	19	19	100.00	0.00	17.00	100.00	2.00	5

3.3.5. Algorithm's performance in recommending medical items individually

To assess the algorithm's performance in predicting the needed quantities of certain medical items, the RMSE was computed for the recommendations of each item individually. Table 12 shows the 20 medical items whose predictions achieved the highest RMSE values. The highest error was obtained in the predictions of the consumptions of 'Compressa N/Esteril Tnt 7,5x7,5cm 30gr' (RMSE = 33.91), followed by 'Luva Nitrilo M' and 'Agulha Diluição 19gx1 1/2 1,10x40' (RMSE = 14.45 and RMSE = 4.05, respectively). These tendencies were also observed in most medical specialties. RMSE values exhibited a strong positive correlation ($r = 0.88$, $p < 0.001$) with the variability in the consumptions of each medical item (Figure 21). Figure 4 from Appendix B demonstrates the differences between the total/average medical consumption and the total/average recommended quantity of each medical item.

Moreover, three performance measures (precision, recall and F1 measure) were computed regarding the predictions of the 20 most frequently registered medical items. As it can be seen in Figure 22, the top 17 items, which corresponded to 61.78% of the medical item records and 68.88% of the medical consumptions, were recommended with a precision, recall and F1 measure above 0.65, 0.90 and 0.80, respectively.

Table 12: *Top 20 medical items for which the predictions of their quantities achieved the highest RMSE values.*

Medical Item	RMSE
Compressa N/Estéril Tnt 7,5x7,5cm 30gr	33.91
Luva Nitrilo M	14.45
Agulha Diluição 19gx1 1/2 1,10x40	4.05
Compressa Estéril Tnt 10x10cm 40gr Pact 5	3.41
Seringa 5ml (2 peças)	2.67
Avental Plástico 80x125cm	2.48
Seringa 10ml (2 peças)	2.34
Cloreto de Sódio 0,9% IV Amp 10 ml	2.20
Luva Vinyl M	2.07
Sistema Administração Soro	1.84
Alcoól 70o Cut Fr 250ml	1.74
Esponja Higiene	1.65
Seringa 2ml (2 peças)	1.61
Tampa Vermelha Luer Lock Torneira e Seringa	1.55
Tampa Torneira 3 Vias	1.53
Luva Nitrilo L	1.45
Insulina humana 100 U.I./ml Ação curta Sol inj Fr 10 ml IV SC	1.40
Mascara Cirúrgica Descartável	1.19
Água para preparações injetáveis Sol inj Fr 10 ml	1.18

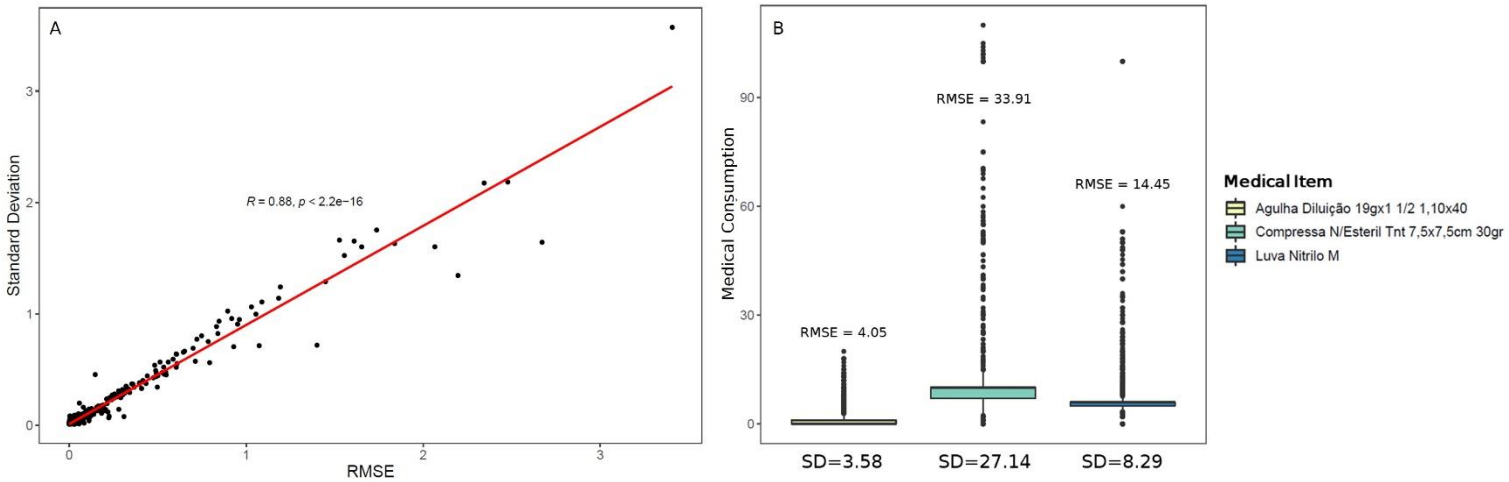


Figure 21: A) *Correlation between the RMSE in the recommendations performed by the algorithm and the standard deviation of the consumptions of each medical item (does not include the medical items represented in B for a better visualization);* ***B)*** *Distribution and standard deviation (SD) of the consumptions of the three medical items for which the predictions achieved the higher values of RMSE.*

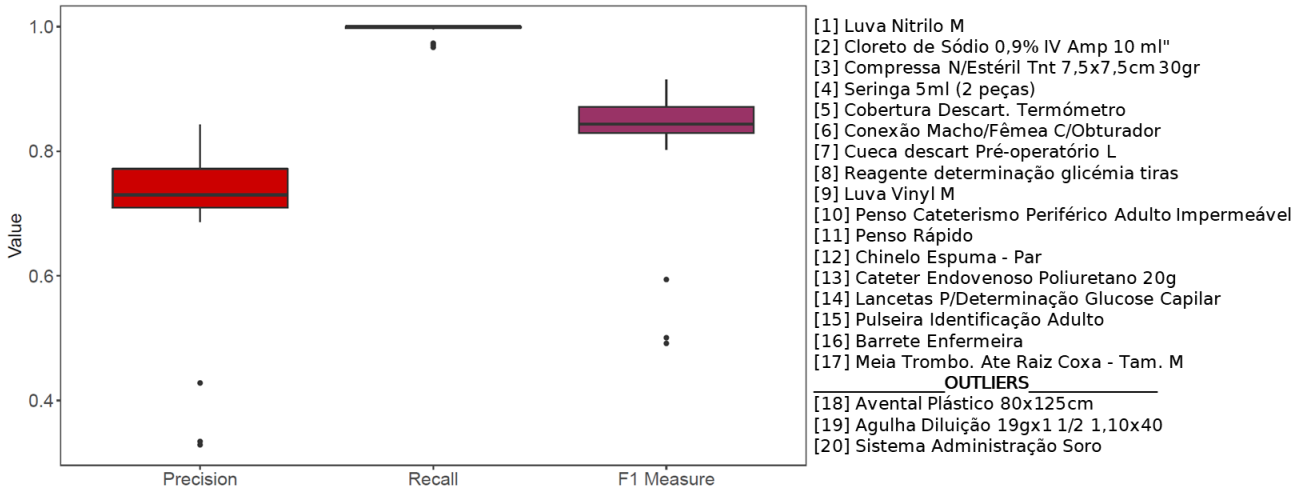


Figure 22: Left) *Distribution of three performance measures (precision, recall and F1 measure) regarding the predictions of the 20 most frequently registered medical items;* ***Right)*** *twenty most frequently registered medical items ranked by their number of records.*

In all medical specialties, except for 'Cirurgia Pediátrica', there was a cut-point in the medical items ranking from which the performance of the algorithm in recommending those items diminished substantially. In most medical specialties the items above the respective cut-point covered >65% medical item records and/or >55% of the medical consumptions. Half of the 20 most frequently registered items in 'Cirurgia Pediátrica' were not recommended once for the test set (Table 1, Appendix B). Additionally, significant differences were detected when comparing the distribution of the three performance measures regarding the predictions in all medical specialties (Precision: $H = 89.28$, $p < 0.001$; Recall: $H = 16.02$, $p = 0.09$; F1 measure: $H = 88.32$, $p < 0.001$) (Figure 23). Predictions for 'Oftalmologia' and 'Otorrinolaringologia' presented a significantly lower precision and F1 measure when compared to other medical specialties, while predictions for 'Cirurgia Maxilo-Facial' demonstrated significantly higher values of precision and F1 measure (Table 1, Table 2, Appendix B).

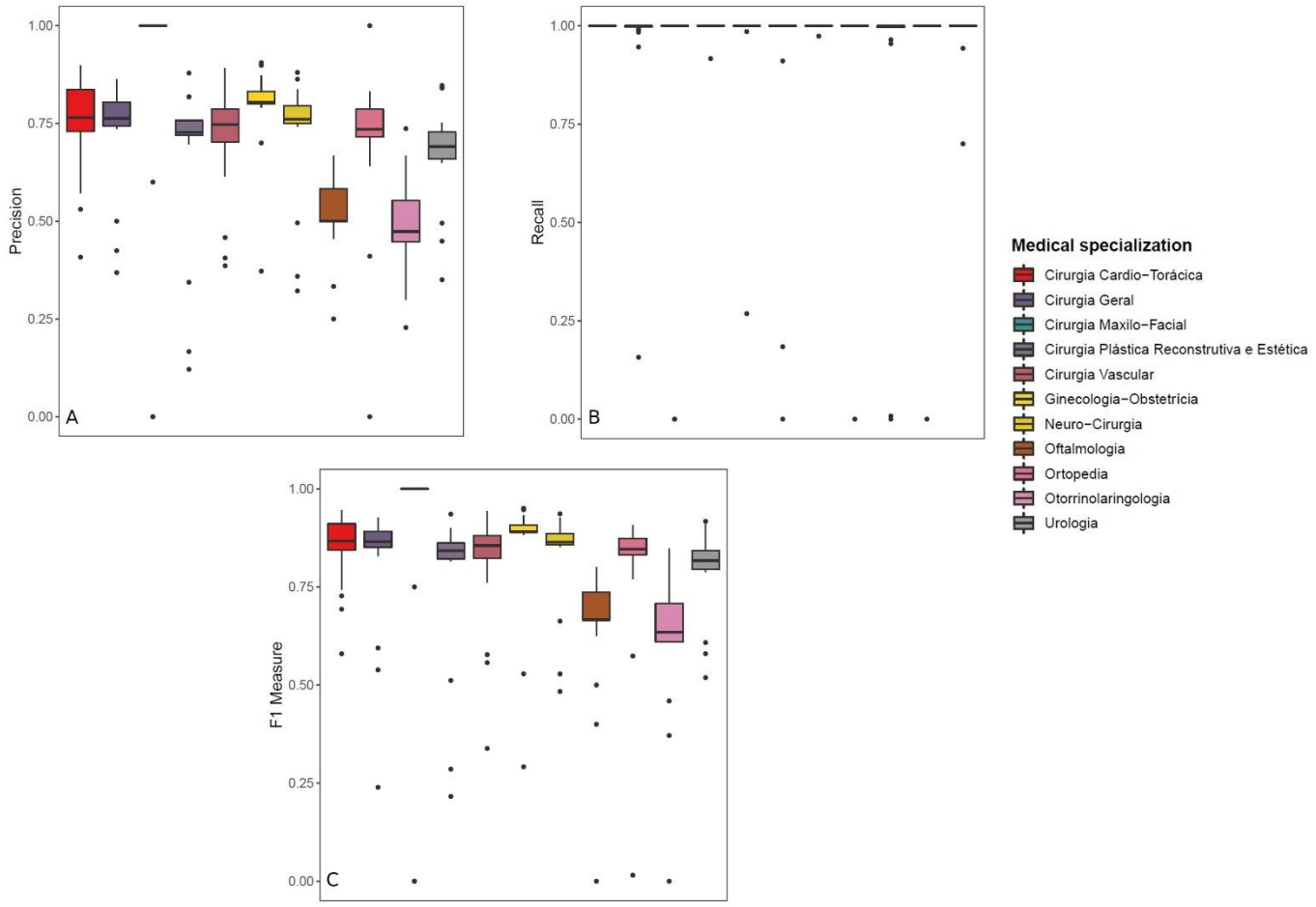


Figure 23: *Distribution of three performance measures (A - precision, B - recall and C - F1 measure) regarding the predictions of the 20 most frequently registered medical items in each medical specialty. 'Cirurgia Pediátrica' is not represented due to the high number of medical items that were not recommended once, by the algorithm.*

3.3.5.1. Variability in the recommandation algorithm's results for each medical sepcialization

To address the differences observed in the algorithm's performance in recommending medical items for surgeries of each medical specialty, the influence of the sample sizes (number of surgeries) were assessed. None of the four performance measures (RMSE, precision, recall and F1 measure) exhibited a significant correlation with the sample size of each medical specialty (Figure 5, Appendix B).

Furthermore, the variability of the contributions for the recommendations was explored. For that purpose, the similarity scores between surgeries, regarding patients' features and consumptions, were analyzed. Figure 24 describes the similiarities of surgeries from 'Cirurgia Pediátrica' and 'Cirurgia Maxilo-Facial' when compared to the

respective neighborhoods. Both medical specialties presented patient features similar (scores above 0) to the ones from other medical specialties. In 'Cirurgia Pediátrica' patient features similarities were mostly below 0.25. On the other hand, patient features from 'Cirurgia Maxilo-Facial' tended to be more similar to each other than to the other medical specialties. Additionally, medical item records and consumptions from all medical specialties were more similar to those from 'Cirurgia Maxilo-Facial' than to the ones from 'Cirurgia Pediátrica'.

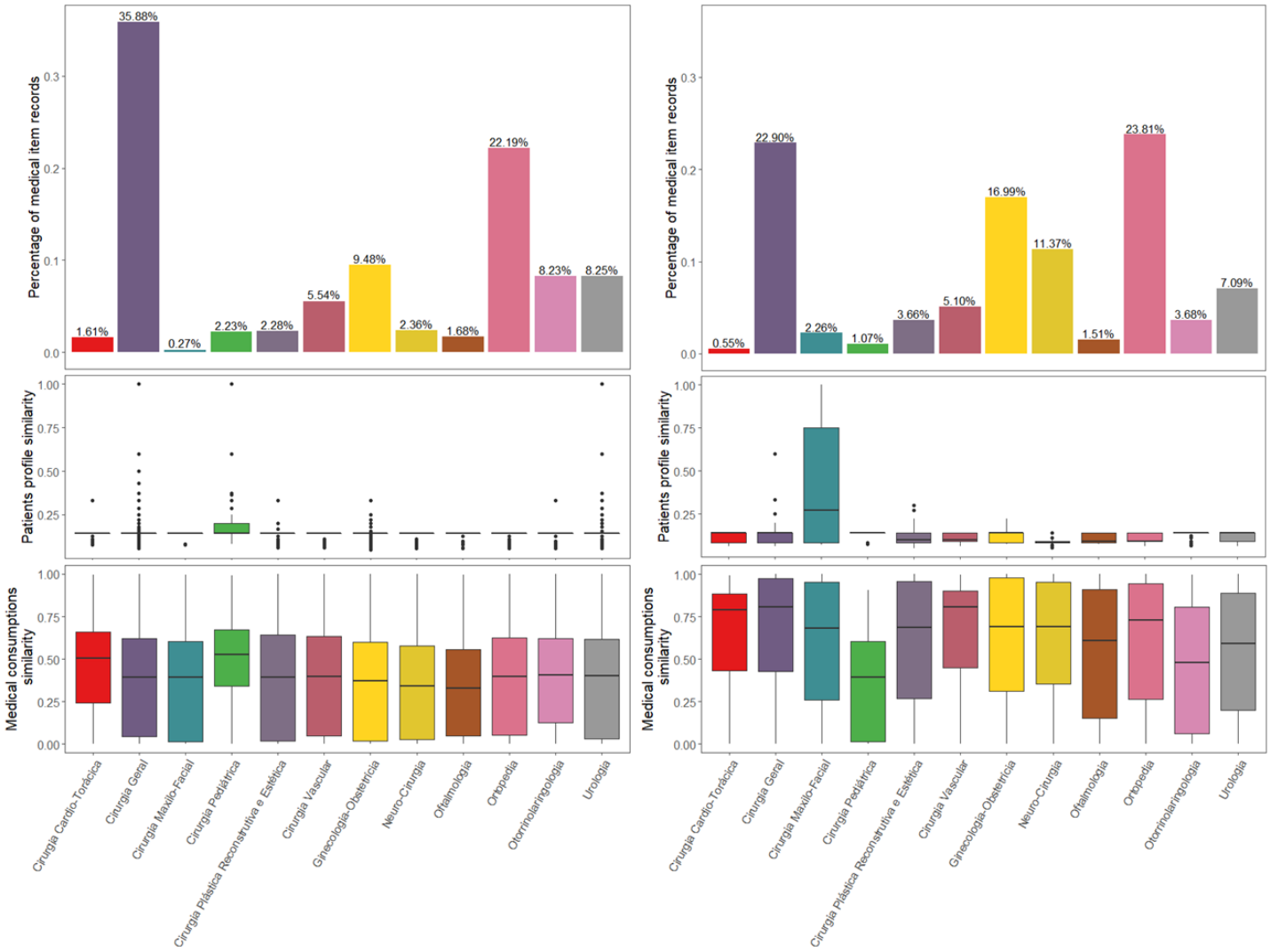


Figure 24: *Similiarities between surgeries from all medical specialties and surgeries from 'Cirurgia Pediátrica' (left) or 'Cirurgia Maxilo-Facial' (right). Only similarity scores above 0 were considered, i.e., surgeries that could contribute to the recommendations, for both medical specialties, performed by the algorithm.* ***Upper graphs)*** *Percentage of surgeries with patient features similar to the ones in 'Cirurgia Pediátrica' or 'Cirurgia Maxilo-Facial', from each medical specialty;* ***Middle graphs)*** *Distribution of the similarity scores between the patient features from 'Cirurgia Pediátrica' or 'Cirurgia Maxilo-Facial' and each medical specialty;* ***Lower graphs)*** *Distribution of the similarity scores between the medical item records/consumptions from 'Cirurgia Pediátrica' or 'Cirurgia Maxilo-Facial' and each medical specialty.*

3.3.6. Health professionals' evaluation

Table 13 reveals the answers for the questionnaire present in Appendix A given by two nurses of the hospital. Besides, some remarks were made in order to improve the results obtained by the algorithm, which are enumerated below:

- From the three variables used as an input for the algorithm, the surgical procedures are the most relevant to define which are the medical item records needed in the first day of hospitalization. For instance, the 15 cases that were provided to the nurses included repeated main surgical procedures. The exact same suggestions of records were given by the nurses for those cases, independently of the variation in the diagnosis or the estimated duration of hospitalization;
- Some rules to ensure the relationship between specific pairs of items should be included. For example, it does not make sense to register materials for intravenous administration if there is no need to use an intravenous medication;
- The algorithm seems to have some limitations in recommending drugs for the treatment of the inpatient.

Regarding the comparison between the algorithm recommendations and the suggestions of medical item records made by the nurses, a RMSE of 3.222, a precision of 0.606, a recall of 0.557, and a F1 measure of 0.581 were achieved. Table 14 shows the resulting confusion matrix. Additionally, Table 15 summarizes the performance of the algorithm's recommendations when compared to the real medical item records and the nurses suggestions, as well as the proximity of the nurse's suggestions to the real records.

Table 13: *Answers from both nurses to the questionnaire.*

Question	Answers	
	Nurse 1	**Nurse 2**
1) How do you rate the capacity of the recommendation system in recommending medical items correctly?	8	10
2) How do rate you the capacity of the recommendation system in recommending the quantities of the medical items correctly?	6	10
3) How many recommendations would you feel comfortable in registering?	15*	15*
4) After knowing the recommendations performed by the system, would you use it to assist you in registering medical items expended in the first day of hospitalization of a patient?	Yes*	Yes*
5) After knowing how the recommendation system works, would you use it to assist you in registering expended medical items?	No**	Yes
6) If the recommendation system explained why a given recommendation was made, would it make a difference in your choice to use it?	No***	Yes
7) Considering that the recommendations present a certain error rate, would you use this tool to support you in registering the standard consumptions in the first day of hospitalization after an elective surgery?	Yes	Yes

**After the corrections that were suggested.*
***When it comes to the recommendation of drugs.*
****Believes it is not necessary to explain why a given recommendations was formulated.*

Table 14: *Confusion matrix for the comparison between algorithm recommendations and the suggestions of medical item records made by the nurses.*

	Recommended	Not Recommended
Used	200	159
Not Used	130	*

**True negatives were not quantified since they were not necessary for this analysis.*

Table 15: *RMSE, precision, recall, and F1 measure when the algorithm's recommendations, real medical item records and nurses' suggestions are compared. The first comparison includes the 1272 surgeries from the test set, while the others include 15 surgeries selected considering the nurses criteria.*

Comparison	**Perfomance Measures**			
	RMSE	**Precision**	**Recall**	**F1 Measure**
Recommendations/ Real Medical Item Records (Test set)	6.901	0.608	0.729	0.663
Recommendations/ Nurses' Suggestions (15 surgeries)	3.222	0.606	0.557	0.581
Nurses' Suggestions/ Real Medical Item Records (15 surgeries)	5.663	0.277	0.477	0.351
Recommendations/ Real Medical Item Records (15 surgeries)	6.550	0.427	0.739	0.541

4. Discussion

The rapid adoption of EHRs by the healthcare systems (rates >90% in several countries) [116,117] led to the creation of large and heterogeneous databases [117]. Consequently, the growing capacity of data storage forces the development of methodologies that make use of that information so that it can be used to improve healthcare [117]. Big data methods, such as AI models, have the potential and are being widely used to assist healthcare professionals in the clinical decision-making [117,118], optimize the efficiency of healthcare tasks [89,118], and reduce the human-driven errors [118].

In this work, a recommendation algorithm was developed to perform data-driven predictions of the medical item records needed in the first day of hospitalization after an elective surgery, based on a CF method. The aim of this project is to optimize the time spent by healthcare professionals in the medical item registration process, which is a time-consuming daily responsibility that can influence their workload [89]. Additionally, during the period of the available data, approximately 25% of the surgeries did not present medical item records in the first day of hospitalization (Figure 14). This can arise from several reasons, such as oversight due to work overload from the nurses, lack of time/resources or by the fact that those surgeries were performed at the end of the day where the consumables are not justified. Therefore, the proposed recommendation algorithm is also intended to reduce the inconsistencies observed in the medical item registration process.

The classic CF approach in the e-commerce and entertainment domains consider that "if users shared the same interests in the past, then they would have similar tastes" [9,21,22]. In brief, the users' ratings/behavior in their shopping/searching on the internet are monitored and compared to other users, to predict if a given item would be of interest. Here, the following adaptation of that approach to the health domain is taken into account: "if patients share similar disease profiles/health conditions, then they would have similar treatments/healthcare services" [9]. Thus, analogously to other studies, the "user's profile" consisted in the patients' features which were used to formulate the recommendations [47,75,80,87], in particular to compute the similarities between patients.

From the 13699 surgeries performed during the stipulated period, only 5086 were considered for the development of the recommendation algorithm. Initially, three

inclusion criteria were applied to the dataset. Thus, only elective interventions were selected, since in a real scenario it is easier to collect all the patient's data associated to a planned surgery compared with an urgent intervention. Furthermore, surgeries that did not need post-operative hospitalization or were associated to more than one main surgical procedure were excluded. The first criterion ensured that only data from hospitalized patients would be considered, while the latter was applied to remove the complexity that those cases would add in the algorithm's predictions. Nonetheless, the proposed algorithm is a starting point, being the inclusion of more complex data important to enable the formulation of recommendations for a broader set of scenarios. Moreover, a proportion of surgeries did not present medical item records in the first day of hospitalization. It is possible that those records could have been registered cumulatively with records from subsequent days of hospitalization or could not have been registered at all. Hence, such surgeries were not included in the development of the recommendation algorithm since they could provide inaccurate data and, consequently, lead to an overestimation of the medical consumptions.

Although the medical consumptions analysis revealed that each patient registered an average of approximately 104 medical item records and consumed an average of 319 items during the whole period of hospitalization, the distribution of the data was highly dispersed (SD = 244.86 and SD = 839.11, respectively) indicating a high variability in the number of records and consumptions per patient. The same tendencies were observed in the first day of hospitalization, however, in a lower magnitude since only approximately 17% of both medical item records and consumptions were represented. This is further supported by the variability of records and consumptions observed within and between medical specialties (Table 7). In contrast, Figure 13 and Figure 15 show that most consumptions are associated to a small and consistent set of medical items. Interestingly, the item distribution regarding the frequency in which they were registered follows the rule of 80/20 (Pareto's Principle), which states that approximately 80% of the outcomes result from approximately 20% of the causes [37,119]. In fact, in the first day of hospitalization, 15.10% of the medical items covered 95.40% of the medical item records. Thus, a great proportion of the different medical items were consumed in low frequencies in the first day of hospitalization, representing the long tail of the medical item distribution [37,38,102].

Table 5 reveals the percentage of surgeries with missing medical item records in the first day of hospitalization concerning each medical specialty that perform elective surgeries in the hospital. Even though all medical specialties exhibited room for improvement, the results in Table 5 evidence a differentiated necessity in applying the proposed recommendation algorithm. For instance, 'Ortopedia' had the highest proportion of surgeries without records in the first day of hospitalization (N = 436, 26.27%), while 'Cirurgia Cardio-Torácia' only presented 16 surgeries (0.96%) with that condition. Hence, the first seems to need more the application of the algorithm to mitigate the lack of medical item records compared to the latter. It is expected that medical specialties that perform more surgeries are more likely to bring up errors in the medical item registration. Nonetheless, medical item registration tended to fail more within specialties with lower number of surgeries.

To optimize an algorithm's performance, its hyperparameters must be tuned so that the best settings are selected for the testing step [94,99–102]. In fact, this process is an important phase in the development of a recommendation system [101,102], not only for the hyperparameter tunning but also to select the best algorithm in cases where multiple models were built [102]. In this work, three hyperparameters (SM, ST, RT) were tuned by a 4-fold stratified cross-validation. The stratified random sampling of the folds secured the medical specialties ratio observed in the training set, which avoids a biased evaluation of the algorithm since the "samples proportion is an unbiased estimate of the population proportion" [99]. Moreover, Kohavi *et al.* [120] suggested a 10-fold stratified cross-validation concerning real world datasets, since lower k-folds increase the variability of the estimations. Nonetheless, the results of the 4-fold cross-validation (Table 8) demonstrated low variability in the performance measures between folds (SD $\in$ [0.000, 0.199]), indicating that the incorporation of more folds in each iteration would lead to an increase of the computational cost without significantly changing the results.

Furthermore, the variation of each hyperparameter's values influenced a specific group of performance measures. On one hand, the RT affected the *usage prediction* measures (precision, recall, FPR and F1 measure), since it defines the cutoff that will separate the classes of recommended and not recommended medical items. On the other hand, the SM and ST influenced the RMSE values, due to their direct role in the calculation of the Qsi. The SM and ST were built considering the assumptions that distinct similarity measures would find different neighborhoods for a new patient and that

the higher the similarity between two patients in terms of their features, the closer are both patients regarding their medical item records, respectively. However, the differences in the RMSE values are very mild between similarity measures and the increasing of the ST tend to slightly improve the quality of the predictions. Consequently, other methodologies could be implemented to select the neighborhood of a patient and, subsequently, assess the corresponding effect in the algorithm's performance. For instance, clustering methods were introduced in the CF as an alternative to the similarity computation in order to overcome some hurdles, such as the computational cost and data sparsity [121,122].

Moreover, the implementation of a ST seemed to hinder the algorithms' running, since at a certain ST the predictions were impaired for some patients due to the absence of similarity scores above the chosen value (Table 8). Therefore, instead of applying a threshold to select similarity scores that should be accounted for the predictions, the neighborhoods could be defined by considering only the *k*-top similarity scores associated to a new patient [21]. This approach would then restrict the neighborhoods, avoiding the attenuation of the contribution of some strong relations by many weak ones or the contribution of a small neighborhood, without impairing the predictions for new patients that only have similarity scores below a given ST. It is often reported in the literature that the best *k* lies between 20 and 50 [36], nevertheless, the "new hyperparameter" could also be determined by a cross-validation process [21].

The overall performance of the recommendation algorithm (Table 9) demonstrated that it performed correct predictions in 60.9% of the positive recommendations and was able to predict correctly 72.9% of the medical items in real records. Moreover, a low FPR (1.7%) was achieved, however, the imbalance between the classes of used and not used items (higher prevalence of not used items) may increase the probability of predicting TN (Figure 2, Appendix B) and, consequently, lead to an underestimation of the measure. Measures that summarize the information of more than one performance measure (F1 measure and AUC) are also useful in the evaluation of the algorithm [101,102]. Both the AUC (0.856) and the F1 measure (0.663) indicate a satisfactory performance of the algorithm predictions. Nevertheless, in the presence of an imbalanced dataset the selection and interpretation of those measurements should be taken with caution [105]. For instance, the AUC quantifies the capacity of the algorithm in ranking the scores of positive and negative classifications without accounting for the classification,

which might lead to misleading conclusions in terms of imbalanced datasets [105,123]. In contrast, values like the F1 measure [103] seem to be more adequate to assess the quality of the algorithms' predictions in such scenarios.

Regarding the predictions' accuracy, the RMSE reveals that the estimated quantities deviate an average of approximately 7 units from the real quantities of each record.Although the algorithm may mitigate the cases where the manual registration of medical item records is not possible, the observed deviations can go above or below the real medical consumptions (Figure 4, Appendix B). Thus, further changes in the algorithm should be considered to decrease the error. It is noteworthy that the RMSE disproportionately penalizes large errors [101,102,124] and may lead to an overestimation of the prediction errors due to the existence of some outliers in the residuals distribution (Figure 1, Appendix B).

As it is shown in Table 10, the algorithm revealed various levels of performance depending on the medical specialty of the surgery for which a recommendation was formulated. Initially, it was expected that the algorithm would have better results in the recommendations for the most preponderant medical specialties (i.e., with a higher number of performed surgeries) due to the availability of more information. This hypothesis was based on the assumption that the features and medical item records associated to surgeries performed in the same medical specialty were more similar to each other than to the surgeries performed in different medical specialties. However, neither the number of surgeries per medical specialty was significantly correlated with the algorithm's performance (Figure 5, Appendix B) nor the recommendations considered only data from surgeries of the same medical specialty as the new patient's surgery (Figure 24). The recommendations for 'Cirurgia Vascular', 'Cirurgia Geral' and 'Cirurgia Cardio-Torácica' presented the highest variabilities in the medical consumptions per surgery which may have induced the increased values of RMSE. Despite that, the recommendations for some medical specialties with lower medical consumptions' variability achieved higher values of RMSE (e.g., 'Oftalmologia') than other specialties that presented higher variability (e.g., 'Neuro-Cirurgia'). Hence, the predictions' quality might be influenced by other factors, as it will be further discussed. Interestingly, the best performance was obtained by the recommendations for 'Cirurgia Maxilo-Facial', where only 21 surgeries were performed. Even though, a good performance for this medical specialty does not have a substantial impact in the objectives that are being

addressed, exploring the corresponding data may give hints of how to improve the recommendations for other medical specialties.

Besides the overall performance, which is informative about the approximation of the recommendations and the real medical consumptions, exploring the actual impact that the algorithm would have in the hospital's resources management is also relevant. The evaluation of the recommendations when different portions of the medical items used in the test set were incorporated in the algorithm, ordered by the frequency in which they were registered, allowed to assess the variation of four performance measures between scenarios (Figure 19). For instance, the best values of precision, recall and F1 measure were achieved when only 5% of the medical items were considered, being a decrease of those measures observed in further segments of the algorithm. It is noteworthy that the aforementioned set of items covers ~75% and ~85% of all medical item records and consumptions, respectively. Thus, the algorithm has a moderate capacity in correctly identifying the need for items that correspond to the majority of the consumed resources in the first day of a patient's hospitalization. Howbeit, a clear limitation is perceived when there was an attempt to predict the need for medical items that were registered below a given frequency, as it can be seen by the sharp decay in the performance measures from the addition of 5% to 10% of the items. This matter will be discussed with more detail in the next paragraphs.

Furthermore, the variation of the AUC has shown an unexpected behavior. Interestingly, it increased with the addition of less popular items in the algorithm, in contrast to the other measures. Hence, this result demonstrates an improvement in the ability to discriminate both classes (i.e., recommended and not recommended items) when the evolution of the precision, recall and F1 measure says otherwise, representing a practical example of the possible misleading interpretations that AUC may give rise to in cases of imbalanced datasets [105,123]. AUC quantifies the capacity of the algorithm in ranking the scores (i.e., the probability of a given instance being from the positive class) of positive classifications above negative classifications, overlooking the threshold that differentiates the classes [105,106,123]. Thus, mainly in scenarios of imbalanced datasets, the ranking ability of the algorithm may be satisfactory which leads to high AUC values, while the classification of the instances may not be adequate. In this case, the proportion of instances in which an item should not be recommended is way more elevated than the opposite. The inclusion of less popular items has progressively widened the difference

between the percentages of both classes (Figure 2, Appendix B), which possibly increased the number of instances that should not be recommended ranked below the instances that should be recommended. Consequently, the frequency of correct rankings also increased which produced a raise of the AUC, even though the classifications by the algorithm have progressively deteriorated.

The accuracy of the suggestions of a recommendation system is the main focus when it comes to its evaluation [102,125–127]. However, it has been recognized that more quality criteria, such as the coverage, are needed to describe the performance of a recommendation algorithm [101,102,125–127]. Figure 20 illustrates the algorithm's coverage variation similarly to the segmented approach represented by Figure 19. As expected, the cumulative addition of progressively less consumed items led to the increase of the percentage of items that were not recommended once by the algorithm (up to 78.46%). Moreover, all popular items (short head) were recommended, while the algorithm only suggested 17.57% of the less consumed items (long tail). This scenario has been previously reported in the literature [37,101,102,119,125–128] in which long tail items are rarely or never recommended due to the sparsity of their usage data (i.e., medical item's table), impairing the algorithm's coverage, usage predictions and accuracy. Even though the portion of medical items that were not recommended only corresponds to <6% of the medical consumptions, the future improvement of the coverage is important to make sure that the records of more specific but not less important cases are also considered by the algorithm.

An evaluation of the recommendations of individual medical items was performed for a more detailed comparison between the algorithm's results and the real medical consumptions. It was observed that recommendations of medical items that are consumed more heterogeneously presented higher values of RMSE (Table 12 and Figure 21). Interestingly, the predictions of the three most consumed items led to considerably high values of RMSE. Since these items correspond to a substantial portion of the medical item records and consumptions, some improvements should be considered to decrease their RMSE. Nevertheless, it is important to underline that the existence of a small amount of outliers in the data (Figure 1, Appendix B and Figure 21B), which may occur due to specific medical consumptions that are different from the usual or errors in the registration process, strongly influence the computation of the RMSE [101,102,124]. Therefore, it is possible that some prediction errors are being overestimated. Furthermore, the expression

that calculates the medical consumptions tends to attenuate the outliers' impact in the predictions, leading to recommendations closer to quantities that were registered more often.

Regarding the classification component of the algorithm, a relatively good performance was achieved in identifying if some of the popular items, which represent most of the medical item records and consumptions, should or should not be recommended (precision ≥ 0.65, recall ≥ 0.90, and F1 measure ≥ 0.80). However, below the 17th most frequently used medical item, the quality of the algorithm's predictions sharply decreased. The observed segregation of the medical items into two groups might be a consequence of the long tail problem, since the recommendation of an item depends on the prediction of its needed quantity (i.e., if the recommended quantity is ≥ 0.50 or not). Hence, the lower the frequency in which an item is used, the more sparse is its data, driving to less accurate predictions [37,38] and, consequently, impairing the quality of the recommendations. These results could motivate the decomposition of the algorithm to solve two subproblems (i.e., recommendation of the popular items and/or recommendation of the items in the long tail). A possible approach could be an adaptation of the method proposed by Park, Y-J. *et al*. to "leverage the long tail" [38], in which the algorithm to recommend less popular items is based on clustering methodologies.

The previous analysis was also performed to test the quality of the algorithm's recommendation for the first day of hospitalizations after surgeries of each medical specialty. In all medical specialties, except for 'Cirurgia Pediátrica', a shift in the algorithm's performance in recommending medical items at a given cut-point was also observed. Half of the 20 most frequently registered medical items in 'Cirurgia Pediátrica' were not recommended once for the test set (Table 1, Appendix B) which, together with the results in Table 10, indicates that the proposed algorithm may not be adequate to formulate recommendations for this medical specialty due to its particularities. For instance, the predictions for 'Cirurgia Pediátrica' were based on data from not only known pediatric patients, but also from known patients that also underwent surgeries from other medical specialties (Figure 24). Thus, the recommendations were computed with contributions from medical consumptions of adults that may be distinct from the pediatric consumptions, causing noise in the estimations. The addition of some features, such as the age of the patients, could restrict the neighborhood of such cases and, consequently, decrease its intra-variability. On the other hand, the comparison between the other 11

medical specialties (Figure 23) revealed differences in the precision ($H = 89.28$, $p < 0.001$) and F1 measure (F1 measure: $H = 88.32$, $p < 0.001$) of the algorithm in identifying recommendable and not recommendable popular items (Table 2, Table 3, Appendix B). Wherefore, the algorithm leads to different proportions of false positives depending on the medical specialty.

Taking this into account together with the results displayed in Table 10, the neighborhoods of new patients from 'Cirurgia Pediátrica' and 'Cirurgia Maxilo-Facial' (i.e., the medical specialty for which the algorithm had the worst and best performance, respectively) were explored (Figure 24) to pinpoint some factors that might influence the quality of the recommendations. Notoriously, the neighborhoods of new patients from 'Cirurgia Maxilo-Facial' tended to contribute for the recommendations with more similar consumption patterns and, therefore, with less intra-variability than the neighborhoods of 'Cirurgia Pediátrica'. Furthermore, the similarity scores between features and consumptions of new patients and known patients that underwent a surgery in 'Cirurgia Maxilo-Facial' tended to be higher than the ones between new patients from 'Cirurgia Maxilo-Facial' and known patients from other medical specialties, being the contributions stronger and more accurate in the first case compared to the latter. Although not so clearly, the same trend was observed in 'Cirurgia Pediátrica', which reinforces the idea of a better contribution of known patients that had a surgery in the same medical specialty of the new patient. Consequently, for the future it should be evaluated the possibility of building a recommendation system that is specialty specific.

Building a well-founded recommendation system is not enough to make users trust and accept the advantages of such algorithms [108,129]. Trust is defined by Parasuraman, R. and Miller, C. as the "users' willingness to believe in the information from a system or make use of its capabilities" [129], hence, it depends not only on the accuracy of the recommendations [101,130], but also on the users' beliefs about the system and the user-system interaction [108,131]. Here, two nurses were provided with the recommendations formulated by the algorithm for 15 surgeries. Afterwards, they were asked to make their own recommendations for each case and to answer a questionnaire (Appendix A) so that it was possible to assess their confidence in the system. Regarding the ability of the algorithm to predict which medical items are needed (Question 1) and the respective quantities (Question 2), nurse 2 was more optimistic about the outputs than nurse 1, however, both would accept all 15 recommendations (Question 3) after the application of

the suggestions/comments present in the results. The referred remarks suggest the decrease in the complexity of the data that is used to find the neighborhood of a new patient, the implementation of rule-based methods to ensure that some items are linked by their usage relationship, and the need to give more relevance to the drugs that should be administered in the first day of hospitalization of an inpatient. Furthermore, before knowing the conceptual functioning of the algorithm both nurses stated that they would use the recommendation system to help in the medical consumptions registration (Question 4). A better understanding of the algorithms and providing explanations of why a recommendation was made may improve the users trust [101,108]. It also facilitates the identification of possible errors or components that should be upgraded [20]. For instance, nurses 1 and 2 had divergent opinions about using the system after knowing how it works (Question 5). Nurse 2 would still use the algorithm, while nurse 1 considers that the way it was built may not be adequate to predict the inpatients need for medication. Moreover, implementing explanations to justify a recommendation does not always help to improve users trust [108], as it can be seen by the answer of nurse 1 who considers that it would not be necessary (Question 6). However, in cases such as nurse 2 who stated that a clarification of a given recommendation would be helpful, a careful and clear description should be formulated since "poorly designed explanations" can have a negative effect in terms of trust in the system [20]. Lastly, although the algorithm is not 100% accurate, both nurses would still use it to support them in registering the medical consumptions of the most frequently consumed medical items (Question 7).

Regarding the comparison between the algorithm's suggestions and the nurses' recommendations, a lower RMSE was achieved than the one obtained in the comparison between the recommendations and the real medical item records. Moreover, the usage prediction measures were worse in the second scenario, except for recall which was higher than the first case. These results reveal that the system is closer to the medical item records that should be registered in the healthcare professionals' point of view. It should be noted that the nurses only gave their input about 15 surgeries which is a small sample compared with the test set used to evaluate the algorithm. Therefore, the observed results may change with the approximation in the number of surgeries used in both validation methods.

Lastly, the results of the comparison between the nurses' suggestions and the real medical item records reveal inconsistencies not only regarding the medical items that

should be recommended but also in terms of their quantities. Even though this analysis probably does not give a full insight of the differences between the real records and the nurses' suggestions due to the low number of surgeries that were explored, it seems that there is a gap between the empirical perception of the healthcare professionals and what is really consumed/registered. This may be derived from errors in the data that occurred during the registration process and, therefore, the nurses' input should be used to improve data quality and, consequently, reach the best possible recommendation system.

5. Conclusions

The results of this project suggest that it is possible to predict and recommend the medical consumptions needed in the first day of hospitalization of a patient to a certain extent using some of their features as predictors. In fact, the algorithm demonstrated a moderate performance in recommending the most frequently consumed medical items which correspond to most of the medical item records and consumptions in the hospital during the selected period. However, the data as it is, and the proposed model still present limitations in discriminating scenarios in which less popular medical items are necessary or not. Thus, exploring different approaches and methodologies to complement or modify the algorithm should be considered.

A more detailed analysis of the obtained results revealed that the model's performance in recommending medical items depends on the medical specialty in which the surgery will be performed. Since the amount of data available for each medical specialty did not correlate with the recommendations' quality, there are other specialty-specific factors that influence the algorithm's performance. Among others, these might be age-related factors and/or the intrinsic variability of the medical items and consumptions needed after the surgeries executed in each medical specialty. Therefore, in the future a specialty-specific adaptation of the present algorithm should be tested.

In the perspective of the healthcare professionals, there is some degree of trust in the results of the algorithm, although there were deviations from what they believed to be the best group of medical item records for a patient who underwent a given surgery. The application of some suggestions and rules in the algorithm was considered crucial, indicating the need for extra steps in the model's tuning to achieve the full confidence of recommendation system's end-users. Moreover, the real medical item records showed some deviations when compared with the nurses' perspective. Hence, although the predictions are formulated in a data-driven way, the observations made by the healthcare professionals must be taken into account to mitigate inconsistencies in the data and, consequently, reduce the bias in the results.

In summary, despite the limitations inherent to an academic project, the developed work concluded successfully an end-to-end project in a near-real world scenario, from

data analysis, model development and end-user test/feedback. Further work is required to scale the developments and measure its impact in the nurses' workload.

6. Final Remarks

The opportunity to develop a project in Hospital da Luz Learning Health. during my curricular internship allowed me to grow in the field of data analysis. In particular, it allowed me to widen my knowledge about different methods to deal and analyze real world data. Moreover, the preliminary theoretical research that I had to perform regarding recommendation systems, the contact with colleagues from other institutions, and the attendance in some of the several seminars provided by Hospital da Luz Lisboa helped me realize the growing relevance of the Artificial Intelligence in the health domain.

The continuous sharing of results, ideas, solutions, and suggestions for my project, both in the weekly remote sessions and during the period in which I worked in Hospital da Luz Learning Health in person, led to an improvement of my critical thinking, communication and troubleshooting abilities. Additionally, by listening and intervening in the discussion of my colleagues' projects I was able to apply my statistical and machine learning knowledge in other contexts.

Developing a data-driven decision-making tool requires a rigorous management of the data and a good knowledge of the programming software that are being used. Taking that into account, during this work I gradually became more aware of the attention needed in the phase of data preparation, which is essential to maximize its potential in providing information. Furthermore, it allowed me to learn and improve my programming skills not only in Python but also in R.

Lastly, the analysis of the obtained results increased the desire to study more deeply the theory underlying the methodologies of model evaluation in order to identify their advantages and limitations in different scenarios. Moreover, the follow up from my advisors taught me not to restrict my analysis by only performing methodologies by the book. Data can be visualized and investigated in more than one perspective, enabling the extraction of different types of information which in turn can answer to different questions.

References

1. Luz Saúde. Luz Saúde: vision, mission and values. (2022). Available at: https://www.hospitaldaluz.pt/en/about-us/luz-saude-vision-mission-and-values. (Accessed: 22nd July 2022)

2. Luz Saúde. Hospital da Luz Learning Health. (2022). Available at: https://www.hospitaldaluz.pt/learninghealth/en/learning-health/about-us/hospital-da-luz-learning-health. (Accessed: 22nd July 2022)

3. Luz Saúde. Professional training areas. (2022). Available at: https://www.hospitaldaluz.pt/learninghealth/en/training/training-areas-for-professionals. (Accessed: 22nd March 2022)

4. Luz Saúde. Research Areas. (2022). Available at: https://www.hospitaldaluz.pt/learninghealth/en/research/research-areas. (Accessed: 22nd March 2022)

5. Montaner, M., López, B. & de la Rosa, J. L. A Taxonomy of Recommender Agents on the Internet. *Artif. Intell. Rev.* **19**, 285–330 (2003).

6. Fayyaz MahsaAU - Nawara, DinaAU - Ibrahim, AhmedAU - Kashef, RashaTI - Recommendation Systems: Algorithms, Challenges, Metrics, and Business Opportunities, Z.-E. Recommendation Systems: Algorithms, Challenges, Metrics, and Business Opportunities. *Applied Sciences* **10**, (2020).

7. Aggarwal, C. C. An Introduction to Recommender Systems BT - Recommender Systems: The Textbook. in (ed. Aggarwal, C. C.) 1–28 (Springer International Publishing, 2016). doi:10.1007/978-3-319-29659-3_1

8. Ricci, F., Rokach, L. & Shapira, B. Introduction to Recommender Systems Handbook BT - Recommender Systems Handbook. in (eds. Ricci, F., Rokach, L., Shapira, B. & Kantor, P. B.) 1–35 (Springer US, 2011). doi:10.1007/978-0-387-85820-3_1

9. Tran, T. N. T., Felfernig, A. & Tintarev, N. Recommender systems in the healthcare domain: state-of-the-art and research issues. *ACM Trans. Interact. Intell. Syst.* **11**, 171–201 (2021).

10. Goldberg, D., Nichols, D., Oki, B. M. & Terry, D. Using Collaborative Filtering to Weave an Information Tapestry. *Commun. ACM* **35**, 61–70 (1992).

11. Wiesner, M. & Pfeifer, D. Health recommender systems: concepts, requirements, technical basics and challenges. *Int. J. Environ. Res. Public Health* **11**, 2580–2607 (2014).

12. Hardesty, L. The history of Amazon's recommendation algorithm. *Amazon Science* 1 (2019). Available at: https://www.amazon.science/the-history-of-amazons-recommendation-algorithm. (Accessed: 12th March 2022)

13. Davidson, J. *et al.* The YouTube Video Recommendation System. in *Proceedings of the Fourth ACM Conference on Recommender Systems* 293–296 (Association for Computing Machinery, 2010). doi:10.1145/1864708.1864770

14. Koren, Y., Bell, R. & Volinsky, C. Matrix Factorization Techniques for Recommender Systems. *Computer (Long. Beach. Calif).* **42**, 30–37 (2009).

15. ACM. The ACM conference on Recommender Systems. (2020). Available at: https://recsys.acm.org/. (Accessed: 12th March 2022)

16. Jannach, D., Zanker, M. & Konstan, J. Special issue on Recommender Systems. *AI Commun.* **21**, 95–96 (2008).

17. Felfernig, A., Friedrich, G. & Schmidt-Thieme, L. Guest Editors' Introduction: Recommender Systems. *IEEE Intell. Syst.* **22**, 18–21 (2007).

18. Virvou, M. & Tsihrintzis, G. A. Special Issue on Advances in Recommender Systems. *Intell. Decis. Technol.* **9**, 219–220 (2015).

19. Schafer, J. Ben, Frankowski, D., Herlocker, J. & Sen, S. Collaborative Filtering Recommender Systems BT - The Adaptive Web: Methods and Strategies of Web Personalization. in (eds. Brusilovsky, P., Kobsa, A. & Nejdl, W.) 291–324 (Springer Berlin Heidelberg, 2007). doi:10.1007/978-3-540-72079-9_9

20. Herlocker, J. L., Konstan, J. A. & Riedl, J. Explaining Collaborative Filtering Recommendations. in *Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work* 241–250 (Association for Computing Machinery, 2000). doi:10.1145/358916.358995

21. Desrosiers, C. & Karypis, G. A Comprehensive Survey of Neighborhood-based Recommendation Methods BT - Recommender Systems Handbook. in (eds. Ricci, F., Rokach, L., Shapira, B. & Kantor, P. B.) 107–144 (Springer US, 2011). doi:10.1007/978-0-387-85820-3_4

22. Aggarwal, C. C. Neighborhood-Based Collaborative Filtering BT - Recommender Systems: The Textbook. in (ed. Aggarwal, C. C.) 29–70 (Springer International Publishing, 2016). doi:10.1007/978-3-319-29659-3_2

23. Koren, Y. & Bell, R. Advances in Collaborative Filtering BT - Recommender Systems Handbook. in (eds. Ricci, F., Rokach, L., Shapira, B. & Kantor, P. B.) 145–186 (Springer US, 2011). doi:10.1007/978-0-387-85820-3_5

24. Aggarwal, C. C. Model-Based Collaborative Filtering BT - Recommender Systems: The Textbook. in (ed. Aggarwal, C. C.) 71–138 (Springer International Publishing, 2016). doi:10.1007/978-3-319-29659-3_3

25. Thakkar, P., Varma, K., Ukani, V., Mankad, S. & Tanwar, S. Combining User-Based and Item-Based Collaborative Filtering Using Machine Learning BT - Information and Communication Technology for Intelligent Systems. in (eds. Satapathy, S. C. & Joshi, A.) 173–180 (Springer Singapore, 2019).

26. Zhao, Z.-D. & Shang, M. User-Based Collaborative-Filtering Recommendation Algorithms on Hadoop. in *2010 Third International Conference on Knowledge Discovery and Data Mining* 478–481 (2010). doi:10.1109/WKDD.2010.54

27. Sarwar, B., Karypis, G., Konstan, J. & Riedl, J. Item-Based Collaborative Filtering Recommendation Algorithms. in *Proceedings of the 10th International Conference on World Wide Web* 285–295 (Association for Computing Machinery, 2001). doi:10.1145/371920.372071

28. Zhang, J., Lin, Z., Xiao, B. & Zhang, C. An optimized item-based collaborative filtering recommendation algorithm. in *2009 IEEE International Conference on Network Infrastructure and Digital Content* 414–418 (2009). doi:10.1109/ICNIDC.2009.5360986

29. Do, M.-P., Nguyen, D. & Nguyen, L. *Model-based approach for Collaborative Filtering*. (2010).

30. Milli, M. & Bulut, H. The Effect of Neighborhood Selection on Collaborative Filtering and a Novel Hybrid Algorithm. *Intell. Autom. Soft Comput.* **23**, 261–269 (2017).

31. Hug, N. k-NN inspired algorithms. (2015). Available at: https://surprise.readthedocs.io/en/stable/prediction_algorithms_package.html. (Accessed: 2nd September 2022)

32. Lacic, E., Kowald, D. & Lex, E. Neighborhood Troubles: On the Value of User Pre-Filtering To Speed Up and Enhance Recommendations. *CoRR* **abs/1808.0**, (2018).

33. Gong, S., Ye, H. & Tan, H. Combining Memory-Based and Model-Based Collaborative Filtering in Recommender System. in *2009 Pacific-Asia Conference on Circuits, Communications and Systems* 690–693 (2009). doi:10.1109/PACCS.2009.66

34. Nilashi, M., Ibrahim, O. & Bagherifard, K. A recommender system based on collaborative filtering using ontology and dimensionality reduction techniques. *Expert Syst. Appl.* **92**, 507–520 (2018).

35. Zarzour, H., Maazouzi, F., Soltani, M. & Chemam, C. An Improved Collaborative Filtering Recommendation Algorithm for Big Data BT - Computational Intelligence and Its Applications. in (eds. Amine, A., Mouhoub, M., Ait Mohamed, O. & Djebbar, B.) 660–668 (Springer International Publishing, 2018).

36. Herlocker, J., Konstan, J. A. & Riedl, J. An Empirical Analysis of Design Choices in Neighborhood-Based Collaborative Filtering Algorithms. *Inf. Retr. Boston.* **5**, 287–310 (2002).

37. Yin, H., Cui, B., Li, J., Yao, J. & Chen, C. Challenging the Long Tail Recommendation. *Proc. VLDB Endow.* **5**, 896–907 (2012).

38. Park, Y.-J. & Tuzhilin, A. The Long Tail of Recommender Systems and How to Leverage It. in *Proceedings of the 2008 ACM Conference on Recommender Systems* 11–18 (Association for Computing Machinery, 2008). doi:10.1145/1454008.1454012

39. Park, Y.-J. The Adaptive Clustering Method for the Long Tail Problem of Recommender Systems. *IEEE Trans. Knowl. Data Eng.* **25**, 1904–1915 (2013).

40. Chen, Y., Li, X. & Zhang, S. Structured Latent Factor Analysis for Large-scale Data: Identifiability, Estimability, and Their Implications. *J. Am. Stat. Assoc.* **115**, 1756–1770 (2020).

41. Isinkaye, F. O., Folajimi, Y. O. & Ojokoh, B. A. Recommendation systems: Principles, methods and evaluation. *Egypt. Informatics J.* **16**, 261–273 (2015).

42. Gower, S. M. Netflix Prize and SVD. in (2014).

43. George, T. & Merugu, S. A scalable collaborative filtering framework based on co-clustering. in *Fifth IEEE International Conference on Data Mining (ICDM'05)* 4 pp. (2005). doi:10.1109/ICDM.2005.14

44. Shani, G., Heckerman, D. & Brafman, R. I. An MDP-Based Recommender System. *J. Mach. Learn. Res.* **6**, 1265–1295 (2005).

45. Tsoukalas, A., Albertson, T. & Tagkopoulos, I. From data to optimal decision making: a data-driven, probabilistic machine learning approach to decision support for patients with sepsis. *JMIR Med. informatics* **3**, e11 (2015).

46. Hug, N. Matrix Factorization-based algorithms. (2015). Available at: https://surprise.readthedocs.io/en/stable/matrix_factorization.html. (Accessed: 13th September 2022)

47. Stark, B., Knahl, C., Aydin, M., Samarah, M. & Elish, K. O. BetterChoice: A migraine drug recommendation system based on Neo4J. in *2017 2nd IEEE International Conference on Computational Intelligence and Applications (ICCIA)* 382–386 (2017).

doi:10.1109/CIAPP.2017.8167244

48. Wang, D., Liang, Y., Xu, D., Feng, X. & Guan, R. A content-based recommender system for computer science publications. *Knowledge-Based Syst.* **157**, 1–9 (2018).

49. Meteren, R. Van & Someren, M. Van. Using Content-Based Filtering for Recommendation. *ECML/MLNET Work. Mach. Learn. New Inf. Age* 47–56 (2000).

50. Lops, P., de Gemmis, M. & Semeraro, G. Content-based Recommender Systems: State of the Art and Trends BT - Recommender Systems Handbook. in (eds. Ricci, F., Rokach, L., Shapira, B. & Kantor, P. B.) 73–105 (Springer US, 2011). doi:10.1007/978-0-387-85820-3_3

51. Aggarwal, C. C. Content-Based Recommender Systems BT - Recommender Systems: The Textbook. in (ed. Aggarwal, C. C.) 139–166 (Springer International Publishing, 2016). doi:10.1007/978-3-319-29659-3_4

52. Pazzani, M. J. & Billsus, D. Content-Based Recommendation Systems BT - The Adaptive Web: Methods and Strategies of Web Personalization. in (eds. Brusilovsky, P., Kobsa, A. & Nejdl, W.) 325–341 (Springer Berlin Heidelberg, 2007). doi:10.1007/978-3-540-72079-9_10

53. Lops, P. *et al.* Learning Semantic Content-Based Profiles for Cross-Language Recommendations. in *Proceedings of the First Workshop on Personalised Multilingual Hypertext Retrieval* 26–33 (Association for Computing Machinery, 2011). doi:10.1145/2047403.2047409

54. Christian, H., Agus, M. P. & Suhartono, D. Single Document Automatic Text Summarization Using Term Frequency-Inverse Document Frequency (TF-IDF). *ComTech* **7**, (2016).

55. Aberg, J. Dealing with Malnutrition: A Meal Planning System for Elderly. in *AAAI Spring Symposium: Argumentation for Consumers of Healthcare* (2006).

56. Felfernig, A. & Burke, R. Constraint-Based Recommender Systems: Technologies and Research Issues. in *Proceedings of the 10th International Conference on Electronic Commerce* (Association for Computing Machinery, 2008). doi:10.1145/1409540.1409544

57. Felfernig, A., Friedrich, G., Jannach, D. & Zanker, M. Developing Constraint-based Recommenders BT - Recommender Systems Handbook. in (eds. Ricci, F., Rokach, L., Shapira, B. & Kantor, P. B.) 187–215 (Springer US, 2011). doi:10.1007/978-0-387-85820-3_6

58. Smyth, B. Case-Based Recommendation BT - The Adaptive Web: Methods and Strategies of Web Personalization. in (eds. Brusilovsky, P., Kobsa, A. & Nejdl, W.) 342–376 (Springer Berlin Heidelberg, 2007). doi:10.1007/978-3-540-72079-9_11

59. Aggarwal, C. C. Knowledge-Based Recommender Systems BT - Recommender Systems: The Textbook. in (ed. Aggarwal, C. C.) 167–197 (Springer International Publishing, 2016). doi:10.1007/978-3-319-29659-3_5

60. BRIDGE, D., GÖKER, M. H., McGINTY, L. & SMYTH, B. Case-based recommender systems. *Knowl. Eng. Rev.* **20**, 315–320 (2005).

61. Doulaverakis, C., Nikolaidis, G., Kleontas, A. & Kompatsiaris, I. GalenOWL: Ontology-based drug recommendations discovery. *J. Biomed. Semantics* **3**, 14 (2012).

62. Doulaverakis, C., Nikolaidis, G., Kleontas, A. & Kompatsiaris, I. Panacea, a semantic-enabled drug recommendations discovery framework. *J. Biomed. Semantics* **5**, 13

(2014).

63. Faiz, I., Mukhtar, H. & Khan, S. An integrated approach of diet and exercise recommendations for diabetes patients. in *2014 IEEE 16th International Conference on e-Health Networking, Applications and Services (Healthcom)* 537–542 (2014). doi:10.1109/HealthCom.2014.7001899

64. Mahmoud, N. & Elbeh, H. IRS-T2D: Individualize Recommendation System for Type2 Diabetes Medication Based on Ontology and SWRL. in *Proceedings of the 10th International Conference on Informatics and Systems* 203–209 (Association for Computing Machinery, 2016). doi:10.1145/2908446.2908495

65. Shimada, K. *et al.* Drug-recommendation system for patients with infectious diseases. *AMIA ... Annu. Symp. proceedings. AMIA Symp.* **2005**, 1112 (2005).

66. Shen, Y. *et al.* An ontology-driven clinical decision support system (IDDAP) for infectious disease diagnosis and antibiotic prescription. *Artif. Intell. Med.* **86**, 20–32 (2018).

67. Bhareti, K. *et al.* A Literature Review of Recommendation Systems. in *2020 IEEE International Conference for Innovation in Technology (INOCON)* 1–7 (2020). doi:10.1109/INOCON50539.2020.9298450

68. Gavgani, V. Z. Health Information Need and Seeking Behavior of Patients in Developing Countries' Context; an Iranian Experience. in *Proceedings of the 1st ACM International Health Informatics Symposium* 575–579 (Association for Computing Machinery, 2010). doi:10.1145/1882992.1883086

69. Fox, S. & Duggan, M. Health Online 2013. (2015). Available at: https://www.pewresearch.org/internet/2013/01/15/health-online-2013/. (Accessed: 25th March 2022)

70. Langford, A. T., Roberts, T., Gupta, J., Orellana, K. T. & Loeb, S. Impact of the Internet on Patient-Physician Communication. *Eur. Urol. Focus* **6**, 440–444 (2020).

71. Ross, M. K., Wei, W. & Ohno-Machado, L. 'Big data' and the electronic health record. *Yearb. Med. Inform.* **9**, 97–104 (2014).

72. Quinn, M. *et al.* Electronic health records, communication, and data sharing: challenges and opportunities for improving the diagnostic process. *Diagnosis (Berlin, Ger.* **6**, 241–248 (2019).

73. Mazur, L. M., Mosaly, P. R., Moore, C. & Marks, L. Association of the Usability of Electronic Health Records With Cognitive Workload and Performance Levels Among Physicians. *JAMA Netw. open* **2**, e191709 (2019).

74. Rehman, F. *et al.* Diet-right: A smart food recommendation system. *KSII Trans. Internet Inf. Syst.* **11**, 2910–2925 (2017).

75. Bankhele, S., Mhaske, A., Bhat, S. & V., S. A Diabetic Healthcare Recommendation System. *Int. J. Comput. Appl.* **167**, 14–18 (2017).

76. Donciu, M., Ionita, M., Dascalu, M. & Trausan-Matu, S. The Runner -- Recommender System of Workout and Nutrition for Runners. in *2011 13th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing* 230–238 (2011). doi:10.1109/SYNASC.2011.18

77. Narducci, F. *et al.* A Recommender System for Connecting Patients to the Right Doctors in the HealthNet Social Network. in *Proceedings of the 24th International Conference on World Wide Web* 81–82 (Association for Computing Machinery, 2015).

doi:10.1145/2740908.2742748

78. Han, Q., Ji, M., Troya, I. M. de R. de, Gaur, M. & Zejnilovic, L. A Hybrid Recommender System for Patient-Doctor Matchmaking in Primary Care. in *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)* 481–490 (2018). doi:10.1109/DSAA.2018.00062

79. Hussein, A. S., Omar, W. M., Li, X. & Ati, M. Efficient Chronic Disease Diagnosis prediction and recommendation system. in *2012 IEEE-EMBS Conference on Biomedical Engineering and Sciences* 209–214 (2012). doi:10.1109/IECBES.2012.6498117

80. Nasiri, M., Minaei, B. & Kiani, A. Dynamic Recommendation: Disease Prediction and Prevention Using Recommender System. *Int J Basic Sci Med* **1**, 13–17 (2016).

81. Davis, D. A., Chawla, N. V, Christakis, N. A. & Barabási, A.-L. Time to CARE: a collaborative engine for practical disease prediction. *Data Min. Knowl. Discov.* **20**, 388–415 (2010).

82. Chen, R.-C., Jiang, H. Q., Huang, C.-Y. & Bau, C.-T. Clinical Decision Support System for Diabetes Based on Ontology Reasoning and TOPSIS Analysis. *J. Healthc. Eng.* **2017**, 4307508 (2017).

83. Mustaqeem, A., Anwar, S. M., Khan, A. R. & Majid, M. A statistical analysis based recommender model for heart disease patients. *Int. J. Med. Inform.* **108**, 134–145 (2017).

84. Park, J.-H., Baek, J.-H., Sym, S. J., Lee, K. Y. & Lee, Y. A data-driven approach to a chemotherapy recommendation model based on deep learning for patients with colorectal cancer in Korea. *BMC Med. Inform. Decis. Mak.* **20**, 241 (2020).

85. Choi, G. H. *et al.* Development of machine learning-based clinical decision support system for hepatocellular carcinoma. *Sci. Rep.* **10**, 14855 (2020).

86. Wang, J. X., Sullivan, D. K., Wells, A. C. & Chen, J. H. ClinicNet: machine learning for personalized clinical order set recommendations. *JAMIA open* **3**, 216–224 (2020).

87. Chen, J. H., Podchiyska, T. & Altman, R. B. OrderRex: clinical order decision support and outcome predictions by data-mining electronic medical records. *J. Am. Med. Inform. Assoc.* **23**, 339–348 (2016).

88. Corny, J. *et al.* A machine learning-based clinical decision support system to identify prescriptions with a high risk of medication error. *J. Am. Med. Inform. Assoc.* **27**, 1688–1694 (2020).

89. Chen, Y.-T., Chiu, Y.-C., Teng, M.-L. & Liao, P.-H. The effect of medical material management system app on nursing workload and stress. *BMC Nurs.* **21**, 19 (2022).

90. Prasad, N., Cheng, L.-F., Chivers, C., Draugelis, M. & Engelhardt, B. E. A Reinforcement Learning Approach to Weaning of Mechanical Ventilation in Intensive Care Units. *CoRR* **abs/1704.0**, (2017).

91. National Center for Health Statistics. International Classification of Diseases,Ninth Revision (ICD-9). (2022). Available at: https://www.cdc.gov/nchs/icd/icd9.htm. (Accessed: 10th April 2022)

92. National Confidential Enquiry into Perioperative Deaths. The NCEPOD Classification of Intervention. *Ncepod* 1–3 (2004). Available at: www.ncepod.org.uk/pdf/NCEPODClassification.pdf? (Accessed: 10th April 2022)

93. Ordem dos Médicos. Código de Nomenclatura. Available at: https://ordemdosmedicos.pt/codigo-de-nomenclatura/. (Accessed: 10th April 2022)

94. Weerts, H. J. P., Mueller, A. C. & Vanschoren, J. Importance of Tuning Hyperparameters of Machine Learning Algorithms. *CoRR* **abs/2007.0**, (2020).

95. Jaccard, P. THE DISTRIBUTION OF THE FLORA IN THE ALPINE ZONE.1. *New Phytol.* **11**, 37–50 (1912).

96. Choi, S. & Cha, S. A survey of Binary similarity and distance measures. *J. Syst. Cybern. Informatics* 43–48 (2010).

97. Fkih, F. Similarity measures for Collaborative Filtering-based Recommender Systems: Review and experimental comparison. *J. King Saud Univ. - Comput. Inf. Sci.* (2021). doi:https://doi.org/10.1016/j.jksuci.2021.09.014

98. Breese, J. S., Heckerman, D. & Kadie, C. Empirical Analysis of Predictive Algorithms for Collaborative Filtering. in *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence* 43–52 (Morgan Kaufmann Publishers Inc., 1998).

99. Berrar, D. Cross-Validation. in (eds. Ranganathan, S., Gribskov, M., Nakai, K. & Schönbach, C. B. T.-E. of B. and C. B.) 542–545 (Academic Press, 2019). doi:https://doi.org/10.1016/B978-0-12-809633-8.20349-X

100. Refaeilzadeh, P., Tang, L. & Liu, H. Cross-Validation BT - Encyclopedia of Database Systems. in (eds. LIU, L. & ÖZSU, M. T.) 532–538 (Springer US, 2009). doi:10.1007/978-0-387-39940-9_565

101. Shani, G. & Gunawardana, A. Evaluating Recommendation Systems BT - Recommender Systems Handbook. in (eds. Ricci, F., Rokach, L., Shapira, B. & Kantor, P. B.) 257–297 (Springer US, 2011). doi:10.1007/978-0-387-85820-3_8

102. Aggarwal, C. C. Evaluating Recommender Systems BT - Recommender Systems: The Textbook. in (ed. Aggarwal, C. C.) 225–254 (Springer International Publishing, 2016). doi:10.1007/978-3-319-29659-3_7

103. Lipton, Z. C., Elkan, C. & Naryanaswamy, B. Optimal Thresholding of Classifiers to Maximize F1 Measure BT - Machine Learning and Knowledge Discovery in Databases. in (eds. Calders, T., Esposito, F., Hüllermeier, E. & Meo, R.) 225–239 (Springer Berlin Heidelberg, 2014).

104. Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. Lett.* **27**, 861–874 (2006).

105. Saito, T. & Rehmsmeier, M. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLoS One* **10**, e0118432 (2015).

106. Hanley, J. A. & McNeil, B. J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143**, 29–36 (1982).

107. Bamber, D. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *J. Math. Psychol.* **12**, 387–415 (1975).

108. Cramer, H. *et al.* The effects of transparency on trust in and acceptance of a content-based art recommender. *User Model. User-adapt. Interact.* **18**, 455–496 (2008).

109. Lyerly, S. B. The average spearman rank correlation coefficient. *Psychometrika* **17**, 421–428 (1952).

110. Artusi, R., Verderio, P. & Marubini, E. Bravais-Pearson and Spearman Correlation Coefficients: Meaning, Test of Hypothesis and Confidence Interval. *Int. J. Biol. Markers* **17**, 148–151 (2002).

111. Schober, P., Boer, C. & Schwarte, L. A. Correlation Coefficients: Appropriate Use and Interpretation. *Anesth. Analg.* **126**, (2018).

112. Kruskal, W. H. & Wallis, W. A. Use of Ranks in One-Criterion Variance Analysis. *J. Am. Stat. Assoc.* **47**, 583–621 (1952).

113. Mann, H. B. & Whitney, D. R. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *Ann. Math. Stat.* **18**, 50–60 (1947).

114. Nachar, N. The Mann-Whitney U: A Test for Assessing Whether Two Independent Samples Come from the Same Distribution. *Tutor. Quant. Methods Psychol.* **4**, 13–20 (2008).

115. Abdi, H. The Bonferonni and Šidák Corrections for Multiple Comparisons. *Encycl. Meas. Stat.* **3**, (2007).

116. HealthIT.gov. Non-federal Acute Care Hospital Electronic Health Record Adoption. (2017). Available at: https://www.healthit.gov/data/quickstats/non-federal-acute-care-hospital-electronic-health-record-adoption.

117. Giordano, C. *et al.* Accessing Artificial Intelligence for Clinical Decision-Making. *Front. Digit. Heal.* **3**, 645232 (2021).

118. Johnson, K. B. *et al.* Precision Medicine, AI, and the Future of Personalized Health Care. *Clin. Transl. Sci.* **14**, 86–93 (2021).

119. Wright, A. & Bates, D. W. Distribution of Problems, Medications and Lab Results in Electronic Health Records: The Pareto Principle at Work. *Appl. Clin. Inform.* **1**, 32–37 (2010).

120. Kohavi, R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. in *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2* 1137–1143 (Morgan Kaufmann Publishers Inc., 1995).

121. Sarwar, B. M., Karypis, G., Konstan, J. & Riedl, J. Recommender systems for large-scale e-commerce: Scalable neighborhood formation using clustering. in *Proceedings of the fifth international conference on computer and information technology* **1**, 291–324 (2002).

122. Shichang, Z. Research on Recommendation Algorithm Based on Collaborative Filtering. in *2021 2nd International Conference on Artificial Intelligence and Information Systems* (Association for Computing Machinery, 2021). doi:10.1145/3469213.3470399

123. Berrar, D. & Flach, P. Caveats and pitfalls of ROC analysis in clinical microarray research (and how to avoid them). *Brief. Bioinform.* **13**, 83–97 (2012).

124. Wang, W. & Lu, Y. Analysis of the Mean Absolute Error (MAE) and the Root Mean Square Error (RMSE) in Assessing Rounding Model. *IOP Conf. Ser. Mater. Sci. Eng.* **324**, 12049 (2018).

125. Shi, L. Trading-off among Accuracy, Similarity, Diversity, and Long-Tail: A Graph-Based Recommendation Approach. in *Proceedings of the 7th ACM Conference on Recommender Systems* 57–64 (Association for Computing Machinery, 2013). doi:10.1145/2507157.2507165

126. Ge, M., Delgado-Battenfeld, C. & Jannach, D. Beyond Accuracy: Evaluating Recommender Systems by Coverage and Serendipity. in *Proceedings of the Fourth ACM Conference on Recommender Systems* 257–260 (Association for Computing Machinery, 2010). doi:10.1145/1864708.1864761

127. Fouss, F. & Saerens, M. Evaluating Performance of Recommender Systems: An

Experimental Comparison. in *2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology* **1**, 735–738 (2008).

128. Pang, J., Guo, J. & Zhang, W. Using Multi-objective Optimization to Solve the Long Tail Problem in Recommender System BT - Advances in Knowledge Discovery and Data Mining. in (eds. Yang, Q., Zhou, Z.-H., Gong, Z., Zhang, M.-L. & Huang, S.-J.) 302–313 (Springer International Publishing, 2019).

129. Parasuraman, R. & Miller, C. A. Trust and Etiquette in High-Criticality Automated Systems. *Commun. ACM* **47**, 51–55 (2004).

130. McNee, S. M., Lam, S. K., Konstan, J. A. & Riedl, J. Interfaces for Eliciting New User Preferences in Recommender Systems BT - User Modeling 2003. in (eds. Brusilovsky, P., Corbett, A. & de Rosis, F.) 178–187 (Springer Berlin Heidelberg, 2003).

131. Lee, J. D. & See, K. A. Trust in Automation: Designing for Appropriate Reliance. *Hum. Factors* **46**, 50–80 (2004).

Appendix A

Table 1: *Variables selected from each database.*

Diagnostics' Database		
Variable	**Description**	**Type**
NHC	Encrypted user ID	Code
Número Episódio	Encrypted event (hospital admission) code in the hospital system	Code
Número Proposta	Encrypted surgery in the hospital system	Code
ID Cirurgia	Encrypted surgical procedure code, in the hospital system, associated to a surgery	Code
Tipo Diagnóstico	Type of diagnostic: • 'Principal' – main diagnostic • 'Associado' – secondary diagnostic related to the main diagnostic	Categorical
Diagnóstico	Diagnostic code (ICD-9) [A]	Categorical
Surgical Procedures' Database		
Variable	**Description**	**Type**
NHC	Encrypted user ID in the hospital system	Code
Número Episódio	Encrypted event (hospital admission) code in the hospital system	Code
Número Proposta	Unique surgery code in the hospital system (one event can present >1 surgery code)	Code
ID Cirurgia	Surgical procedure code, in the hospital system, associated to a surgery (each surgery can present >1 surgical procedure)	Code
Grau de Prioridade	Urgency degree of the patient's intervention [92]: • Elective intervention - planned or booked surgery. • Expeditive intervention – stable patient who needs an early surgery which is not life threatening. • Urgent intervention – surgery for acute medical conditions or clinical deterioration that may be life-threatening.	Categorical
Especialidade	Medical specialty of the surgery	Categorical
Duração Prevista Internamento	Prediction of the days of hospitalization after the surgery	Categorical
Tipo de procedimento	Type of surgical procedure:	Categorical

	• 'Principal' – main surgical procedure • 'Associado' – secondary surgical procedure related to the main surgical procedure	
OM	Code of the surgical procedure from *Código de Nomenclatura de Actos Médicos* [93]	Categorical
ICD	Code of the surgical procedure from ICD-9 [91]	Categorical
Medical Items' Database		
Variable	**Description**	**Type**
NHC	Encrypted user ID in the hospital system	Categorical
Número Episódio	Encrypted event (hospital admission) code in the hospital system	Categorical
Data do Episódio	Encrypted date and time of the event	Date
Data do Consumo	Encrypted date and time of the medical item records associated to a surgery	Date
Código do Artigo	Encrypted medical item ID that was consumed	Categorical
Nome do Artigo	Name of the medical item that was consumed	Categorical
Quantidade	Quantity of a medical item that was consumed	Numerical

QUESTIONNAIRE

Este questionário destina-se a avaliar a sua confiança nas recomendações sugeridas por um algoritmo, no registo de consumos de artigos médicos e respetivas quantidades, utilizadas no primeiro dia de internamento de um doente sujeito a uma cirurgia eletiva.

1. **De 0 a 10, como avalia a capacidade de o sistema de recomendação sugerir os artigos médicos corretos?**

2. **De 0 a 10, como avalia a capacidade de o sistema de recomendação sugerir as quantidades corretas dos artigos médicos?**

3. **Das 15 cirurgias, para quantas registaria as recomendações (conjunto de itens e respetivas quantidades) sugeridas pelo sistema?**
   (Ex: Registaria as sugestões feitas pelo sistema em 7 das 16 cirurgias → 7/16.)

   ___ / 15

4. **Depois de conhecer as recomendações feitas pelo sistema, utilizaria o mesmo para o ajudar no processo de registo de artigos médicos?**

5. **O seguinte texto explica de forma breve como é que o sistema de recomendação faz as previsões dos artigos médicos e respetivas quantidades:**
   O sistema de recomendação utiliza dois tipos de informação recolhidos em doentes prévios:
   a. Perfil do doente – inclui os seus diagnósticos, os procedimentos cirúrgicos a que foi sujeito e o tempo previsto de internamento.

b. Informação sobre os consumos - inclui os artigos médicos e respetivas quantidades que foram utilizados no primeiro dia dos doentes que já foram sujeitos a uma cirurgia eletiva.

As previsões dos artigos médicos, necessários para o primeiro dia de internamento de um novo doente, são feitas através de dois passos principais:

1) Comparação do perfil do novo doente com o perfil de doentes cuja informação dos consumos já é conhecida, para ser determinado o nível de semelhança entre eles;

2) Cálculo das quantidades necessárias dos artigos médicos, baseado nos consumos feitos pelos doentes conhecidos, cujo perfil apresenta semelhanças ao perfil do novo doente, e respetiva recomendação.

Após conhecer o funcionamento do sistema de recomendação, utilizaria o mesmo para o ajudar no registo de artigos médicos?

○ ○
Sim Não

6. **Se no momento da sugestão dos artigos médicos e respetivas quantidades o sistema explicasse o porquê das suas recomendações, isso ajudaria à sua decisão de o utilizar?**

○ ○
Sim Não

7. **As quantidades dos artigos médicos sugeridas pelo sistema apresentam taxas de erro relativamente aos consumos reais. Além disso, o sistema tende a ter mais dificuldade em sugerir artigos menos frequentemente utilizados. Tendo isto em conta, utilizaria esta ferramenta como apoio ao registo dos consumos standard (mais frequentemente utilizados) no primeiro dia de internamento após uma intervenção eletiva?**

○ ○
Sim Não

Appendix B

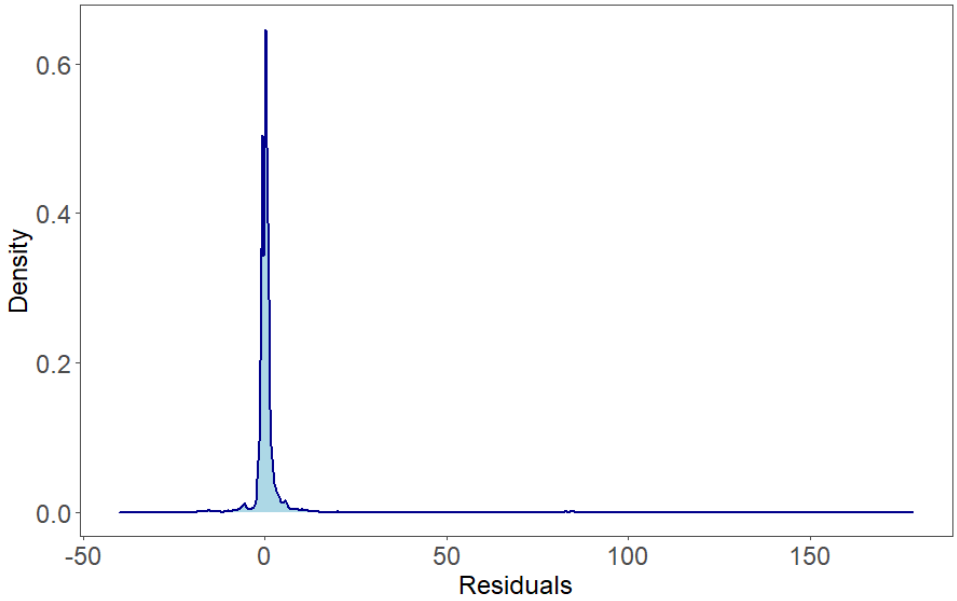


Figure 1: *Distribution of the predictions' residuals.*

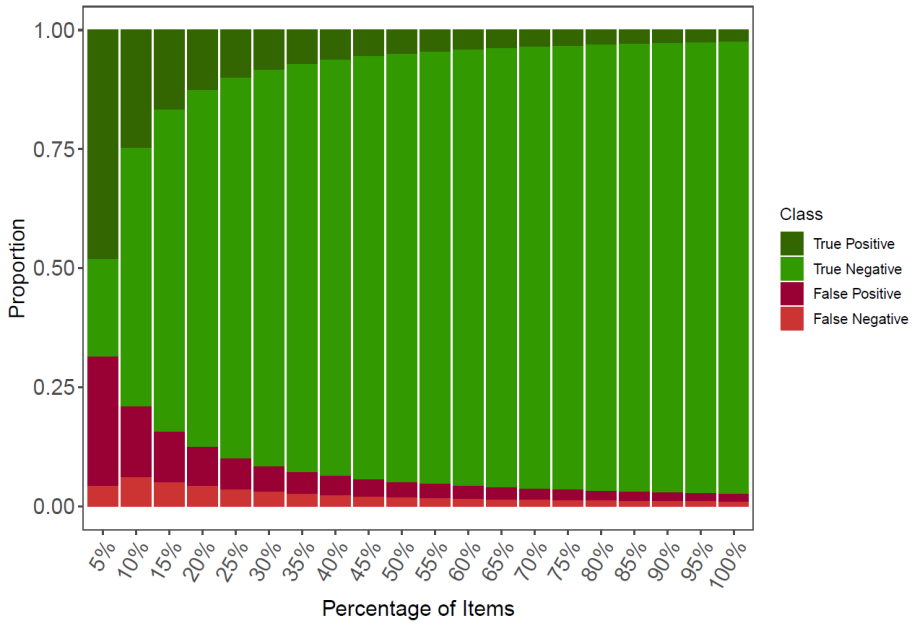


Figure 2: *Proportion of TP, TN, FP, and FN of the algorithm recommendations with the cumulative addition of medical items (ordered by their frequency of records). In each step, 5% of the 602 distinct medical items that could be recommended are added to the algorithm (x-axis).*

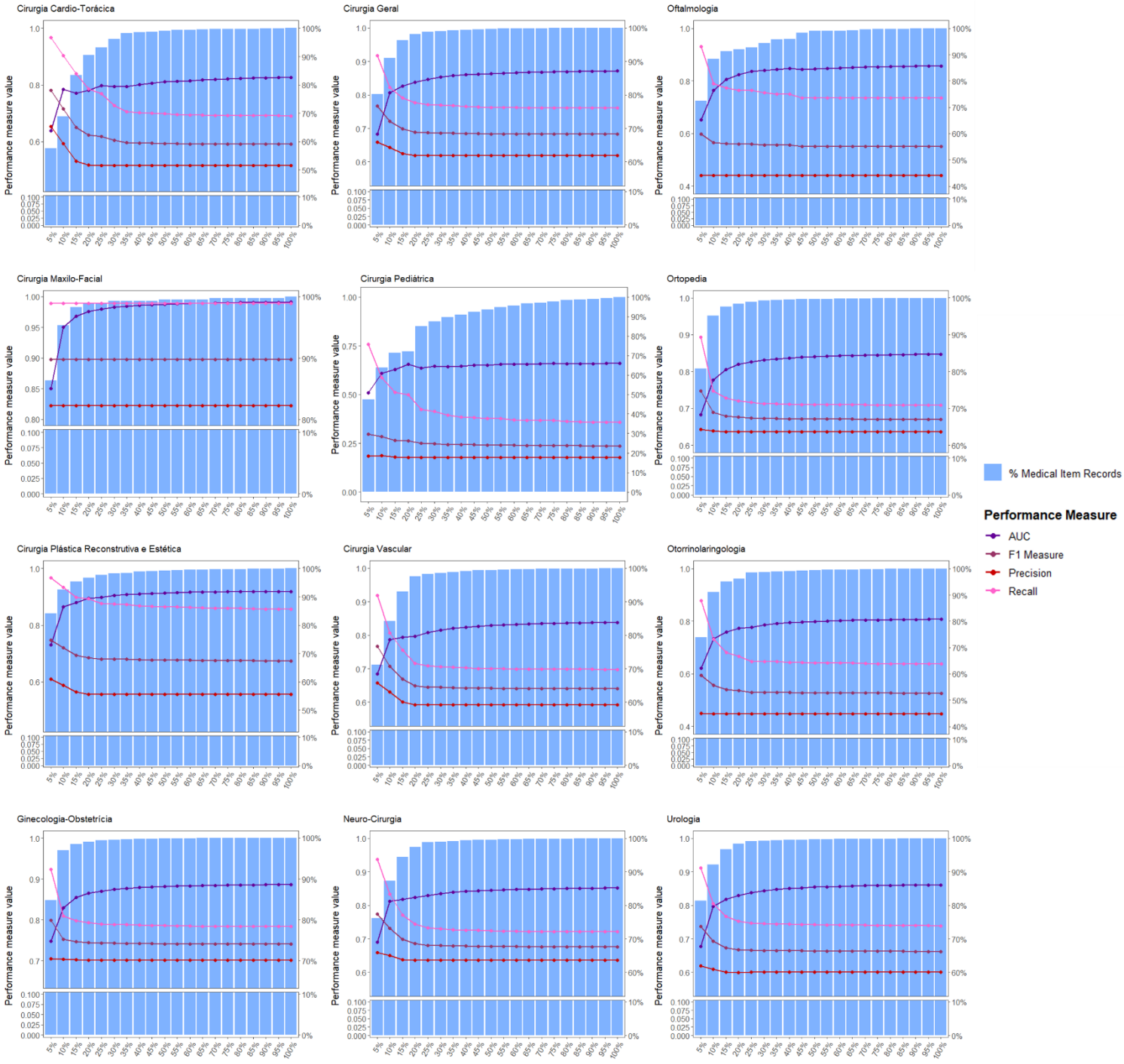


Figure 3: *Variation of four performance measures (AUC, F1 measure, precision, and recall) with the cumulative addition of medical items (ordered by their frequency of records) in the recommendation algorithm, per medical specialty. In each step, 5% of the 602 distinct medical items that could be recommended are added to the algorithm (x-axis). The percentage of medical item records in medical specialty of the test set, covered by the items added in each step (right y-axis), are represented by the bars in the background. Note that the test set did not use all medical items, wherefore the medical records and consumptions only refer to items that were used in a given medical specialty of the test set. AUC - Area under the receiving operating characteristic curve.*

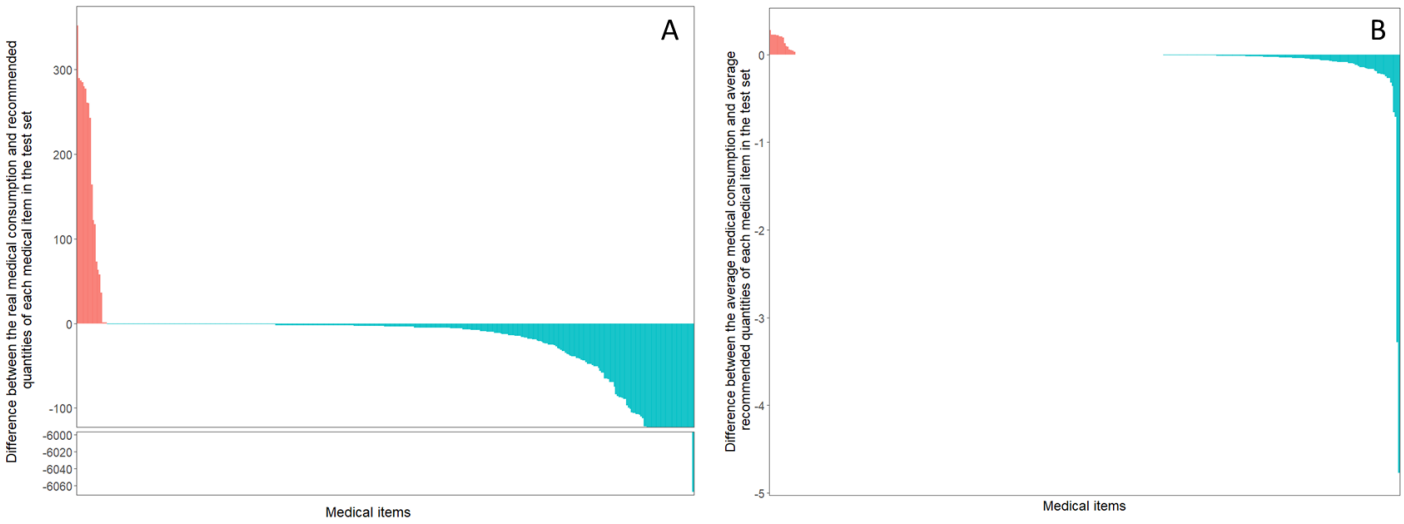


Figure 4: *Comparison between the real medical consumptions and the quantities recommended by the algorithm in the test set.* ***A)*** *Difference bewteen the real medical consumptions of each medical item and the respective recommended quantities;* ***B)*** *Difference bewteen the average medical consumption of each medical item and the respective average quantity recommended by the algorithm.*

Table 1: *Results of the recommendations of the 20 most frequently registered items in 'Cirurgia Pediátrica' performed by the algorithm.*

Medical Item	Precision	Recall	F1 Measure
Paracetamol 10 mg/ml Sol inj Fr 100 ml IV	Not recommended	0.000	0.000
Agulha Diluição 19gx1 1/2 1,10x40	0.636	1.000	0.778
Cloreto de Sódio 0,9% IV Amp 10 ml	0.727	1.000	0.842
Seringa 5ml (2 peças)	0.455	1.000	0.625
Compressa N/Esteril Tnt 7,5x7,5cm 30gr	0.455	1.000	0.625
Seringa 10ml (2 peças)	0.545	1.000	0.706
Cobertura Descart. Termómetro	0.455	1.000	0.625
Seringa 20ml	1.000	0.333	0.500
Seringa 2ml (2 peças)	0.000	0.000	-
Luva Nitrilo M	0.455	1.000	0.625
Cetorolac 10 mg/1 ml Sol inj Fr 1 ml IM IV	Not recommended	0.000	-
Prolongamento Venoso Pediátrico 150cm c/ Luer Lock	Not recommended	0.000	-
Alcool 70 º 1000 ml	Not recommended	0.000	-
Sacarose, Solução 24% 1,5ml unidose	Not recommended	0.000	-
DIazepam 5 mg Comp	Not recommended	0.000	-
Oxibutinina 5 mg Comp	Not recommended	0.000	-
Metamizol magnésico 2000 mg/5 ml Sol inj Fr 5 ml IM IV	Not recommended	0.000	-
Midazolam 15 mg/3 ml Sol inj Fr 3 ml IM IV	Not recommended	0.000	-
Sistema Infusora	0.500	0.250	0.333
Água para preparações injetáveis Sol inj Fr 10 ml	Not recommended	0.000	-

Table 2: *Multiple comparisons (bonferroni's corrections) between the precision of the predictions regarding the 20 most frequently registered medical items in each medical specialty. Green cell – $p < 0.05$.*

Table 3: *Multiple comparisons (bonferroni's corrections) between the F1 measure of the predictions regarding the 20 most frequently registered medical items in each medical specialty. Green cell – $p < 0.05$.*

Pairwise Wilcoxon Test – Precision

	Cirurgia Cardio-Torácica	Cirurgia Geral	Cirurgia Maxilo-Facial	Cirurgia Plástica	Cirurgia Vascular	Ginecologia-Obstetrícia	Neuro-Cirurgia	Oftalmologia	Ortopedia	Otorrinolaringologia
Cirurgia Geral	1.000	-	-	-	-	-	-	-	-	-
Cirurgia Maxilo-Facial	0.003	0.003	-	-	-	-	-	-	-	-
Cirurgia Plástica Reconstrutiva e Estética	1.000	1.000	0.003	-	-	-	-	-	-	-
Cirurgia Vascular	1.000	1.000	0.003	1.000	-	-	-	-	-	-
Ginecologia-Obstetrícia	1.000	0.760	0.005	0.036	0,702	-	-	-	-	-
Neuro-Cirurgia	1.000	1.000	0.003	0.519	1,000	0,597	-	-	-	-
Oftalmologia	0.001	0.002	0.001	0.007	0.005	0.000	0.005	-	-	-
Ortopedia	1.000	1.000	0.008	1.000	1.000	0.151	1.000	0.001	-	-
Otorrinolaringologia	0.001	0.002	0.001	0.027	0.005	0.000	0.002	1.000	0.003	-
Urologia	1.000	1.000	0.003	1.000	1.000	0.028	0.578	0.009	1.000	0.008

Pairwise Wilcoxon Test – F1 Measure

	Cirurgia Cardio-Torácica	Cirurgia Geral	Cirurgia Maxilo-Facial	Cirurgia Plástica	Cirurgia Vascular	Ginecologia-Obstetrícia	Neuro-Cirurgia	Oftalmologia	Ortopedia	Otorrinolaringologia
Cirurgia Geral	1.000	-	-	-	-	-	-	-	-	-
Cirurgia Maxilo-Facial	0.003	0.003	-	-	-	-	-	-	-	-
Cirurgia Plástica Reconstrutiva e Estética	1.000	1.000	0.003	-	-	-	-	-	-	-
Cirurgia Vascular	1.000	1.000	0.003	1.000	-	-	-	-	-	-
Ginecologia-Obstetrícia	1.000	0.760	0.005	0.036	0.702	-	-	-	-	-
Neuro-Cirurgia	1.000	1.000	0.003	0.519	1.000	0.597	-	-	-	-
Oftalmologia	0.001	0.002	0.001	0.007	0.005	0.000	0.005	-	-	-
Ortopedia	1.000	1.000	0.008	1.000	1.000	0.151	1.000	0.001	-	-
Otorrinolaringologia	0.001	0.002	0.001	0.027	0.005	0.000	0.002	1.000	0.003	-
Urologia	1.000	1.000	0.003	1.000	1.000	0.028	0.578	0.009	1.000	0.008

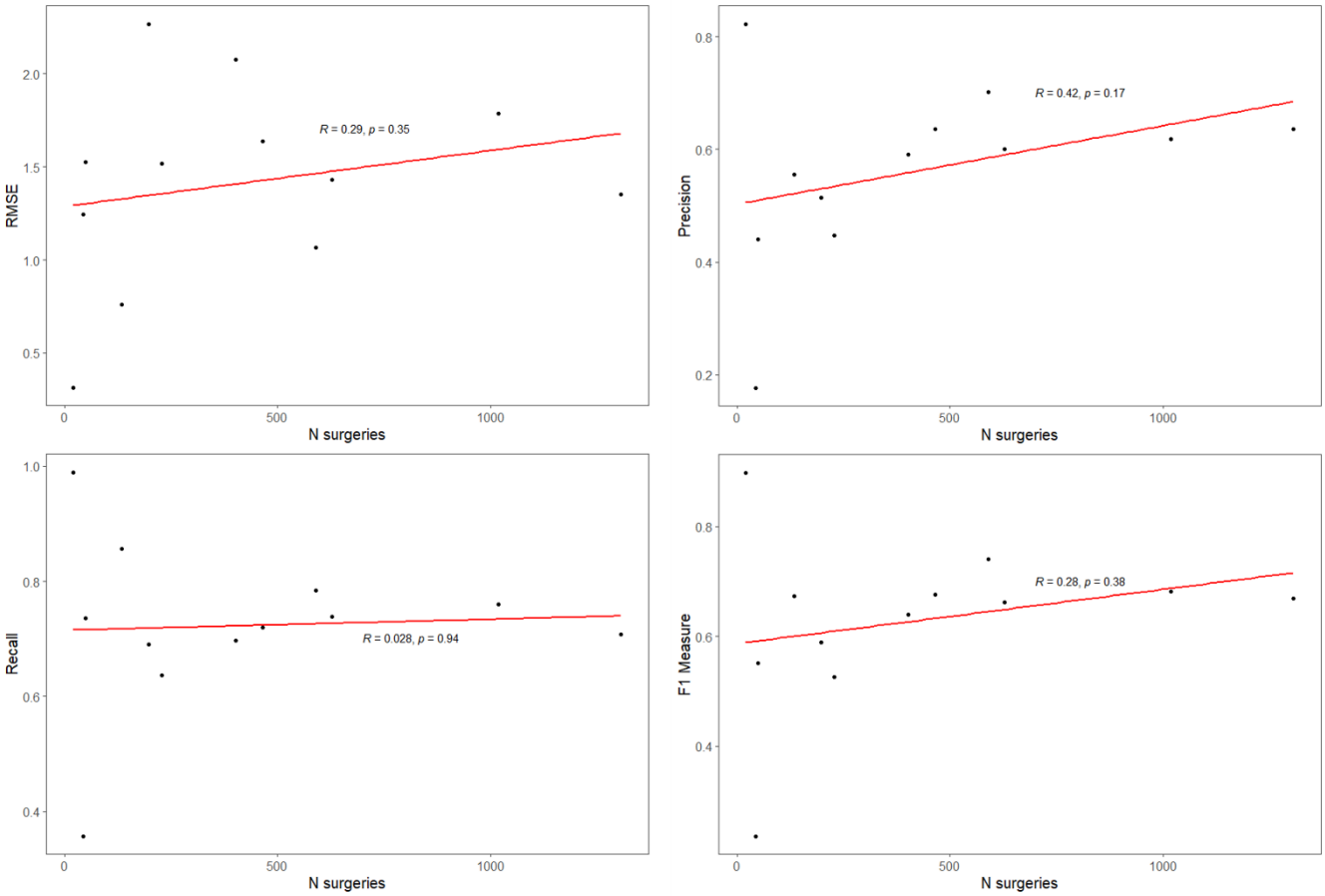


Figure 5: *Correlations between four performance measures (RMSE, precision, recall and F1 measure) and the sample size (number of surgeries) of each medical specialty.*