UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN

BAYESIAN NETWORKS FOR EFFICIENT RECOMMENDATION OF DIAGNOSTIC EXAMS ON EMERGENCY ROOM

TESIS PARA OPTAR AL GRADO DE MAGISTER EN CIENCIAS, MENCIÓN COMPUTACIÓN
MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL EN COMPUTACIÓN

JORGE ANDRÉS QUINTEROS SCHOLZ

PROFESOR GUÍA:
NELSON BALOIAN TATARYAN

MIEMBROS DE LA COMISIÓN:
CLAUDIO GUTIÉRREZ GALLARDO
GONZALO NAVARRO BADINO
ELIANA SCHEIHING GARCIA

SANTIAGO DE CHILE
2019

ESTE TRABAJO HA SIDO FINANCIADO EN PARTE POR LA EMPRESA JAPONESA ALLM INC.

Resumen

El sistema de salud público japonés incluye seguros médicos obligatorios que subsidian todo tipo de procedimiento. Esto provoca que algunos profesionales en duda a veces encarguen exámenes innecesarios, especialmente en departamentos como la sala de urgencias (SU) donde hay restricciones de tiempo y personal, generando así costos adicionales al sistema. Surge así el desafío de desarrollar un modelo de inteligencia artificial tome los síntomas de un paciente entrando a la SU y recomiende los exámenes más adecuados. Esta tesis presenta el desarrollo de un modelo así que prueba dos tipos de Redes Bayesianas (RB): una red BN2O completa similar a las ampliamente usadas para el problema de diagnóstico general, y un mucho más simple *Arreglo de Árboles Bayesianos* (AAB) que trata cada par observación-enfermedad independientemente. A pesar de que hay mucha literatura sobre RBs para diagnóstico médico, este trabajo es una contribución ya que está enfocado en sugerir exámenes basándose en probabilidades pre-examen, y también porque en esencia es un sistema experto con muy poca participación directa de expertos médicos. Distintas fuentes de información y vocabulario fueron montadas junto a algunos datos médicos para construir la base de conocimiento requerida, la cual fue post-procesada y revisada manualmente con ayuda de un experto médico. Los modelos fueron probados en enfermedades de alta relevancia y una ligera evaluación de la exactitud de los exámenes recomendados fue llevada a cabo por 2 médicos diferentes (de Japón y de Chile) comparando ambos modelos, y comparando distintos tipos de pacientes. Los resultados sugieren sorpresivamente que el modelo más simple (AAB) mostró mejores resultados que el más completo modelo BN2O. Las comparaciones entre tipos de pacientes arrojaron muy pocas diferencias entre sexo y edad, probablemente debido a la falta de modelación de factores de riesgo y restricciones en los datos. Los resultados apunta en una buena dirección pero dejan la puerta abierta a una evaluación más completa así como a varias mejoras. Las conclusiones generales son positivas, pero el camino es largo antes de que un sistema como este pueda ser implementado en un escenario real.

Abstract

The Japanese public health system relies upon a mandatory insurance scheme that subsidizes every procedure. This can cause practitioners in doubt to order unnecessary exams, especially in places like the emergency room (ER) where time and personnel constraints apply, generating additional costs for the system. This opens the challenge of developing a computer model using Artificial Intelligence that, given a patient's symptoms upon entering the ER, recommends the most appropriate exams to increase the accuracy of the diagnosis. This thesis presents the development of such a tool using Bayesian Networks (BN). Two models were evaluated: a BN2O network model similar to the ones usually used for the general diagnosis problem, and a simpler *Bayesian Trees* (BT) array model that treats findings-disease relationships independently. Although there is a lot of literature on BN for medical diagnosis, this work is a contribution as it suggests useful exams based on pre-test probabilities, and also because essentially it is an expert system with very little input from medical experts. Different sources such as the Human Symptom-Disease Network (HSDN), the Unified Medical Language System (UMLS), medical data from the Tsuoyama Hospital in Japan, among others were put together to build the required knowledge base, with post-processing and manual revision of the database undertaken with help of a medical expert. Both models were tested on relevant diseases and a light evaluation of the accuracy of the recommended exams was performed by 2 different physicians (from Japan and Chile) for both models, and also for different patient types. Results surprisingly suggest that the simpler BT model performs better than the slower, but more complete BN2O model. Comparisons between patient types showed very little difference between sex and age, probably as a consequence of the limited modeling of risk factors and restrictions in the data. Results indicate a general good direction, but leave the door open for a more complete evaluation as well as several improvements. General conclusions are positive as the results are promising, but there are substantial steps to take before a system like this could be implemented in a real life scenario.

A mis padres. Gracias por todo.

Agradecimientos

Quiero partir agradeciendo inmensamente a mi profesor guía Nelson por todo su apoyo y confianza en mi desde el primer minuto y durante todo este tiempo. Agradezco también enormemente a la gente de Allm, en particular a Horacio y a Teppei, por la oportunidad que me dieron de trabajar con ellos en un proyecto tan interesante, por confiarme los datos médicos y por sus generosas invitaciones. Quiero agradecer también al Dr. Álvaro Henríquez por su gran ayuda con la interpretación de la información médica, la evaluación de resultados y por su orientación y retroalimentación práctica en general. También agradezco a la Dr. Atsuna Matsumoto por darse el tiempo de evaluar los resultados también. Agradezco a todo el resto de quienes me ayudaron con el paper y otros detalles, y finalmente agradezco mucho a mis padres, y amigos por apoyarme, presionarme y por supuesto también distraerme cuando fuera pertinente.

Jorge Quinteros Scholz

Contents

List of Figures

Chapter 1

Introduction

Artificial Intelligence for Medicine

Has been existed since computers began the idea of developing *artificially intelligent* systems to assist human labor. In the particular field of medicine, the belief that computer systems could one day support or even replace physicians on their decision making process—something that sounds novel even today—has been around since at least the early 1950s [6].

Compared to other areas where it is relatively simple to picture a program performing a task normally done by humans, like classification of documents, customer service or video games, the medical domain can be very restricted and sensitive. The overall knowledge and the complex structure of this knowledge required by a medical expert (and therefore by a computer in some way) to be able to make a confident diagnosis or medical decision was not something easily programmable for decades. Additionally, the ethical dilemma present when systems like this go into production, as well as the issues concerning access to private medical data are even subjects of books, conferences and medical literature, so it has been undoubtedly an important obstacle for many ambitious projects. To add more, it is known that the medical community has been historically quite skeptical and reluctant to such projects because of prevailing traditionalism and general distrust, among other reasons. For all of these reasons, it has been very difficult for most of these attempts to perform their desired task correctly, hence a very limited number of these projects have made it to production in actual clinical environments.

Despite the above, it is generally understood that the decision making process of a medical expert can be described as *algorithmic* as it is generally based on gathering as much information as possible and then estimating disease probabilities based on that information. Usually many iterations of these steps are required to determine the presence of a disease and be able to treat it effectively [7]. Within this scheme, an effective diagnosis takes into account convincing probability levels of one particular disease being present over another. A physician with the knowledge to determine whether or not a patient has a certain disease must take into account various components: symptoms, signs, risk factors like age or sex, and laboratory or image test results.

Clinical Decision Support Systems (CDSS) are implementations of at least some part of this process. They will be addressed in more detail in Chapter 3 but the key idea is that they serve different purposes like general diagnosis, disease risk estimation or exam recommendation. How they are built (medical data, coded knowledge, etc.) depends greatly on their purpose and the real-world setting where they are immersed.

The Japanese Emergency Department Case

The Japanese public health system requires every resident in the country to enroll in one of the two main types of insurance programs: the National Health Insurance and the Employees' Health Insurance [8]. The system is heavily subsidized by the government so both of these insurances cover a great percentage (around 70%) of almost every medical procedure including consultations, image or laboratory tests, hospitalizations and even medicines. In particular, they cover a great part of the cost of often very expensive diagnostic exams.

Not only in Japan, but almost everywhere, the emergency room (ER) of a hospital is often an overcrowded place where medical procedures are tied to strict time and personnel constraints [9, 10]. Practitioners have to make decisions towards diagnosing as fast as possible based on patient' information, or more specifically, *findings*: presence of acute symptoms,

medical history, physical examinations and laboratory/radiology tests [11]. Under pressure, and due to lack of specific knowledge and/or fear of making medical mistakes, practitioners sometimes order exams that may not be necessary to reach a precise diagnosis. In public health systems like the Japanese one, where most of the cost is refunded anyway, this translates into heavy monetary costs. This is also a recurrent problem in many countries, and it ends up decreasing the quality of health care, increasing its costs and even exposing patients to secondary effects of unnecessary exams [12].

From a medical point of view the process involved in emergency visits is pretty well understood. To reach a medical diagnosis practitioners need to explore symptoms and signs of a patient, on what is called a *clinical examination*. In most cases this procedure is enough to reach a correct diagnosis but sometimes it only results in a spectrum of possible diagnoses, something that is known as *differential diagnosis*. Diagnostic exams are then required to achieve a more specific diagnosis within all the possibilities resulting from the differential diagnosis.

It is in this scenario where an artificial intelligence application could be of great help. For example, it could suggest that practitioners perform certain diagnostic exams based on symptoms and vital signs obtained from the clinical examination. The combination of a patient's information taken on arrival to the emergency room along with pre-test probabilities calculated from this information could enable a program with the required knowledge to suggest the most useful diagnostic exams for that specific scenario. As long as these recommendations are made fast and accurately, the suggested exams would eventually help to clarify the correct diagnosis or at least reduce the number of possibilities in a differential diagnosis.

Objective of this Work

With the problem introduced, the general objective of this work: the idea is to address the exam recommendation problem for emergency departments, specifically for Japanese hospi-

tals, by developing a CDSS that is able to make exam recommendations as fast and accurately as possible. It is important that it is specifically for Japanese hospitals because medical data, and therefore most of the computer programs oriented to diagnosis, vary greatly between different countries and ethnic groups. What somehow works in Japan, could be very wrong in the U.S. or Chile and vice versa. For this work we assume that at least these genotypic and phenotypic population differences between the country's hospitals are negligible.

The approach taken in this work is common in the literature of Clinical Decision Support Systems: *Bayesian networks*. The decision to approach the problem with them was made based on their extensive use and documentation within CDSS literature, their flexibility to incorporate data from different sources and also on the personally available hospital data at the time of the conception of this project. Chapter 3 offers a more detailed argument for this. The baseline design is typical for diagnosis-support models: a set of findings which can be present or not due to the manifestation of different diseases is established and used to calculate the most probable ones via multiple instances of the Bayes' theorem.

To achieve this we can divide our goal into more specific objectives which can be stipulated as follows:

1. Gather raw and processed data for diseases and their corresponding symptoms and exams, match it and then transform it using methods from both medical and CDSS literature to obtain the probabilities required by a Bayesian network model.

2. Implement two different models to perform the pre-test diagnosis (from where the exam recommendation is going to take place): one traditional BN model with the usual mathematical assumptions, which is expected to be more precise but slower, and one simplified model that is more like an array of 'Bayesian trees', which is expected to be less precise but runs much faster.

3. Based on these pre-test probabilities, establish a diagnostic exam recommendation method that takes into account the many possible diseases, the potentially different results these tests may have (depending on the quality of the data gathered), and how the probabilities change between these different outcomes.

4. Try out and compare the efficiency of the two models, see how the recommended exams may change depending on the model and patients' personal information, and analyze the results in the context of the emergency room and our problem in general.

Hypothesis

Finally, we formally state our hypothesis for this work as follows:

We believe *it is possible to build an exam recommendation application using Bayesian networks along with simple prevalence and findings data, where it is possible to find on average 50% correctly recommended exams within the ranking's top 5*, making it a potential actual contribution to the emergency room situation.

By simple prevalence and findings data, we mean the opposite of robust medical data sets directly connecting the input with the output, which are required by Machine Learning algorithms.

Additional Remarks

It is important to mention that although there is a lot of work done using Bayesian networks for medical diagnosis, not much has been done with a focus on exam recommendation. A lot of the literature focuses on giving a final diagnosis by inputting a set of findings (symptoms *and* exam results). We believe that even though that sounds desirable from a computer science perspective, it is much more unlikely to be integrated and of actual use in a real world setting, as was implied earlier. Also, a substantial part of the projects involving Bayesian networks as CDSS have been built with help from many medical experts and through dozens of hours of professional work, thus incorporating their knowledge directly. The work presented here is a contribution as its focus is on suggesting useful exams based on pre-test probabilities, but probably also because it was built almost exclusively using medical data and other sources of freely available information, with very little direct input from medical experts.

This project originates from a Japanese public fund granted to the company Allm Inc. to research possible uses of artificial intelligence in hospital procedures and health care in general. Some of the data used comes from public hospitals in Japan and access to it was specifically allowed to Allm for this purpose.

An initial iteration of this research was published in early 2018 [13]. This thesis consists of a newer, more complete iteration of that work and is organized like this: a medical and technical background is presented with a discussion of related work. It then continues with an extensive explanation of the methodology and then medically evaluated results as much as it was possible. Finally these results are discussed and a conclusion including potential future work is presented.

Chapter 2

Background

2.1 Clinical Decision Support Systems

There is an extensive history of collaboration between Medical and Artificial Intelligence experts, dating as far as 1954 [6]. According to Wyatt et al. (1991) Clinical Decision Support Systems (CDSS) are "active knowledge systems which use two or more items of patient data to generate case-specific advice" [14]. Figure 2.1 shows a simple diagram of a CDSS like the one we aim to develop. The idea behind these kinds of systems aimed at clinical support is old, but this definition is still relevant nowadays.

Among the most notorious CDSS along the lines of this work are IBM's Watson [15], an A.I. system originally developed to participate in the TV quiz show *Jeopardy!* but later aimed at providing health-care support, as well as the Quick Medical Reference (QMR) [16] and DxPlain [17] systems, which are knowledge-wise very complete systems that have been used in real-world settings.

The aforementioned systems all fall roughly in one of the two main categories for approaching A.I. These are the Knowledge-based or Expert approach, which is the "classic" approach and the one used to build the CDSSs used as an example, and the newer and extremely popular nowadays Machine Learning approach. The latter involves computers learning by themselves what decisions to make by being fed large amounts of data.

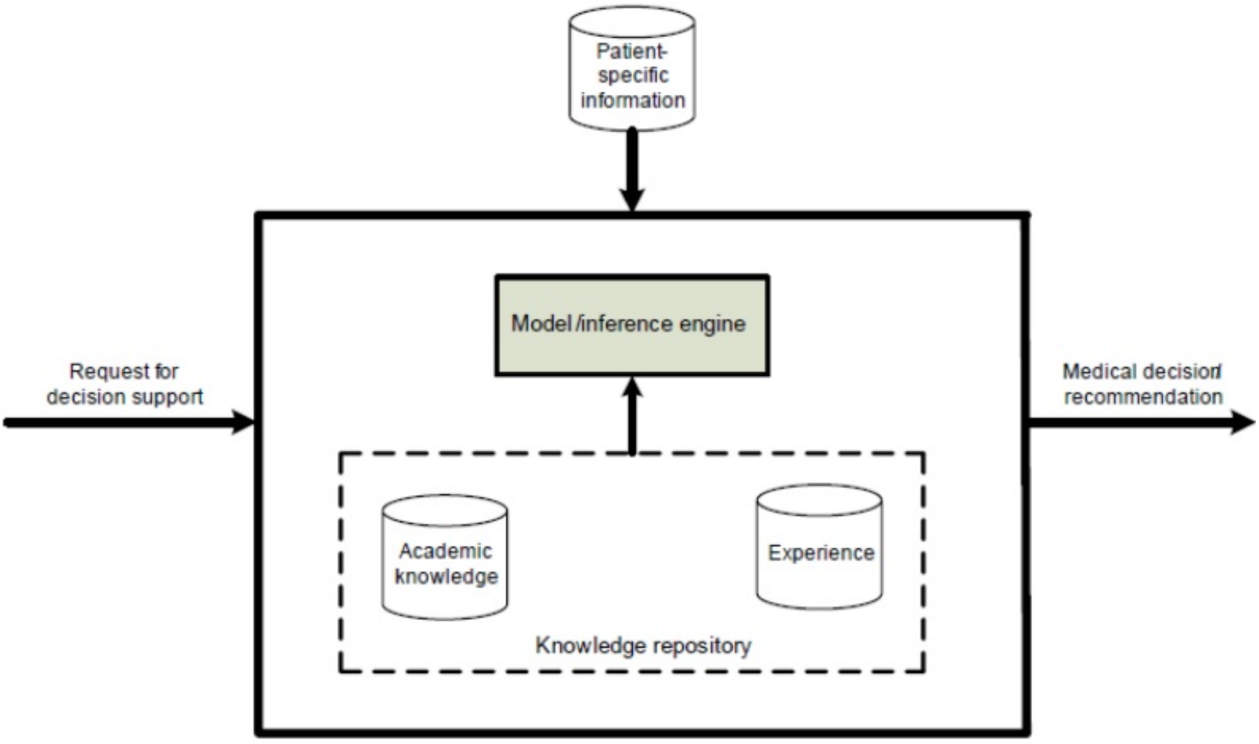


Figure 2.1: Diagram of a typical Clinical Decision Support System (CDSS). *Source: SlideShare.net. 2016-04. CDSS - Suresh Arora. Available from https://www.slideshare.net/SureshArora1/cdss-61348325.*

Knowledge-based expert systems, like the work of Sánchez et al. (1979) using fuzzy set theory [18], and many others based on sets of rules [19] and Bayesian Networks [3] have historically dominated over Machine Learning approaches, like those based on Artificial Neural Networks [20] or Decision Trees [21]. This is mainly because the latter methods became popular more recently and also because it has always been more difficult to acquire [22] and manipulate [23] large amounts of medical data (due to legal and ethical reasons, as mentioned in Chapter 1). Table 2.1 shows a large list of medical systems of different kinds that were developed only during the 90's. It can easily be observed how knowledge or rule-based systems dominate.

With the recent boom of the so-called *Big Data* field some new powerful machine learning algorithms have been developed. *Deep Learning* networks or *Random Forests* can attain good results in classification and regression problems, among others. But although these algorithms can solve problems in a variety of domains, having access to enough medical data

of reasonable quality is still a difficult task. Thus, not many revolutionary solutions using these tools have been developed yet, especially for the case of general medical diagnosis, where the set of possible outcomes is not limited to the classification of 1 or 2 diseases, but hundreds. For this reason, the subject of diagnosis-support systems is still very linked to knowledge-based methods.

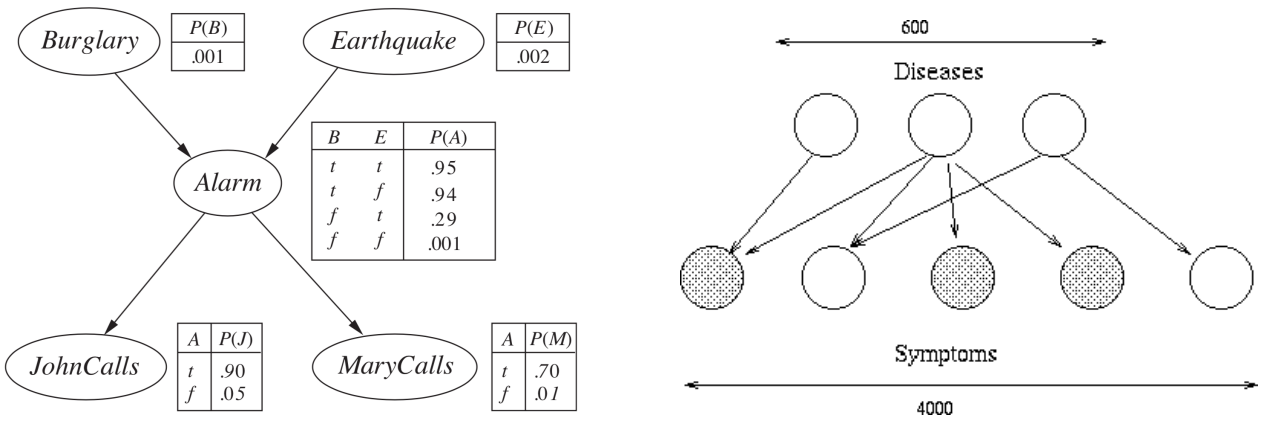
2.2 Bayesian Networks

Taking all of this into account, on the Knowledge-based side Bayesian networks appear as one of the most recurrent approaches to these problems: Quick Medical Reference's significant adaptation to Bayesian Networks, QMR-DT or QMR-BN [3], appeared only a few years after Bayesian networks were first introduced by Pearl in 1985 [25]. Other works applying very similar foundation were developed during the following decades even until recently [26]. One big challenge we face choosing this approach (and in general with expert approaches) is actually being able to set up a robust and hopefully complete knowledge base. Most of the systems based on Bayesian networks like the ones mentioned have their knowledge base built with strong help of medical experts or are adaptations of another system's knowledge (although BNs can also have their arcs and probabilities inferred from data [27]).

2.2.1 General Bayesian Networks

A Bayesian network [28] is an extensively applied probabilistic modeling tool, especially used in domains involving variable levels of uncertainty as they arose out of an attempt to add probabilities to expert systems. BNs have been used in a wide range of areas like information retrieval [29], risk analysis [30], sports betting [31], hardware and system diagnostics [32] and many more [33].

According to the definition presented by Zagorecki et al. (2013) [26] a Bayesian network is a probabilistic graphical model that represents a set of random variables (nodes) and their conditional dependencies (arcs) via a directed acyclic graph. Nodes with no incoming arcs



(a) (b)

Figure 2.2: Examples of Bayesian networks. **(a)** Commonly used example of a simple BN: an alarm can go off because of an earthquake or a burglary and cause John to call or Mary to call. Each Conditional Probability Table is shown next to its corresponding node. For parent variables only the prior probability is required, but for children variables each possible outcome regarding the parents has to be specified. *Source: "Artificial Intelligence: A Modern Approach" by Stuart Russel and Peter Norvig. Person, 3rd Edition. December 2009. Page 512, fig. 14.2. [1]* **(b)** Representation of 2-level symptom-disease network commonly used for diagnosis where presence of diseases causes the manifestation of one or more different symptoms. Numbers represent usual amount of nodes required for each category. *Source: "An introduction to graphical models", by Kevin Murphy. 1998. Available from https://www.cs.ubc.ca/ murphyk/Bayes/bnintro.html [2]*

are associated with simple probability distributions called prior probabilities, while nodes with incoming arcs or parents, have multiple conditional probabilities associated which are stored in Conditional Probability Tables (CPT). The key feature of Bayesian networks is that they represent a compact version of the Joint Probability Distributions (JPD) of all the random variables in the model. Following the independence of the variables implied from the model, one can reduce the size of CPTs by trimming the distributions required to specify each JPD. Thus, BNs allows for efficient answering of queries related to arbitrary conditional probabilities involving the variables, such as the probability of a disease given a set of findings, or the findings that maximize the probability of some disease. Figure 2.2-a shows a classical example of a simple Bayesian network representing an alarm system with

binary nodes, clearly showing the values required to specify each CPT. Figure 2.2-b shows the structure of a typical symptom-disease 2-level diagnosis network, or more specifically a *BN2O network* [34], which will be explained in detail in brief.

2.2.2 Related Medical Terminology

Some fundamental concepts used in medicine research are closely related—or can be directly applied—to statistical inference. It is partly this "algorithmic" view of medical processes along with the idea that it is possible to automate them that has motivated so much research in this area, especially within Clinical Decision Support Systems [35].

For this work, a *finding* is defined as a symptom, the outcome of a laboratory examination, analysis of an image or any other kind of diagnostic exam that can move probabilities closer or away from a certain diagnosis. Findings usually have two associated numerical values: *sensitivity*, defined as the likelihood of a person who is ill to have a positive test result, and *specificity*, defined as the likelihood of a healthy person receiving a negative test result [36, 37].

More specifically, let D_i represent disease i and F_u represent finding u. Sensitivity is the probability that finding F_u appears when disease D_i is present, i.e.:

$$S_e := P(F_u|D_i).$$

Equivalently, specificity is the probability that finding u does not appear when disease i is not present, i.e.:

$$S_p := P(\bar{F}_u|\bar{D}_i).$$

These sensitivity and specificity concepts are widely studied in the field of evidence based medicine. Ideally, they are usually statistically estimated or obtained via expert judgment, and are studied for specific finding-disease relations.

However, the numbers that are known are not quite in the form that is explicitly required for the CPTs. For example, two diseases D_1 and D_2 might cause the same finding F to appear. Usually, we can obtain the sensitivity $P(F|D)$ and specificity $P(\bar{F}|\bar{D})$ for the two

diseases; for the CPT, however, we need joint conditional probabilities like $P(F|D_1, D_2)$.

As Nikovski (2000) states [38], there are several reasons why sensitivity and specificity numbers are the easiest to obtain. They are modular and represent objective relationships between a finding and a disease regardless of which other diseases are modeled. The entries in the CPT, on the other hand, depend on all the diseases that the designer of the network has decided to model. In addition, physicians relate easily to this type of information and can provide fairly good estimates.

Fortunately, there is an efficient way to make use of these values under certain assumptions.

2.2.3 BN2O Networks

The most commonly used specific structure of Bayesian network for diagnosis problems is called BN2O [39]. This is a two-layer network with disease nodes as the top layer, finding nodes (symptoms or exam results) placed in the lower layer, and the arcs coming from the disease nodes to the finding nodes. All findings are assumed to be *noisy OR* models. We employ a summarized version of the definition presented by Nikovski (2000) to explain what a noisy OR model is [38].

Noisy OR

For n parent nodes representing diseases, there are only $2n$ sensitivities and specificities available. Determining all 2^n entries in the CPT is an ill-posed problem. The mapping from diseases to findings, however, has a specific structure, which can be exploited. It can be assumed that diseases act independently to produce a finding F, and the net effect of all diseases can be combined to produce the cumulative probability that the finding will be present. The finding can be caused by any disease, similarly to a logical OR gate, although the relationship is not deterministic. Each of the diseases D_i acting alone can cause the finding to appear with probability p_i:

$$p_i = P(F|D_i \ only) = P(F|\bar{D}_1, \bar{D}_2, ..., D_i, ..., \bar{D}_n) \tag{2.1}$$

This non-deterministic OR gate is known under the name noisy-OR and has been studied extensively [40, 41]. The probabilities p_i are also called link probabilities of the noisy-OR gate. Thus, any entry in the CPT can be obtained as:

$$P(F|H) = 1 - \prod_{D_i \in \mathcal{H}^+} [1 - p_i], \tag{2.2}$$

where H is a particular truth assignment to the parents of F, and $\mathcal{H}^+$ is the subset of those nodes in H that are set to true. If $\mathcal{H}^+$ is empty, the product term is assumed to be one and the resulting entry in the CPT is zero. This result is rarely true of any real problem domain. Even when no diseases are known to be present, there might be other unmodeled diseases that can cause the finding. In addition, there is always measurement error and certain number of false positives in the detection of the finding is inevitable.

In order to represent the effect of unmodeled diseases and measurement errors, a leak term p_L is usually employed in noisy OR models. Conceptually, all unmodeled causes for the finding are represented by a disease node L that is always present. The leak probability p_L can be used directly in 2.1 just like any other link probability; L is always in $\mathcal{H}^+$. This way, in order to specify the CPT, only $n+1$ parameters are needed (n link and one leak probabilities).

For a more detailed and possibly more intuitive explanation of the Leaky Noisy-OR, section 14.3 of the book *Artificial Intelligence: A Modern Approach* by Russel and Norvig is highly recommended [1].

Formal Definition

Now we can easily define what constitutes a BN2O network. Henrion (1991) more precisely defined a BN2O network to be one that meets the following statements [42]:

Assumption 1 (MID): Diseases are marginally independent.

Assumption 2 (CIF): All findings are conditionally independent of each other given any hypothesis.

Assumption 3 (LNOG): The effects of multiple diseases on a common finding are combined as a *Leaky Noisy OR Gate*. Suppose S_{df} is the link event that disease d is sufficient to cause finding f. The noisy OR assumption is that finding f will occur if any link event occurs linking a present disease to f, and that these link events are independent (this is sometimes known as *causal independence*). With a *leaky* noisy OR an additional leak event L_t is possible, which can cause f to occur even with no explicit disease present.

Definition 1 (BN2O): The class of bipartite Bayesian networks conforming to Assumptions 1, 2 and 3, are termed BN2O.

Almost every work that uses Bayesian networks models them under these assumptions (otherwise it becomes unmanageable as we have seen). As was mentioned in Chapter 1, most of these attempts using BN2O networks aim for general diagnosis, but virtually none aim to recommend useful exams from the ground of pre-test probabilities.

Since the problem at hand is ultimately giving useful recommendations based on overall sets of possible diseases as fast as it is necessary for an emergency department, we will work on two main things:

1. Test Bayesian models with different degrees of simplification hoping to increase speed at the cost of small accuracy decreases on giving pre-test probabilities
2. This way focus more on the elaboration of an exhaustive, good quality database of exam results, symptoms and findings in general using the limited resources available

2.3 Findings and Exam Recommendation

As was stated earlier, a finding represents any kind of medical observation or medical manifestation of one or more diseases. They can be symptoms, signs, physical exam results, or image or laboratory results. Depending on the source, some findings can be very specific, like *presenting a C-Reactive Protein level higher than 1.0 mg/dL*, and others not that much, like *Abdominal Pain*, both related to the diagnosis of *Appendicitis*.

In the usual diagnostic process, as briefly mentioned in Chapter 1, it is required first to gather information about symptoms and possibly vital signs of a patient before stepping into more complex findings like the outcome complex examinations. Therefore, if the goal is to suggest useful examinations instead of giving a diagnosis after *already* having the exam results (which seems to be the most common approach), we have to make a clear distinction between *symptom* findings and *exam results* findings.

For a particular disease, its related findings and parameters can be typically found in what is called prevalence statistics (which include sensitivity and specificity). The most famous CDSSs built their own private findings databases using dozens of medical experts [16, 17, 26]. Some of this data is partially accessible on research-oriented standardization projects like America's National Institutes of Health's *Unified Medical Language System®* (UMLS) [43]. Commercial solutions also seem to be available. However, most of these solutions either contain information that is too incomplete or does not satisfy our particular problem's constraints. Ideally, a good amount of emergency medical data would be of great help to build the database, but that was not quite available at the time of this project's conception.

Prevalence statistics usually come from research papers studying the effect of different kinds of examinations on diseases. A few websites have intended to extract these values from literature and organize it according to their respective diagnosis, with varying levels of success [44, 45]. Since these compilations of findings come from medical literature, their validity should be at least acceptable, so it comes across as a proper source for the required data.

System	Domain	Knowledge approach
Babic A. et al.	Asymptomatic liver disease	Neural networks
Doering A. et al.	Epilepsy	Neural networks
Dumitra A. et al.	Psychiatric mood disorder	Neural networks
Leao B. et al. (HYCON S II)	CHD and renal disease	Neural networks
Ohno-Machado L. et al.	AIDS	Neural networks
Witte H. et al.	Hepatobiliary disease	Neural networks
Ichimura T. et al.	Hepatobiliary disease	Neural networks Fuzzy rules
Tsumoto S. et al.	Headache and facial pain	Probabilistic rule-based reasoning
Dore L. et al. (CPR)	Hypertension	Object oriented reasoning
Ambrosino R. et al.	Cost-effective health care	Rule-based reasoning
Ryder R.M. et al.	Cardiovascular rehabilitation	Rule-based reasoning
Hogan W.R. et al.	Medicine	Bayesian networks
Haddawy P. et al. (BNG)	Cardiology	Bayesian networks
Hadzikadic M. et al.	Trauma	Standard logistic regression
Kuperman G.J. et al.	Pharmacy	Rule-based reasoning
Boon-Ralleur A.L. et al.	Hyperthyroid	Rule-based reasoning
Smart J.R. et al.	Pathology	Rule-based reasoning
Hosseini-Nezhad S.M. et al.	Pediatric	Neural networks
Spyropoulos B. et al.	Nosology	Case-based reasoning
Astion M.L. et al.	Rheumatic disease	Neural networks
Ozkarahan I.	Health care costs in operation room	Rule-based reasoning
Place J.F. et al.	Medicine	Neural networks
Duong D.V. et al.	Semiotic	IAC neural networks
Mayer-Ohly E. et al. (MEDRISK)	Insurance medical knowledge	Rule-based reasoning
Shahar Y. et al. (RESUME)	Diabetes	Knowledge-based temporal abstraction
Anderson J.D.	Clinical uncertainty	Case-based reasoning
Alendahl K.	Primary health care	Rule-based reasoning
Cl ret M. (ADM)	Cardiology, pathologies	Rule-based reasoning
Ferri F. (CADMIO)	Medicine	Rule-based reasoning
Fiore M. (MUMPS)	Medicine	Rule-based reasoning
Mann G.	Medicine	Object oriented system
Ohmann C.	Acute abdominal pain	Rule-based reasoning
Schmidt R. (ICONS)	Antibiotics therapy	Case-based reasoning
Baatar S. (RGT)	Medicine	Rule-based reasoning
Hofest R. (METABOLICA)	Metabolic disease	Rule-based reasoning
Lau F.	ICU care	Case-based reasoning
Narita Y. (3-AND-3-OR)	Medicine	Rule-based reasoning
Steimann F.	Toxoplasmasis	Rule-based reasoning
Kovacs M. (Auctoriatas)	Psychiatry	Rule-based reasoning

Table 2.1: Examples of medical expert systems of any kind (not only CDSS) developed during the 90's. The table shows each one's domain and knowledge representation. It shows that the majority correspond to knowledge-based systems, which tend to have a wider domain of application than their neural network counterparts. *Adapted from the book "The Handbook of Applied Expert Systems", by Jay Liebowitz. CRC Press, 1st Edition. December 1997. Ch. 6, Appendix. [24].*

Chapter 3

Methodology

In this section the methodology will be detailed as accurately as possible.

Almost the entire programming was done on Python 3.6 using the IDE *Pycharm 2* by JetBrains. Many Python 3 libraries on their most recent versions were used at different stages, including `Numpy`, `MySQLConnection`, `SQLAlchemy`, `Pandas`, `Pickle`, `BeautifulSoup`, `Scikit-learn` and `MeCab`. Some others, like `BayesPy` or `PGMPY` which are Bayesian Network libraries were tested but were eventually replaced by self-implementations. As for the database we used MySQL since the company has its data stored on a MySQL server, making it easier to export and import the relevant data. To access and manipulate the database remotely and locally mostly the software *SequelPro* was used. The UMLS *Metathesaurus* software was used to access medical terminology. What this is and how it can be used will be detailed in section 3.2.

We begin by describing in detail how we modeled the two different Bayesian network models implemented for this particular case. Later, we address step by step how we put together the various sources to build the database and knowledge base. Finally, we describe how the exam recommendation is made to be as optimal as possible.

3.1 Bayesian Network Models

For a program like this to be of actual use in an emergency room setting, it needs to be both accurate and fast. For this reason, two different Bayesian network models were implemented: one is a BN2O Network, described in section 2.2.3, believed to be more accurate but also slower on inference [42], and the other one is a *Bayesian Rules Array* believed to be less accurate, but much faster on inference.

Both models share the same number of disease (parent) nodes, and finding (children) nodes. They differ on the number of active arcs and the interactions between these nodes; in other words, on their inference mechanisms.

3.1.1 BN2O Network

One of the problems with Bayesian networks is that the size of their Conditional Probability Tables (CPTs) grows exponentially with the number of parent nodes, which makes them impractical for domains requiring a large number of nodes like this one. Additionally, let us say there are two diseases D_1 and D_2 related to finding F. Even though joint conditional probabilities like $P(F|D_1, D_2)$ are required to specify the finding's local table, the information that can be normally obtained from medical literature are, at most, sensitivity $P(F|D)$ and specificity $P(\bar{F}|\bar{D})$ values [38]. For this reason, actually specifying the 2^n probability values required for each node with n parents becomes too complex, so additional independence constraints and other techniques are applied.

The first model is a complete BN2O network [39] (Figure 3.1). As explained in Chapter 2, a BN2O is a 2-layer bipartite Bayesian network typically used for the general diagnosis problem. This model assumes that diseases manifest independently into findings in the form of a *Leaky Noisy-OR* [40] (see section 2.2.3). Under this model, it is the added effect of all diseases that combined can produce the cumulative probability of the finding. The finding can be caused by any disease, similarly to a logical OR gate (therefore the name).

Although these assumptions make it much simpler, as we have seen, this type of model

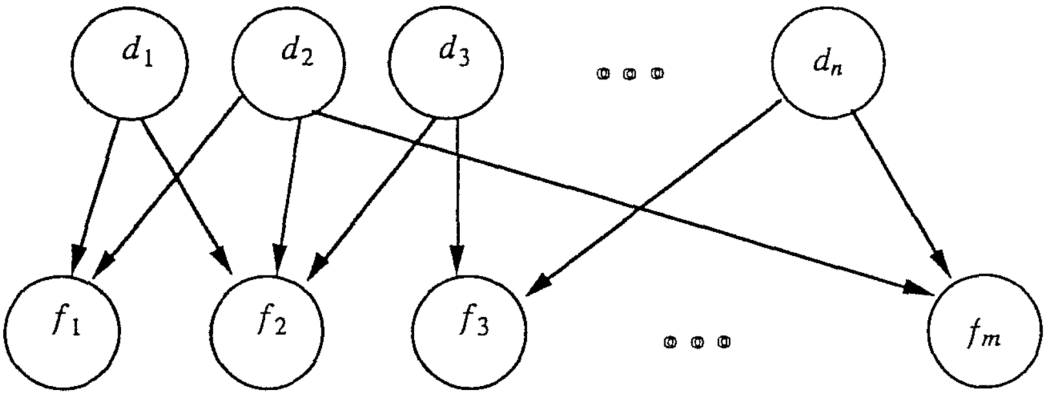


Figure 3.1: 2-level BN2O network diagram where disease nodes are labeled $d_1, ..., d_n$ finding nodes are labeled $f_1, ..., f_m$ and arrows represent conditional dependencies between nodes. *Source: Shwe et al. (1991) [3].*

still does not use findings' sensitivity and specificity values directly. Instead, CPT entries are calculated using *link probabilities* and *leak probabilities*. Nikovski (2000) proposed a way to estimate these parameters using sensitivity and specificity values [38]. Link probabilities are estimated through the equation

$$p_i = \frac{P(F|D_i) - P(F|\bar{D}_i)}{1 - P(F|\bar{D}_i)} = \frac{S_e(F, D_i) - (1 - S_p(F, D_i))}{S_p(F, D_i)}, \tag{3.1}$$

which is well-behaved if a disease D_i actually causes F, meaning findings with link probability < 0 are discarded as they are not really helpful as a diagnostic tool. Having all link probabilities p_i, a leak probability *factor* for each disease D_i can be estimated using the equation

$$p_L^i = 1 - \frac{P(\bar{F}|\bar{D}_i)P(\bar{D}_i)}{\sum_{H|D_i \in \mathcal{H}^-} \prod_{D_j \in \mathcal{H}^+} (1 - p_j) \prod_{D_j \in \mathcal{H}^-} P(\bar{D}_j)} \tag{3.2}$$

where we recall that H is a particular truth assignment to the parents of F, being $\mathcal{H}^+$ the subset of those nodes in H that are set to true and $\mathcal{H}^-$ the subset of those nodes in H that are set to false. These values are then combined by weighted averaging using the negated priors $P(\bar{D}_i) = 1 - P(D_i)$ to obtain the leak probability for a certain finding, i.e., the probability that the finding is present even when no disease is present:

$$p_L = \sum_i [1 - P(D_i)] p_L^i. \tag{3.3}$$

The details on where these formulas come from can be seen in the aforementioned paper [38].

Under this model, findings are connected to many diseases at once, but only prior probabilities, sensitivity and specificity values are required. Nevertheless, doing exact inference in a network this size still remains practically impossible, so it is common to address it using randomized sampling algorithms. A few of them were tested and will be briefly explained, but for reasons that will be made clear, in the end we decided on the *Likelihood Weighting* sampling algorithm.

Randomized Sampling Methods

Direct Sampling is the simplest sampling method. It begins by sampling each variable of the network in turn, in topological order. The probability distribution from which the value is sampled is conditioned on the values already assigned to the variable's parents, i.e., each sampling step depends only on the parent values. It is not hard to see that the samples are generated from the prior joint distribution specified by the network. To obtain an estimate of the probability for each event, the samples of that particular event are averaged over the total number of samples. The estimated probability becomes exact in the large-sample limit. Such an estimate is called *consistent*. Now, the problem with direct sampling is that it does not take into account the evidence, which is very important in this case.

Rejection sampling, on the other hand, takes into account the evidence by simply rejecting the samples that do not agree with it, and then counting how often an event occurs in the remaining samples. This produces a consistent estimate of the true conditional probability. The problem is that it rejects too many samples if the distribution of the evidence values and/or the number of evidence variables is not very small, making the estimation process too long for a complex problem like ours.

Rejection sampling was implemented and tested but it ended up taking several hours to

be able to make reasonable estimates even when using only a couple of symptoms as evidence, so a more efficient approach had to be implemented.

Likelihood weighting improves this inefficiency by generating only events that are consistent with the evidence: it fixes the values for the evidence variables and samples only the non-evidence variables. However, not all events are equal. Each event is weighted by the likelihood that the event accords to the evidence, as measured by the product of the conditional probabilities for each evidence variable, given its parents. Intuitively, events in which the actual evidence appears unlikely should be given less weight.

For any given query with some evidence the process goes as follows: First, the weight of the whole sample w is set to 1.0. For every variable in the network, if it is an evidence variable, the whole sample's weight w is multiplied by that node's probability (given its parents, or simply the prior probability if it is a parent node). If it is not an evidence variable, it simply gets sampled from its distribution. Thus, for each variable, instead of averaging over all the samples equally, each sample has its own weight, which is used to sum over all the samples and then divide it by the total number of samples. This and the other algorithms are explained in more detail in and it is shown that it gives consistent estimates [1].

This algorithm was implemented and performs much better than rejection sampling, being able to give reasonable estimates in just a few minutes. For this reason, this was selected as the default inference algorithm for the BN2O network.

Gibbs sampling is the third sampling algorithm that was implemented. This one is from the family of *Markov Chain Monte Carlo* algorithms and is much more complex than the previous two. For this reason we skip the explanation of how it works (which can be seen in [1]) and also because, upon testing, we realized that it takes an unreasonable amount of time (more than a few hours) when there are too many arcs in the network, as is the case.

3.1.2 Bayesian Trees Array

Since in most of the cases only one disease is present at a time, we attempt to make a much simpler Bayesian network model under this assumption. This model is structured like an *array of independent Bayesian trees*. As in the previous structure, the disease nodes also act independently on each finding, but we take it further by also making findings *manifest independently from each disease*. In other words, each finding connects only to one disease at a time, effectively making each disease-finding tree independent from the rest of the network. Figure 3.2 shows how findings can be 'repeated' but manifest independently for different disease nodes (e.g., findings f_1, f_2 and f_3).

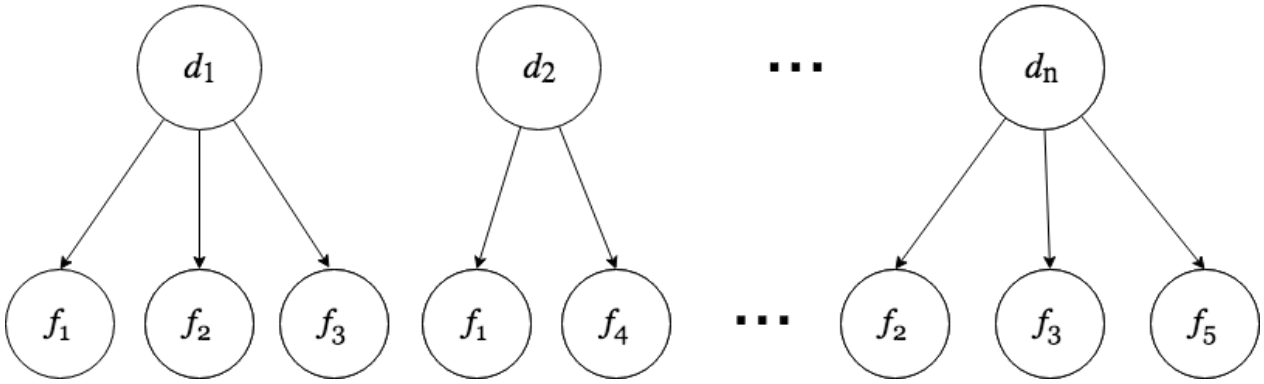


Figure 3.2: Array of independent Bayesian trees. Each finding node connects to only one disease at a time, even if they help diagnose more than one disease.

By doing this, findings' conditional probability tables only depend on one parent node, making sensitivity and specificity values exactly the required conditional probabilities. This way, we can do 'exact inference' by simply applying Bayes' theorem to every disease given the evidence:

$$P(D_i = d_i | F_1 = f_1, \ldots, F_m = f_m) = \frac{P(F_1 = f_1 | D_i = d_i) \cdots P(F_m = f_m | D_i = d_i) P(D_i = d_i)}{\sum P(F_1 = f_1 | D_i) \cdots P(F_m = f_m | D_i) P(D_i))} \tag{3.4}$$

Given that inference on each tree produces probabilities which are independent from the rest, they have to be normalized between all of them to finally output a ranking of the most probable diseases. While it is clear that this method is not as precise as the first one, it turned out to be much faster, something very important in an emergency room scenario.

3.2 Knowledge Base

After having established the architectures to be used, the next and probably most important step is to build a complete, robust and trustworthy knowledge base. As was mentioned, the raw medical data available at the time was simply not enough to perform this task (through machine learning or knowledge representation techniques) so different information sources had to be used for this. Fortunately, the attempt to organize medical information is not something new so there is semi-organized medical data *floating around*; it just needs to be organized into the structure we require.

For each type of node, we describe how the data was obtained and its probabilities estimated. We begin with the parent nodes (diseases) and then continue with the children nodes (symptoms, exam results and others).

3.2.1 Diseases

Structure

To establish the disease nodes for the network, something called the Human Symptoms-Disease Network (HSDN) [4] was used. The HSDN is a big network containing thousands of symptom-disease relationships (see Figure 3.3). These were obtained by counting lexical co-occurrence of MeSH metadata [46] within medical publications from the PubMed portal. It contains 4,219 disease terms linked to 322 symptom terms, which include the number of co-occurrences along with their corresponding TF-IDF score [47]. Both types of terms come from the official MeSH vocabulary, and therefore left untouched and used as a starting point for disease nodes.

Compared to other BNs [3], this number of diseases is too large for practical purposes so some filtering had to be done. First, diseases with too low a total number of co-occurrences (i.e., medical publications) were left out, considered to be either synonyms of diseases already present or simply extremely rare and unlikely to occur. This left us with around 3,600 disease terms.

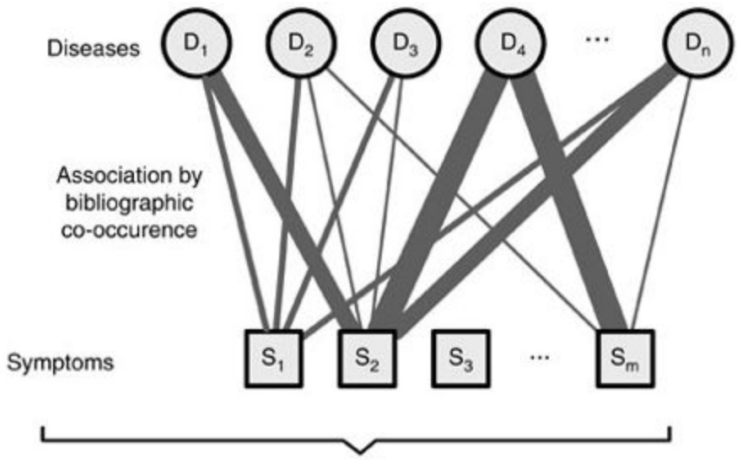


Figure 3.3: Human Symptoms-Disease Network (HSDN). The association between symptoms and diseases are based on their co-occurrence in the MeSH metadata fields of PubMed bibliographic literature database. *Source: Zhou et al. (2014) [4].*

Probabilities

As diseases are parent nodes, only their prior probability (or *prevalence* in medical terms) is required: how probable is it that the cause of an emergency visit is a specific disease?. This value is more or less proportional to the frequency of that disease being diagnosed as a result of a medical visit.

Fortunately, we had available data of more than 1 million medical visits of more than 10 different departments (of which about 190,000 correspond to emergency room) from the Tsuyama Chuo hospital in Okayama, Japan. These records included patients' information, visit dates, outcome disease name in Japanese and—most of the time—department and ICD-10 [48] codes. ICD-10 is the 10th revision of a medical classification list by the World Health Organization (WHO) that categorizes diseases, symptoms, findings and other types of medical information with a hierarchical format, and that is used widely to structure medical information.

To count these records and match them with our diseases obtained from the HSDN, we used the Unified Medical Language System (UMLS) [49], which is a very large database of medical terminology mapped from many different sources (around 30GB on its complete SQL dump version). It includes ICD-10 data, MeSH terminology, SNOMED-CT codes [50]

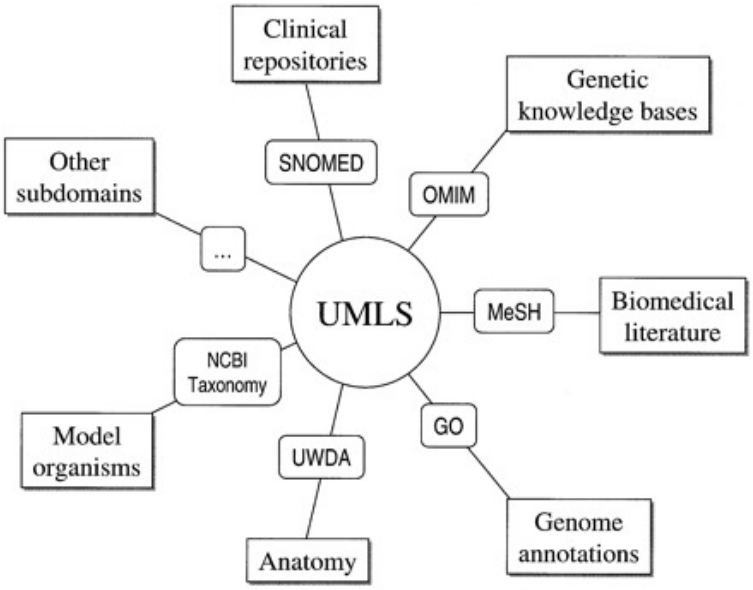


Figure 3.4: The Unified Medical Language System (UMLS), maintained by the U.S. National Library of Medicine is a composition of various biomedical terminologies and contains around 2.5 million concepts and over 12 million relations among these concepts. *Source: Bodenreider (2004) [5].*

and many more in different languages, including Japanese (see Figure 3.4). Since the data here is cross-mapped between the many different information sources, we can easily map the diseases from the records' ICD-10 codes to the MeSH terms of our HSDN disease list. However, many diseases on our list had no related code and the mapping was imperfect for the many diseases we had. To solve this, we put additional codes obtained from other mapping sources not present in the database like the SNOMED-CT to ICD-10 mapping and Wikipedia's WikiData, hence obtaining a longer list of ICD-10 possible codes covering as many diseases as possible. This allowed us to more robustly count the diseases from the medical records to finally get a better estimate of the real frequency of the diseases in a normal Japanese hospital. Diseases with no possible ICD-10 mapping after this process were also ultimately discarded. Tables 3.1-a and 3.1-b show the current most common diseases when considering all medical visits and only emergency room visits.

Frequency values are a good estimate of the prior probability for a disease, but some diseases are exceptionally more common than others, which may be of higher importance. To somehow smooth these values so the difference in prior probabilities is not that abrupt

Nº	Disease Name	Frequency
1	Diabetes Mellitus	0.03106
2	Gastroesophageal Reflux	0.01937
3	Esophagitis, Peptic	0.01932
4	Pneumonia	0.01865
5	Hypertension	0.01804
6	Rhinitis	0.01560
7	Dehydration	0.01549
8	Spondylarthropathies	0.01525
9	Heart Failure	0.01511
10	Bronchitis, Chronic	0.01471
11	Stomach Ulcer	0.01375
12	Respiratory Tract Infections	0.01264
13	Liver Diseases	0.01256
14	Intestinal Polyposis	0.01201
15	Low Back Pain	0.01142
16	Influenza, Human	0.01114
17	Cerebral Infarction	0.01111
18	Angina Pectoris	0.01081
19	Gastrointestinal Diseases	0.010706
20	Enteritis	0.010702

(a)

Nº	Disease Name	Frequency
1	Gastrointestinal Diseases	0.04168
2	Enteritis	0.04167
3	Influenza, Human	0.04131
4	Respiratory Tract Infections	0.03354
5	Pharyngitis	0.03127
6	Craniocerebral Trauma	0.03055
7	Cerebral Hemorrhage	0.02707
8	Dehydration	0.02555
9	Brochitis, Chronic	0.02524
10	Pneumonia	0.02326
11	Common Cold	0.01777
12	Status Asthmaticus	0.01629
13	Rhinitis	0.01457
14	Urticaria	0.01403
15	Cerebral Infarction	0.01384
16	Myocardial Infarction	0.01217
17	Asthma	0.01185
18	Asthma, Aspirin-Induced	0.01181
19	Bronchial Hyperreactivity	0.01180
20	Whiplash Injuries	0.01109

(b)

Table 3.1: Tables showing the most frequent diagnoses in the data consisting of approximately 1 million medical records. (a) considering all kinds of medical visits and (b) considering only emergency room visits. *Example constructed from author's data. Source: Tsuoyama Hospital, Okayama, Japan.*

(causing the same diseases to be part of the pre-test diagnosis over and over again), we applied a smoothing hyperbolic tangent function:

$$P(f) = \begin{cases} A \cdot tanh(T \cdot f - B) + C) & \text{if } f > 0 \\ P_0 & \text{if } f = 0 \end{cases} \tag{3.5}$$

where A, B, C and T are empirically adjusted constants, and P_0 is a *minimum prior probability* to attain for diseases that cannot be discarded yet, but were not even present in the medical records.

3.2.2 Symptoms

Structure

The children nodes represent the findings and these are divided in two: symptoms, which are simply patient complaints and easily observable manifestations that are part of the input from which to make the exam recommendation, and exam results and others, which are what is going to be recommended based on the initial set of symptoms and personal information.

Symptoms were also obtained from the Human Symptoms-Disease Network and its links were used as arcs for the models (figure 3.3). However, the number of symptom links per disease was still too large and some symptoms having a tiny lexical connection (co-occurrence in medical literature) with a certain disease seemed to be not really relevant. This way, the ones with a too low TF-IDF score in relation to others for the same disease (around the bottom 10% of the highest score) were ignored.

Probabilities

The conditional probabilities required for children nodes are sensitivity and specificity, as we have seen earlier. These values are more difficult to calculate than prior probabilities so other techniques had to be used to estimate them.

Fortunately, the TF-IDF co-occurrence score present in the symptom-disease links from the HSDN seems to be an acceptable representation of the relevance of these relationships. Thus, we used these values to estimate sensitivity and specificity values by mapping them through appropriate functions.

Sensitivity $P(F_i|D_j)$ represents the probability that a finding (or symptom in this case) is present when a certain disease is known to be present. Assuming that the TF-IDF co-occurrence score somehow measures the relevance of a symptom-disease pair, we can pass it through a function to estimate sensitivity directly. In this case we used the logarithm function to do this:

$$S_e(\hat{s_{ij}}) = P(F_i|D_j) = A \cdot log(\hat{s} + T) + B \tag{3.6}$$

where $\hat{s}$ is the relative TF-IDF score of the symptom against the highest score of any symptom for the same disease $s_{ij}/s_{max(j)}$, and A, B and T are empirically adjusted constants.

Similarly, specificity $P(\bar{F}_i|\bar{D}_j)$ represents the probability that a symptom is not present when a certain disease is known to be absent. This measures how *specific* a symptom is, or how rarely it gets manifested when the cause is other than that particular disease. Assuming this value is inversely correlated to the presence of a symptom in the total number of diseases, we can estimate it like:

$$S_p(\hat{p_i}) = P(\bar{F}_i|\bar{D}_j) = A \cdot (1 - p_i) + B \tag{3.7}$$

where p_j represents the *presence* or frequency of the symptom as a relevant finding within the set of all diseases, and A and B are empirically adjusted constants. Note that this estimation is much simpler since it only depends on the finding, resulting in the same specificity value for any disease.

3.2.3 Exam Results and Others

Structure

The second type of findings (our children nodes) are the exam results and other types of non-symptom signs. These types of data were not present in the HSDN or the available medical records so they had to be gathered from somewhere else.

Thus, exam results and other findings were scrapped and assembled from 3 different web databases containing information about hundreds of findings (including exam results), related

diseases and, in many cases, sensitivity and specificity values. These websites are *Get The Diagnosis* [44], *Sensitivity-Specificity* [45] and *Essential Evidence Plus* [51]. They were built by members of the medical community to organize these type of data which is extracted from concrete medical studies. Their goal is to make this information and values easily available for anyone. Most of these findings include at least one of sensitivity and specificity values, as well as the associated diagnosis to which the finding is related.

Now, since these websites only include a disease name to relate to each finding, another matching process is required to assign the new findings to our already established disease nodes. In many cases the MeSH disease term that we have matches the name of the scrapped finding's associated disease, but sometimes it is very different. For this reason we used the UMLS service again to search for the correct mapping of diseases, since it includes many alternative names for the same disease component. The list of matched diseases was then manually checked, completed and cleaned to obtain links for as many diagnoses as possible.

Probabilities

These findings' sensitivity and specificity values were most of the time included so the values were stored directly. Some of these values were of ranges, e.g., 50%-70%, and sometimes a finding was repeated, but with different values (coming from different studies); in these cases an average value was stored instead. Finally, when one of the two probability values was missing, an estimated default sensitivity or specificity value was used; if the two values were missing, the finding was discarded.

3.2.4 Clustering

Having completed the knowledge base with findings and diseases, we intend to decrease the number of possible diagnoses because it still remains too high for practical purposes. Since we do not want to lose the information of findings whose related disease is no longer part of the network, the goal is to simply group very similar diseases under one *main* disease.

To do this we used a very common clustering method called K-Means [52]. In simple terms, its goal is to partition n observations into k different clusters where each observation

is assigned to the cluster with the nearest mean, serving as a cluster's 'leader'. The result is a partitioning of the data space into Voronoi cells.

The input data was each disease's various ICD-10 code alternatives gathered before the frequency calculation (Section 3.2.1), as we assumed that similar diseases shared a larger number of ICD-10 codes. The cluster *leader* or prototype (main disease) was set to the one with higher frequency, or the closest to the centroid in case of a tie. At first the number of clusters was set to 2,200 but later decreased to 1,800, which seems to be closer to the practical number of possible diagnoses.

Original Disease	ICD-10 Codes	Cluster Leader
Inflammatory Bowel Diseases	K529	Inflammatory Bowel Diseases
Colitis	K529, K500	Inflammatory Bowel Diseases
Gastrointestinal Diseases	P779, K529, P789, K929, K9429, K319, K3189	Inflammatory Bowel Diseases
Gastroenteritis	P779, K529	Inflammatory Bowel Diseases
Common Cold	J00	Common Cold
Nasopharyngitis	J00, J311	Common Cold
Nephrosis	N289	Nephrosis
Kidney Diseases	N289, N08	Nephrosis
AIDS-Associated Nephropathy	B20, N289	Nephrosis

Table 3.2: Examples of clusters generated using the K-Means algorithm. The first 4 diagnoses all correspond to the same type of very common bowel diseases, the next 2 correspond to common cold and the last 3 are all kidney diseases. In most cases these can be considered equivalent upon recommending pertinent examinations. *Examples constructed from author's data.*

These clusters were finally quickly revised by a physician who suggested a few corrections, and thus it was used as the final list of diseases. Table 3.2 shows a sample of the resulting clusters. For exam recommendation purposes each cluster's diagnoses can be considered equivalent.

3.3 Exam Recommendation

3.3.1 Findings Processing

In order to make the recommendations more robust and more generally applicable, some additional processing and considerations were needed. Most of this procedures were done manually and one-by-one because the extracted medical findings did not present any standardized structure, so no moderately simple script could handle every case.

The findings obtained from internet consisted of signs, physical examinations, image and laboratory exam names (sometimes with results) and symptoms. Since the symptoms were already part of the data and served a different purpose (of being used as an input to calculate pre-test probabilities), the newly obtained symptoms were tagged in order to be merged with possibly already present HSDN symptoms so they can be ignored when doing the exam recommendations. In cases of collision, we used the sensitivity and specificity values of the symptom's version that was extracted from the web databases, since the other version had only an estimate.

On the other hand, the goal is to recommend examinations and these can have different results for different diagnoses, but performing one exam can be potentially useful to more than one disease. For this reason, the obtained findings were manually re-arranged to separately become *exam* and *results* (where possible). By doing this the number of exams was reduced greatly and the information became more understandable. Table 3.3 shows a sample of the resulting findings separated by name and result, along with their corresponding probability values and diagnosis. When no result was specified we used the default result *Abnormal / Present* to indicate the expected outcome.

Additionally, some exams and diseases had sex and/or age restrictions (e.g., men and infants can't develop gynecological diseases, women cannot develop prostatic diseases, some findings' values are only valid for children, etc.). To handle this, we introduced 3 new variables to the findings: *sex* (-1 for female, 1 for male and 0 for either), *minimum age* (0 for the default case) and *maximum age* (150 for the default age).

Exam Name	Exam Result	Diagnosis	Sensitivity	Specificity
CT Scan	Adenopathy	Appendicitis	62%	66%
Ultrasound	*Abnormal / Present*	Pneumonia	93.4%	97.7%
CSF Protein	> 100 (mg/dL)	Meningitis, Bacterial	69%	99%
Chest X-Ray	Pleural Effusion	Pulmonary Embolism	45%	65%
Tachycardia	*Abnormal / Present*	Cardiac Tamponade	77%	100%
CT Scan	*Abnormal / Present*	Cholelithiasis	96%	91%

Table 3.3: Examples of processed findings. Some exams can have different results for different diagnoses. When no result was specified the default result *Abnormal / Present* was used. *Examples constructed from author's data.*

3.3.2 Recommendation Process

To determine which exams improve the diagnostic probability the most for each of the diseases, their marginal probabilities have to be estimated. Then, the recommended exams can be listed according to how much the probability changes when certain exam result is present or absent, and how many different diagnoses they are potentially useful for.

Since calculating the marginal probability for each finding related to the diseases in the list implies running the algorithm through the whole network, an algorithm that takes too long (like the sampling algorithm of the BN2O model) would be very impractical. For this reason, the marginal probability of each exam is calculated applying the Bayesian Trees model's algorithm to the pre-test results (no matter if the original model was BN2O or Bayesian Trees) and then added to the pre-test probabilities to finally organize the recommendation of exams.

With this, the exam recommendation is made as follows:

1. A patient's basic information along with a set of 5 or 6 symptom names is used as input.

2. A list of pre-test disease probabilities is calculated specifying one of the two models above.

3. Based on these list, the non-symptom findings related to each disease are grouped together by examination name by estimating marginal probabilities using the Bayesian Trees model.

4. A final ranking is output based on how many diseases from the list it helps diagnose and also on the increase or decrease of disease probability the presence or absence, respectively, causes.

Diagram of figure 3.5 outlines a rough scheme of the process and the required data in each step.

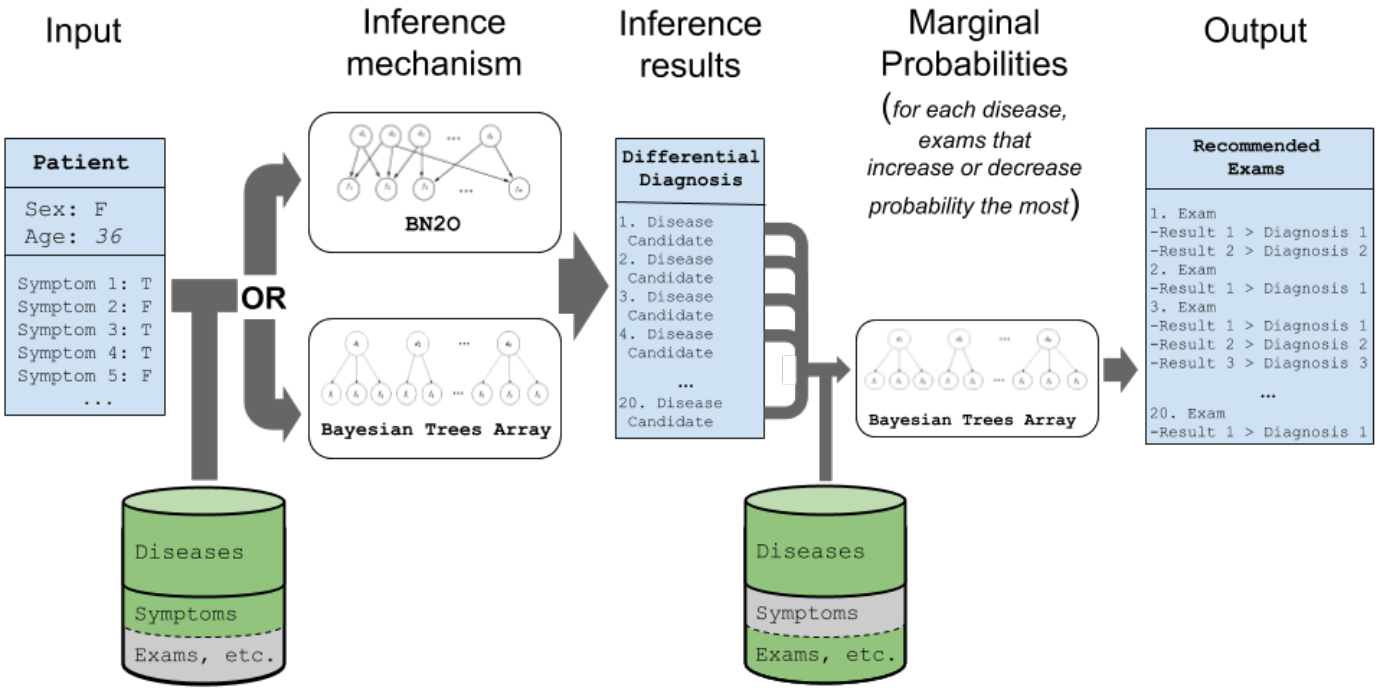


Figure 3.5: The inference process and the data used on each step. First, basic patient data is used as evidence for the first inference call with the specified algorithm using diseases and symptoms data as the BN structure. The output is a list of pre-test disease candidates. For each candidate, its associated exams are tested as evidence on the simpler Bayesian Trees model to determine the exams that increase or decrease probabilities the most. The final output is the list of recommended exams ordered by disease and marginal probability.

Chapter 4

Results and Analysis

4.1 Foreword

Before jumping into the details of the evaluation it is very important to state that to get conclusive results, a much more detailed trial than the one performed is probably required. Compared to other sciences, the medical sciences rely heavily on probabilities of many kinds. Even the judgment of different physicians for the same set of symptoms is far from being a deterministic function, as it is better described as a probability distribution. Medical literature and guidelines are much more consistent in this regard and are the way to go for this type of evaluation, but they still need to be managed and used by medical professionals because of the complexity of some medical cases.

With all this in mind, a complete evaluation would require to have at least a few different experts from different specialties, including emergency room professionals, reviewing different sources of medical literature in a trial with strict settings. This would be highly desirable to prove our hypothesis in a conclusive manner, but reality is that it would take a huge amount of financial and time resources (probably closer to the ones available for a PhD thesis). Unfortunately, this is far from being the case even with financial aid from the aforementioned company so we had to largely scale this process into what we call a preliminary evaluation where we aim to get a taste of the conclusive results in a general first inspection. We are fully aware that this is not the best evaluation to perform for a strong hypothesis as the one stated in the introduction, but we also believe that it achieves our desired goal of showing the

more general direction of these results and some interesting points that can be still analyzed on a first inspection.

That being said, in order to test our system across various dimensions, the desired comparisons had to be established. We want to compare not only both models' accuracy and speed but also how results may change for different inputs: patients of different age and sex. Additionally, we want test cases for different types of diagnoses to have a general idea of how the system works on a subset of the most relevant emergency room visits, such as common cases, high risk-of-death cases and age or sex-restricted cases.

The general evaluation of these results was made by two different physicians. One Chilean and one Japanese, who did not know each other. They kindly accepted our request to take an objective look at and review the results, where medical expert revision is mandatory. For this reason, the number of comparisons and results to be evaluated had to be decided beforehand on a number of cases big enough to draw decent conclusions, but compact enough to take into account both experts' limited availability. Of course, since the number of reviewers is reduced and it is ultimately an important factor it would be desirable to include in the analysis a comparison of their evaluations as well.

We begin by comparing the two models on a subset of the testing data, at the same time we compare the results of the review made by both experts. Once the best-performing algorithm has been chosen, we try out some comparisons on how the results for the same diseases change for different sexes. Finally, we observe any possible change in the results related to the age of the patients.

In this chapter we show some results of these cases mostly in the form of stacked charts and discuss them immediately after. More general conclusions are presented in the next chapter.

4.2 Test cases

Ten diagnoses were chosen as our testing set to cover for the most relevant disease categories. The first seven cases correspond to very common emergency room diagnoses from different specialties like gastroenterology, nephrology and neumology. The next two correspond to common diagnoses within particularly urgent cases, i.e., with an elevated risk of death. The remaining one is a female-restricted pregnancy related disease chosen to address a more restricted case.

These diseases and their related symptoms (presence-of but sometimes absence-of) were obtained from diagnostic-oriented medical literature [53] and were adjusted with help of the Chilean expert, based on his ER documented data and experience. They contain not more than 5 related symptoms each, resembling data that can be extracted from an initial patient interview. This information is presented in Table 4.1.

Finally, to cover for different age ranges we established *range representatives*: 5 patients aged 9, 29, 49, 69 and 89, from both sexes.

Also, the number of pre-test diagnoses as well as recommended exams displayed (and of interest for our study) is limited here to the top 20, but it is of course adjustable.

4.3 Bayesian Network Model Comparison

We begin comparing the two implemented models by testing all the diseases for a generic patient with the same personal information. In this case, we set the patient to be a middle-aged 49-year old male.

4.3.1 BN2O Network

The first model tested was the BN2O Network. As discussed in Chapter 3, given that exact inference on a network of this size is intractable we had to use Likelihood Weighted sampling.

The number of iterations estimated necessary for convergence was estimated to be around 5000. With this number it takes in average 404 seconds in total to calculate pre-test prob-

Diagnostic	Present Symptoms	Absent Symptoms
Gastroenteritis	Abdominal Pain, Diarrhea, Vomiting, Chills, Fever	
Pharyngitis	Cough, Fever, Headache, Odynophagia	
Urinary Tract Infections	Dysuria, Abdominal Pain, Urinary Urgency, Pelvic Pain	
Otitis	Earache, Hearing Loss, Fever	Cough
Allergic Rhinitis	Cough, Dysphagia, Pruritus, Itchy Eyes, Eye Manifestations	
GERD*	Heartburn, Cough, Abdominal Pain	Dyspnea
Stroke	Dysarthria, Hemiplegia	Angina Pectoris, Palpitations
Myocardial Infarction	Chest Pain, Dyspnea, Nausea	Fever
Ectopic Pregnancy	Abdominal Pain, Pelvic Pain, Abdomen Acute, Adnexal Mass	Diarrhea

Table 4.1: Diagnoses to be tested, along with their present and (optional) absent symptoms that serve as input. *GERD stands for Gastroesophageal Reflux Disease. *Source: "Differential Diagnosis in Internal Medicine", by Laso Guzmán, March 1997 [53], adjusted using the Chilean phyisician's ER data and experience.*

abilities and suggest the exams for one diagnostic case. We believe this number is near the limit of practical use in an emergency room scenario. From a medical perspective, anything longer than a few minutes would make the system somehow useless in such circumstances.

Pre-test Diagnosis

Table 4.2 shows a sample of the pre-test probabilities calculated using the BN2O network model, considering input data for two emergency room diseases of high relevance: gastroenteritis and stroke. The recommendation of exams is made based on these rankings.

It is important to note that the implied disease name used as input may be not present in the database as it is written since it could have been merged with other very similar diseases

Nº	Disease (Intended: Gastroenteritis)	Probability
1	Influenza, Human	5.618%
2	Pharyngitis	2.811%
3	Enteritis	2.64%
4	Gastrointestinal Diseases	2.62%
5	Dehydration	1.801%
6	Respiratory Tract Infections	1.707%
7	Malaria, Vivax	1.387%
8	Lymphoma, Non-Hodgkin	1.356%
9	Liver Abscess	1.182%
10	Common Cold	1.05%
11	Arthritis, Infectious	0.999%
12	Sporotrichosis	0.976%
13	Hernia, Hiatal	0.971%
14	Hyperaldosteronism	0.932%
15	Liposarcoma	0.887%
16	Empyema, Tuberculous	0.873%
17	Acquired Hyperostosis Syndrome	0.865%
18	Tuberculosis	0.802%
19	Appendicitis	0.792%
20	Liver Failure, Acute	0.791%

Nº	Disease (Intended: Stroke)	Probability
1	Cerebral Infarction	2.623%
2	Cerebral Hemorrhage	2.078%
3	Mutism	1.347%
4	Gastrointestinal Diseases	1.301%
5	Tongue Diseases	1.296%
6	Intracranial Embolism	1.18%
7	Influenza, Human	1.16%
8	Pulmonary Embolism	1.031%
9	Hypoglycemia	0.947%
10	Pseudoxanthoma Elasticum	0.943%
11	Brain Ischemia	0.878%
12	Enteritis	0.82%
13	Bronchitis, Chronic	0.818%
14	Hypertension	0.815%
15	Thromboembolism	0.812%
16	Pharyngitis	0.787%
17	Syncope, Vasovagal	0.747%
18	Gingivitis, Necrotizing Ulcerative	0.744%
19	Aortic Valve Stenosis	0.734%
20	Myocardial Infarction	0.719%

Table 4.2: Pre-test disease probabilities using the BN2O network model (which uses Likelihood Weighting sampling) for 2 of the 10 test implied diagnoses: Gastroenteritis (left) and Stroke (right). The exam recommendation heuristic is based on these rankings.

(e.g. Gastroenteritis is clearly a Gastrointestinal Disease and is also basically identical to Enteritis for practical purposes, according to a medical expert).

Exam Recommendation

Once the next step of the algorithm outputs the recommended exams based on the already calculated pre-test probabilities, these are classified by the medical experts to fit into one of two categories. These categories were established to more accurately categorize the probabilistic score given by the exams, as this is not 'black or white'. We believe this representation is flexible enough to model this probabilistic nature and simple enough to avoid too much

complexity for the evaluating agent and the analyst.

Exams in **green** correspond to the *golden standard*, which are exams generally recognized to be very useful to reach (or discard) the particular diagnosis and are typically recommended by practitioners in the presence of a set of symptoms like the one established (without considering costs), i.e., in theory they increase or decrease the probability the most.

Exams in **yellow** correspond to exams that can help determine or are related in some way to the diagnosis but are not within the first options. They would not be recommended as first options but are still useful for reaching or discarding a diagnosis. In theory, they increase or decrease the probability in some significant manner.

The amount of exams in the top 5 or top 20 depends on the pre-test diagnoses that were discussed before. The top 5 stacked bar shows how close to the top the recommended exams are (and it is closer to the number of exams a physician could ask for at once). It is fundamental to understand that the height of these bars is intrinsically variable and it only depends on the total number of exams related to a certain diagnosis that are available in the database. What we need to look for when observing these results is the relative presence or absence of exams in the top 20, and more importantly in the top 5, because this means that correct exams are in fact present within the first options, which is our ultimate goal.

Using this information for all of the diseases the list of recommended exams for the test patient can be elaborated. These exam rankings were reviewed by both medical experts and their results can be seen in Figure 4.1 (on the top for the Chilean physician and on the bottom for the Japanese physician). It shows, for each disease, the percentage of the recommended exams that are considered correct (very useful for the particular diagnosis) in green and semi-correct (moderately useful for the particular diagnosis) in yellow, within the ranking's top 5 (left column) and top 20 (right column). Exams considered incorrect (useless for the particular diagnosis) are of course left out.

Figure 4.1: Exam recommendation results when using the BN2O model and the Likelihood Weighting sampling algorithm on the 49-year old male test patient. For each disease, the percentage of the recommended exams that are correct are in green and the semi-correct are in yellow within the ranking's top 5 and top 20. On the top, results according to the Chilean physician's criteria and on the bottom, results according to the Japanese physician's criteria.

4.3.2 Bayesian Trees Array

The second model tested was the Bayesian Trees Array. As discussed in Chapter 3, a form of exact inference is possible when using this network, therefore we used it. It takes in average 6 seconds in total to calculate pre-test probabilities and suggesting the exams for one diagnostic case. This number is very low and would be perfect for a practical scenario.

Pre-test Diagnosis

Table 4.3 shows a sample of the pre-test probabilities calculated using the Bayesian Trees Array network model (which uses 'exact inference'), with input data again for the same pair of relevant emergency room diseases: Gastroenteritis and Stroke.

Exam Recommendation

Figure 4.2 shows, for each disease (at the top for the Chilean physician and the bottom for the Japanese physician), the percentage of the recommended exams that are considered correct (very useful for the particular diagnosis) in green and semi-correct (moderately useful for the particular diagnosis) in yellow within the ranking's top 5 (left column) and top 20 (right column).

4.3.3 Model Comparison Analysis

Pre-test Diagnosis

To break down completely these preliminary results of the two models for the same patient, we first need to observe the pre-test diagnosis results (tables 4.2 and 4.3). These results have a very direct influence on the exam recommendation so if they are not accurate there is a high chance that the exam recommendation is not accurate as well.

An important observation is that these pre-test probability values are very close between each other for a particular case. To understand this it is important to remember first that the model outputs probabilities for virtually every one of the 3600+ diseases in the database, which means that hundreds of diseases not on the list are still possible and therefore have a non-zero probability value. Additionally, in many cases diseases with close probability values

Figure 4.2: Exam recommendation results when using the Bayesian Trees array model on the 49-year old male test patient. For each disease, the percentage of the recommended exams that are correct are in green and semi-correct are in yellow within the ranking's top 5 and top 20. The Chilean physician's criteria is at the top and at the bottom is the Japanese physician's criteria.

Nº	Disease (Intended: Gastroenteritis)	Probability
1	Gastrointestinal Diseases	1.841%
2	Enteritis	1.599%
3	Influenza, Human	1.575%
4	Pharyngitis	1.356%
5	Liver Abscess	1.052%
6	Respiratory Tract Infections	0.773%
7	Bacteremia	0.665%
8	Lymphoma, Non-Hodgkin	0.604%
9	Appendicitis	0.596%
10	Dehydration	0.524%
11	Gonorrhea	0.451%
12	Pancreatitis, Chronic	0.45%
13	Candidiasis	0.447%
14	Liposarcoma	0.424%
15	Pneumonia	0.417%
16	Adenocarcinoma, Follicular	0.404%
17	Liver Neoplasms	0.398%
18	Zygomycosis	0.396%
19	Empyema, Tuberculous	0.391%
20	Gastric Fistula	0.389%

Nº	Disease (Intended: Stroke)	Probability
1	Cerebral Infarction	1.107%
2	Cerebral Hemorrhage	0.665%
3	Intracranial Embolism	0.38%
4	Brain Ischemia	0.311%
5	Tongue Diseases	0.299%
6	Brain Diseases	0.268%
7	Sagittal Sinus Thrombosis	0.266%
8	Mutism	0.264%
9	Pulmonary Embolism	0.257%
10	Leukoencephalitis, Acute Hemorrhagic	0.254%
11	Gastrointestinal Diseases	0.234%
12	Enteritis	0.234%
13	Influenza, Human	0.233%
14	Movement Disorders	0.231%
15	Lipomatosis, Multiple Symmetrical	0.226%
16	Infarction, Middle Cerebral Artery	0.213%
17	Sparganosis	0.204%
18	Respiratory Tract Infections	0.204%
19	Toxoplasmosis, Cerebral	0.195%
20	Pharyngitis	0.194%

Table 4.3: Pre-test disease probabilities using the Bayesian Trees network model for 2 of the 10 test implied diagnoses: Gastroenteritis (left) and Stroke (right). The exam recommendation heuristic is based on these rankings.

happen to be this way because they are in fact similar (or belong to the same physiologic system) and therefore share many common recommended exams. Given that the end goal of the system is to recommend the appropriate exams, this is not a major issue.

Having said this, it is still clear that the pre-test diagnosis turns out to be quite sensitive as its input is at most 5 symptoms. Luckily, the probability values only get very close to each other as we approach the bottom of the list (and therefore the rest of the 'hidden values') but the top values are usually the most relevant for the next steps.

As for Gastroenteritis, we see that while the BN2O model (table 4.2, left) put Influenza and Pharyngitis in a higher position than Gastrointestinal Diseases and Enteritis (which, again, are basically Gastroenteritis for practical purposes), the Bayesian Trees model (table 4.3, right) put it on top. This is already a bit surprising since the former model is more complete and even takes longer to compute its results. This is not the case for the Stroke case where both model output correctly the diagnose on top of the list (tables 4.2 and 4.3, right).

Experts Criteria Comparison

Now, if we take a look at the exam recommendation results in figures 4.1 and 4.2, the first notorious observation that can be made is that results between both experts for the same setting are quite different in general, although they present similarities for some cases.

For some diseases the amount of correct and semi-correct exams in the top 5 as well as in the top 20 are similar between both experts (e.g. UTI by BN2O, Myocardial Infarction by BN2O, Otitis by BN2O, Cholelithiasis by BT, Pharyngitis by BT, etc.), with differences mainly on the proportion of exams each one considered correct and semi-correct on a certain case. Given that these cases all present correct or semi-correct exams within the top 5, they are the most interesting since they not only suggest good performance but also suggest objectivity and consistency on what would be considered the next diagnostic step. Although these graphs do not guarantee that the classified exams are exactly the same, it is safe to assume both experts agree that a subset of exams are adequate for these specific diagnoses. This suggests that it does not matter whether the patient is Chilean or Japanese, the process should be carried on similarly as it is most likely the same around the globe.

That being said, many other cases show clearly different results (Gastroenteritis by BN2O, GERD by BN2O, Allergic Rhinitis by BT, GERD by BT, etc.). For example, figure 4.1 for the BN2O model suggests that in the eyes of the Japanese expert, to diagnose Gastroenteritis the exams on the top of the ranking (i.e., the top 5) are mostly useful, while in the eyes of the Chilean expert they are not useful at all. This can be attributed to many reasons.

First, as it was mentioned in the introduction the recommended exams for the same

disease (or symptoms for that matter) can be very different between countries genotypically and fenotypically different as are Chile and Japan.

Second, despite support from medical guidelines, there are intrinsic medical opinion discrepancies between any group of physicians, even from the same country. These are part of the probabilistic nature of medical decisions as was discussed earlier and are the main reason a complete evaluation requires the opinion of as many experts as possible, so they can be considered a canonical error or variance.

Third, other sources of error like medically unclear or misunderstood outputs (i.e. problems with the original data), the experts' evaluation process itself and others that can never be disregarded.

More generally, we can also see that for the BN2O model (figure 4.1) the Japanese expert shows on average slightly better results, with 8/9 diseases with recommended exams in the top 5 versus 6/9 for the Chilean expert. On the other hand, for the Bayesian Trees model (figure 4.2) the Chilean expert shows on average slightly better results but both have 7/9 diseases with recommended exams in the top 5. Overall, there seem to be lightly more consistency for the Bayesian Trees model.

In general these are good preliminary results with respect to our hypothesis since we can find useful exams at the top of the recommended exams rankings in almost every case. We now discuss differences between models in more detail.

Model Comparison

If we now compare more directly the two models, we can observe that both models (figure 4.2 perform on average similarly. On a first sight, it is hard to tell which one performs *better* for this setting, since for the BN2O model (figure 4.1 the Japanese physician's results are better, but for the Bayesian Trees model the Chilean physician's results are better. We can not make any strong conclusion about which one is better on average, but we can see that there is slightly more consistency for the Bayesian Trees model. In the BN2O model, there are four cases where either physician marks correct exams (green) within the top 5 and the other does not, while for the Bayesian Trees model there are only two of these cases. The Bayesian Trees model also has, for both physicians, at least correct exams within the top

20 for every disease, while this is not the case for the BN2O model. Additionally, we know that the Bayesian Trees model is a much simpler algorithm that takes only a few seconds to calculate its output compared to the BN2O.

The general observation is that the simplest algorithm performs not only faster but even slightly *better*. This could be because of an incorrect choice of sampling size for the sampling method of the BN2O model or simply because the extra complexity is just not necessary and a much simpler method is sufficient enough. Regarding the former, we tried to make sure to find a sampling size where the algorithm had already converged, but is not too large to be impractical. Though this should not be the main cause, we can not completely discard it. It is possible that the sampling size chosen was too small and that much better results could be achieved by increasing it (although a real test for different sizes is hard to do with limited expert time on this algorithm).

As mentioned, making a decisive conclusion as to which algorithm is better is not possible even in the context of our small preliminary study, but we believe the arguments presented let us fairly pick the simpler Bayesian Trees as the winner, considering other factors such as time performance.

In order to continue our preliminary study, we conducted the next experiments using the chosen Bayesian Trees model to be able to use the experts' time more efficiently and to be at least a small step closer to the real potential of the implemented solution.

4.4 Sex Comparison

Since the Bayesian Trees model showed better results in time and surprisingly rather similar results similar in accuracy, the next comparisons are made using this model instead of the BN2O to make better use of the limited time resources.

To see how the recommended exams could change between male and female patients, a review was made using a subset of the diseases on female and male 29-year old patients.

4.4.1 Female

The subset of the chosen diseases that was evaluated on a female patient but also on a male patient includes Urinary Tract Infection (UTI), Allergic Rhinitis, Cholelithiasis, GERD, Stroke and Myocardial Infarction. Figure 4.3 shows, as in the above manner, for each of the 6 diseases the percentages of what are considered by the experts to be correctly and incorrectly recommended exams.

Additionally, Ectopic Pregnancy is a female-restricted disease that was also reviewed on the female patient for the Bayesian Trees model but also for the BN2O model. Figure 4.4 shows the experts' evaluation results for this disease.

4.4.2 Male

The results of the evaluation for the same subset of 6 diseases (excluding Ectopic Pregnancy) but for a 29-year old male patient can be seen in Figure 4.5.

4.4.3 Sex Comparison Analysis

A quick look at figures 4.3 and 4.5 tells us that there are almost no differences in the outcomes of male and female patients for the selected diseases. The differences between experts are practically the same as before, but both considered that there is no significant difference.

The reason for this is most likely the small number of sex restricted diseases and their related exams in the database, but equally or more importantly the low frequency of some of these diseases (which are mostly gynecological or pregnancy-related) compared to the selected relevant cases. What would really make a greater difference in the results would be to properly incorporate sex as a *risk factor* instead of only restrictions. This would calibrate at least prior probabilities to take into account sex differences in a more realistic manner. This will be discussed as future work in the following chapter.

That being said, there is a very slight difference in the Cholelithiasis and GERD bars considered by the Chilean expert. These 2 diseases (along with Urinary Tract Infections) correspond to either nephrology or gastroenterology diagnoses. The symptoms related to

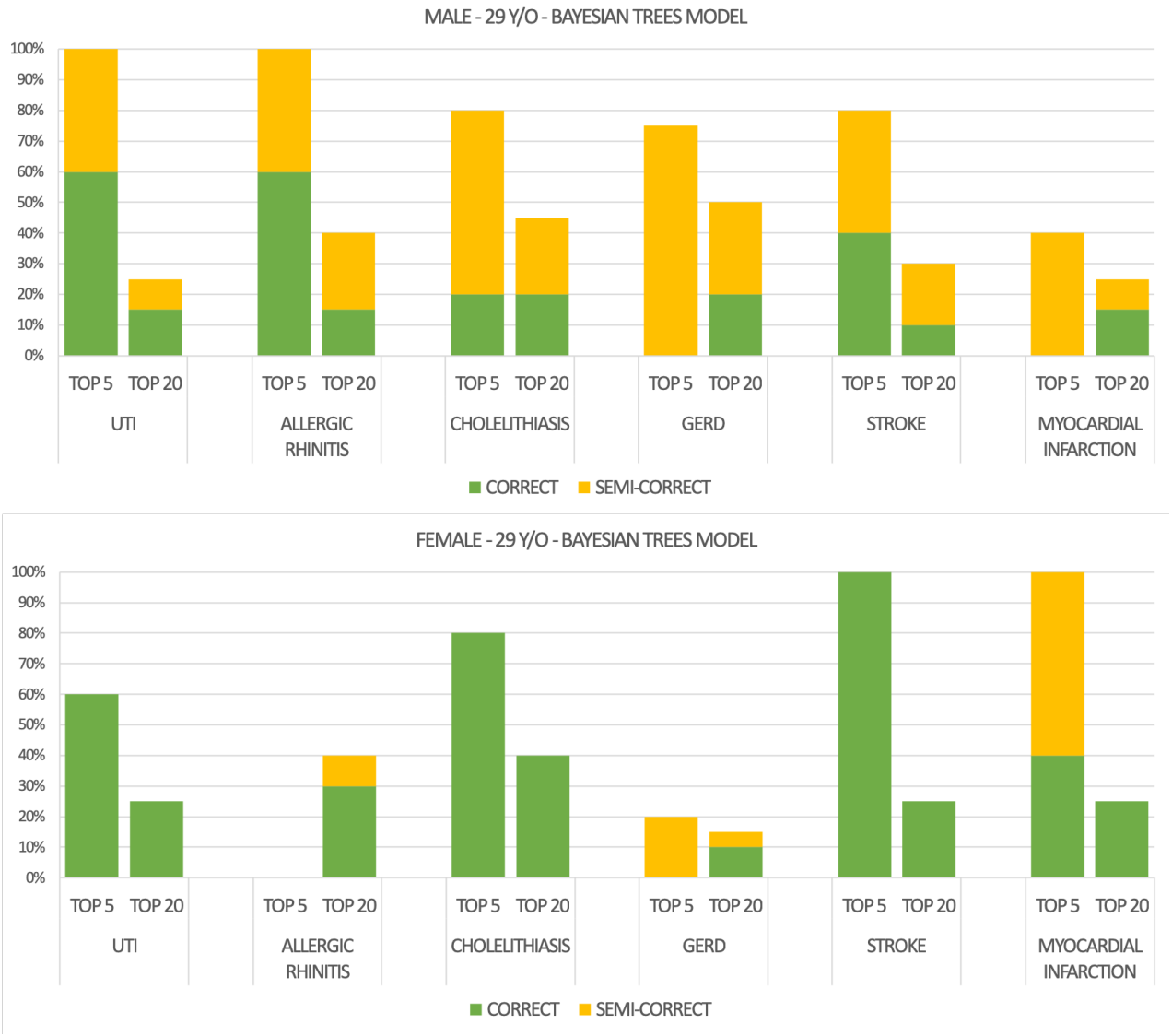


Figure 4.3: Exam recommendation results when using the Bayesian Trees array model on the 29-year old female test patient. For each of the 6 diseases, the percentage of the recommended exams that are correct are in green and semi-correct are in yellow within the ranking's top 5 and top 20. The Chilean physician's criteria is at the top and at the bottom is the Japanese physician's criteria.

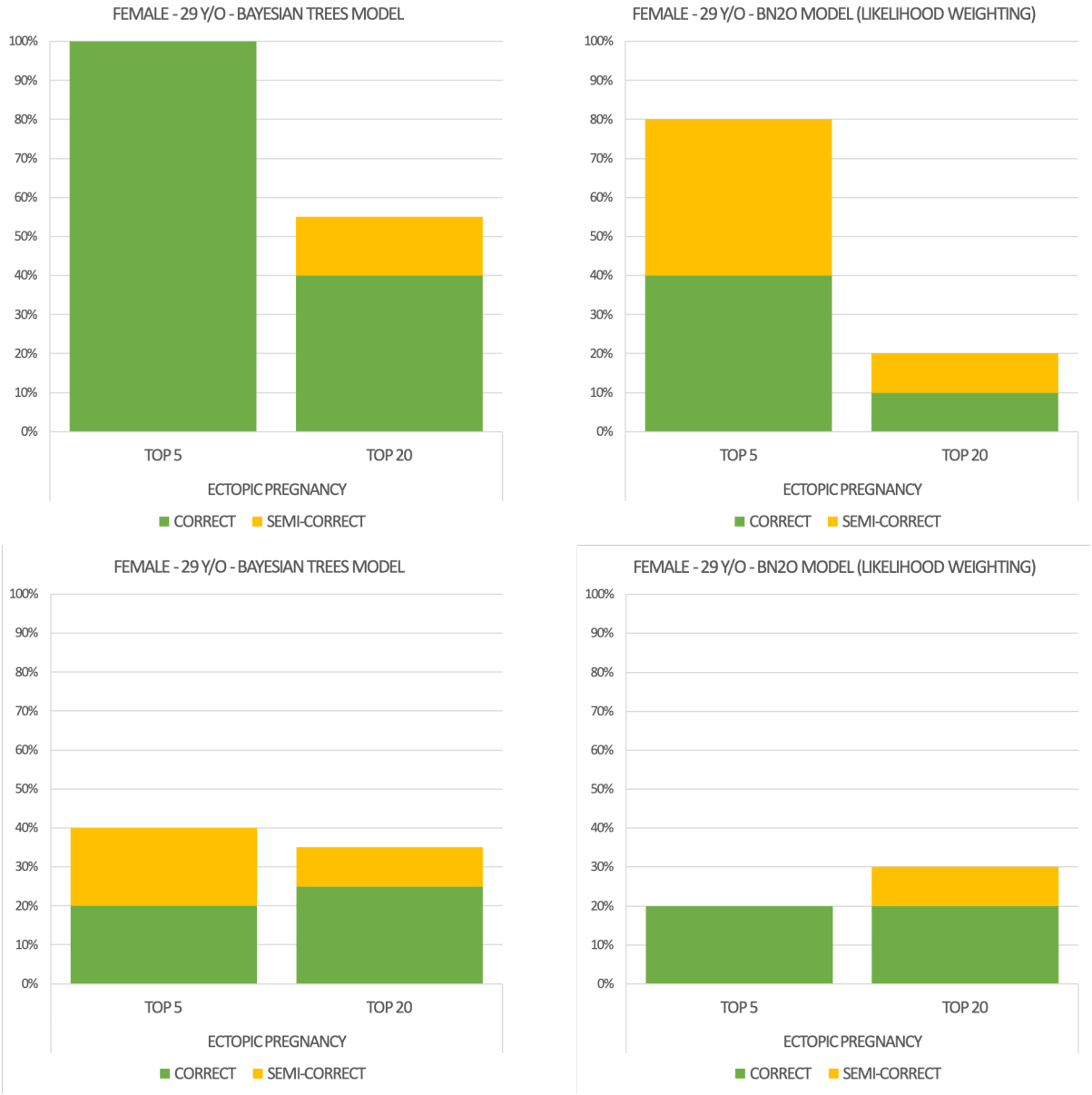


Figure 4.4: Exam recommendation results for Ectopic Pregnancy disease on the 29-year old female test patient when using the Bayesian Trees array model (left) and when using the BN2O model (right). The Chilean physician's criteria is at the top and at the bottom is the Japanese physician's criteria.

Figure 4.5: Exam recommendation results when using the Bayesian Trees array model on the 29-year old male test patient. For each of the 6 diseases, the percentage of the recommended exams that are correct are in green and semi-correct are in yellow within the ranking's top 5 and top 20. The Chilean physician's criteria is at the top and at the bottom is the Japanese physician's criteria.

these diagnoses are mainly in the abdominal area, very similarly to gynecological symptoms, making the algorithm possibly output these female-restricted diseases with a high probability and thus appear on the list. This is an interesting result since it puts in evidence how the symptoms actually manifest in the diseases and how these diseases vary (even if slightly) between male and female patients, but since this is not replicated in the case of the Japanese expert, this remains only a theory.

Upon analyzing these results with the Chilean medical expert, it was observed that in fact the gynecological (or more generally the sex-restricted diagnoses) are not that common in emergency room admissions as the other selected diagnoses, so explicit sex differences in diagnoses (i.e., restrictions) are not statistically fundamental. What may be more relevant but was left out of the scope of the algorithms are the risk factors' effect on the selected diagnoses.

Finally, figure 4.4 shows good results for Ectopic Pregnancy in general. This is also a female-restricted disease but the difference in performance of the two algorithms can be seen more clearly here. Both experts found correct exams within the top 5 of the list. The Bayesian Trees model put 5/5 correct exams on top of the list according to the Chilean physician and some others in the top 20 while the BN2O model only had 2 correct exams in the whole list. The Japanese physician also judged the Bayesian Trees model slightly better. These results contribute to our observation that with the established parameters the Bayesian Trees model outperforms the BN2O model by a small margin (within the scope of our preliminary study).

4.5 Age Comparison

Finally, as was discussed at the beginning of the chapter, we show results for 5 different ages of male patients on the 9 diseases: 9 years old (figure 4.6), 29 years old (figure 4.7), 49 years old (figure 4.8), 69 years old (figure 4.9) and 89 years old (figure 4.10), in order to cover for most of the age ranges.

Note that figure 4.8 (49 years old) is the same as figure 4.2 but was added again for better visualization and comparison. Also, figure 4.7 (29 years old) contains, among others,

the results for the 6 diseases on figure 4.5.

4.5.1 Age Comparison Analysis

Finally, regarding the results for different age ranges we can see that in most cases the accuracy remains unchanged, for both experts, between different-aged patients. As with the rest of the evaluation outcomes for the Bayesian Trees algorithm, results are quite promising and show the presence of correct exams in the top 5 in most of the cases. Moreover, both the Chilean and the Japanese expert indicated that for all of the selected diseases correct exams can be found in the ranking of the recommended top 20.

The only notorious differences between these results can be seen for the 9-year-old and 89-year-old patients. Since the evaluation of whether an exam was correct, semi-correct or incorrect is tied to the experts medical opinion we can only partially hypothesize on the cause of these differences and results in general. That being said, the most intuitive or 'common-sense' possible reason for these differences when compared to the more median-aged patients could be that some of these exams, even if useful for a certain diagnose, might be risky to perform on too young or too old patients, meaning some other type of tests should be required (or simply avoided).

Apart from that observation, we see that for both experts exams remain mostly unchanged for different ages, which is caused by the final exam lists being practically identical between cases. Even if the selected diagnoses are not strictly age-restricted, clearly the potential age restrictions in the exams that exist on the database are not affecting the results greatly for the selected diseases. This could be the result of an insufficient number of age-restricted exams (or restrictions on existing exams) but also of the absence of risk factors in the model and how they could have different probabilities for the same exam-disease pair between different ages.

4.5.2 General Analysis

In summary, one general observation that can be made upon checking the differences across different patient settings is that these differences are very small and almost negligible. For

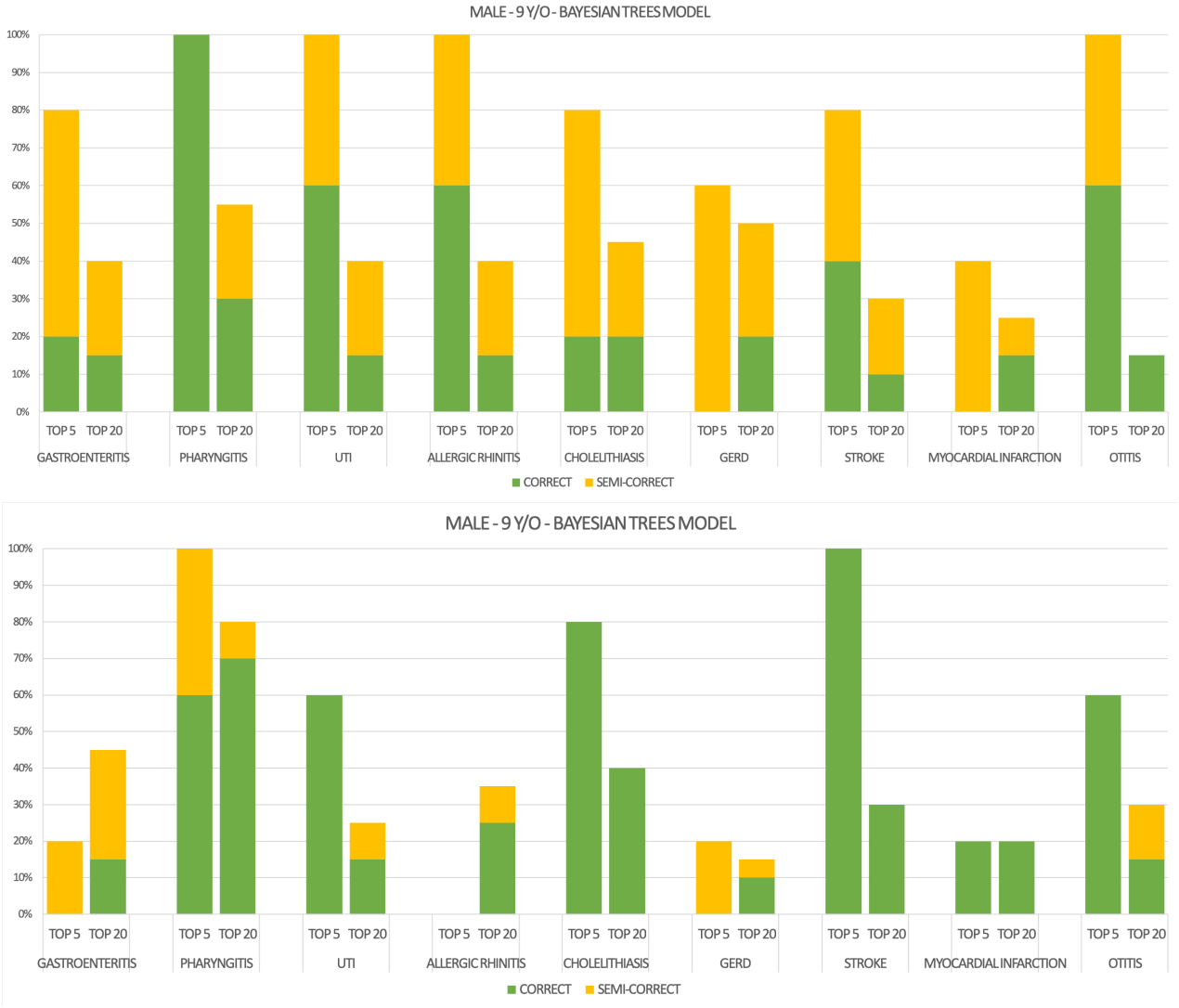


Figure 4.6: Exam recommendation results when using the Bayesian Trees array model on the 9-year old male test patient. The percentage of the recommended exams that are correct are in green and semi-correct are in yellow within the ranking's top 5 and top 20. The Chilean physician's criteria is at the top and at the bottom is the Japanese physician's criteria.

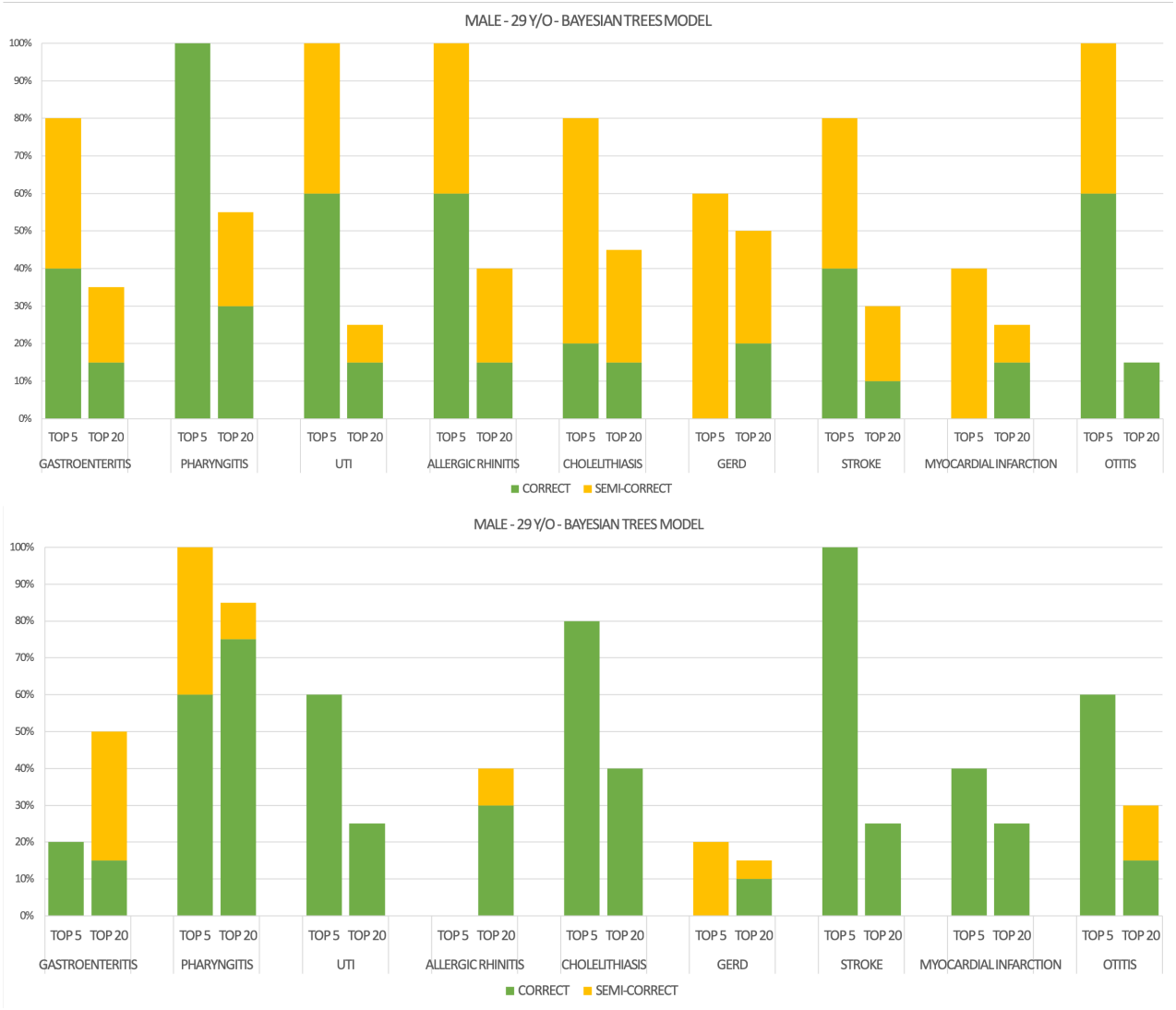


Figure 4.7: Exam recommendation results when using the Bayesian Trees array model on the 29-year old male test patient. The percentage of the recommended exams that are correct are in green and semi-correct are in yellow within the ranking's top 5 and top 20. The Chilean physician's criteria is at the top and at the bottom is the Japanese physician's criteria.

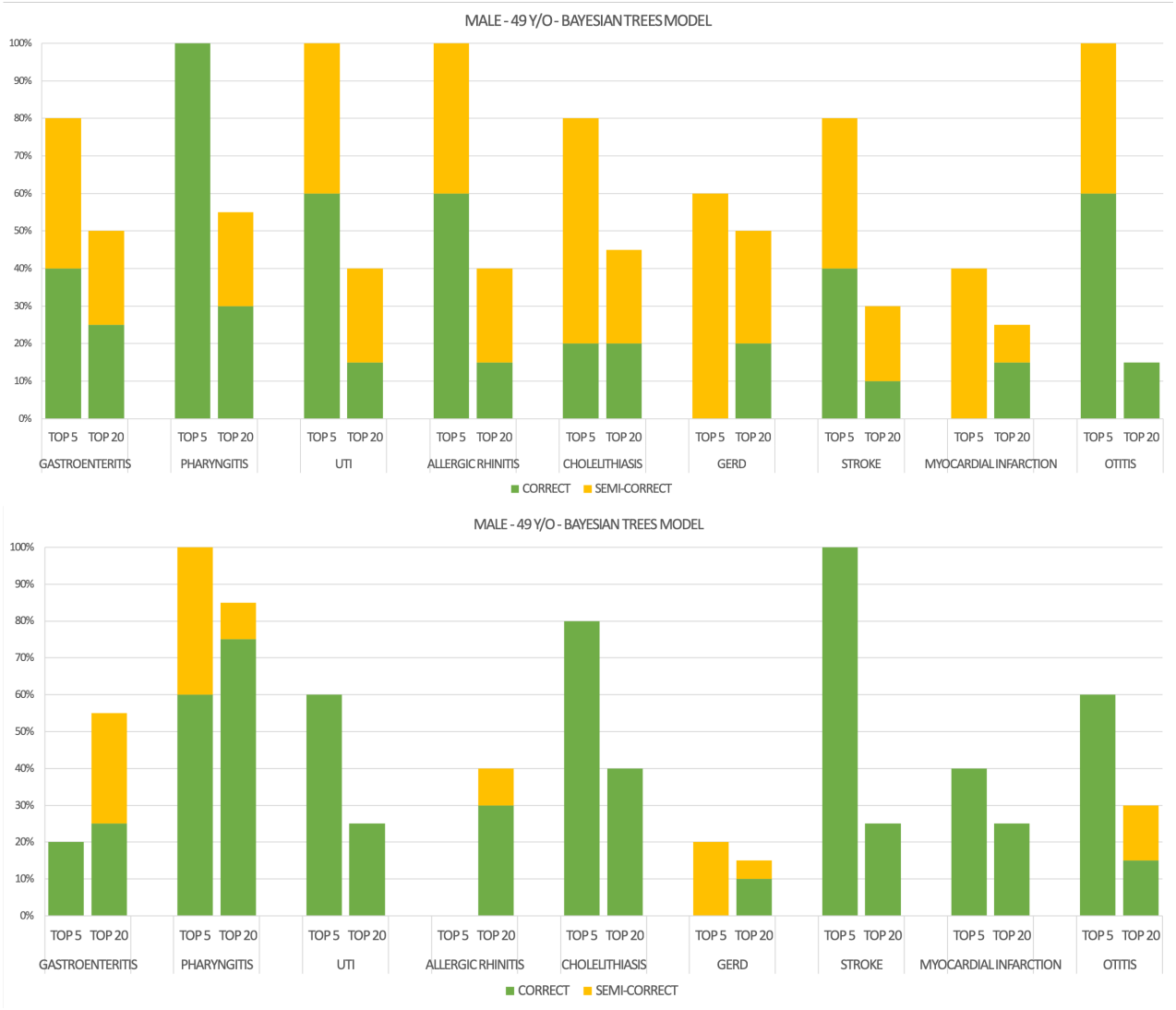


Figure 4.8: Exam recommendation results when using the Bayesian Trees array model on the 49-year old male test patient. The percentage of the recommended exams that are correct are in green and semi-correct are in yellow within the ranking's top 5 and top 20. The Chilean physician's criteria is at the top and at the bottom is the Japanese physician's criteria.

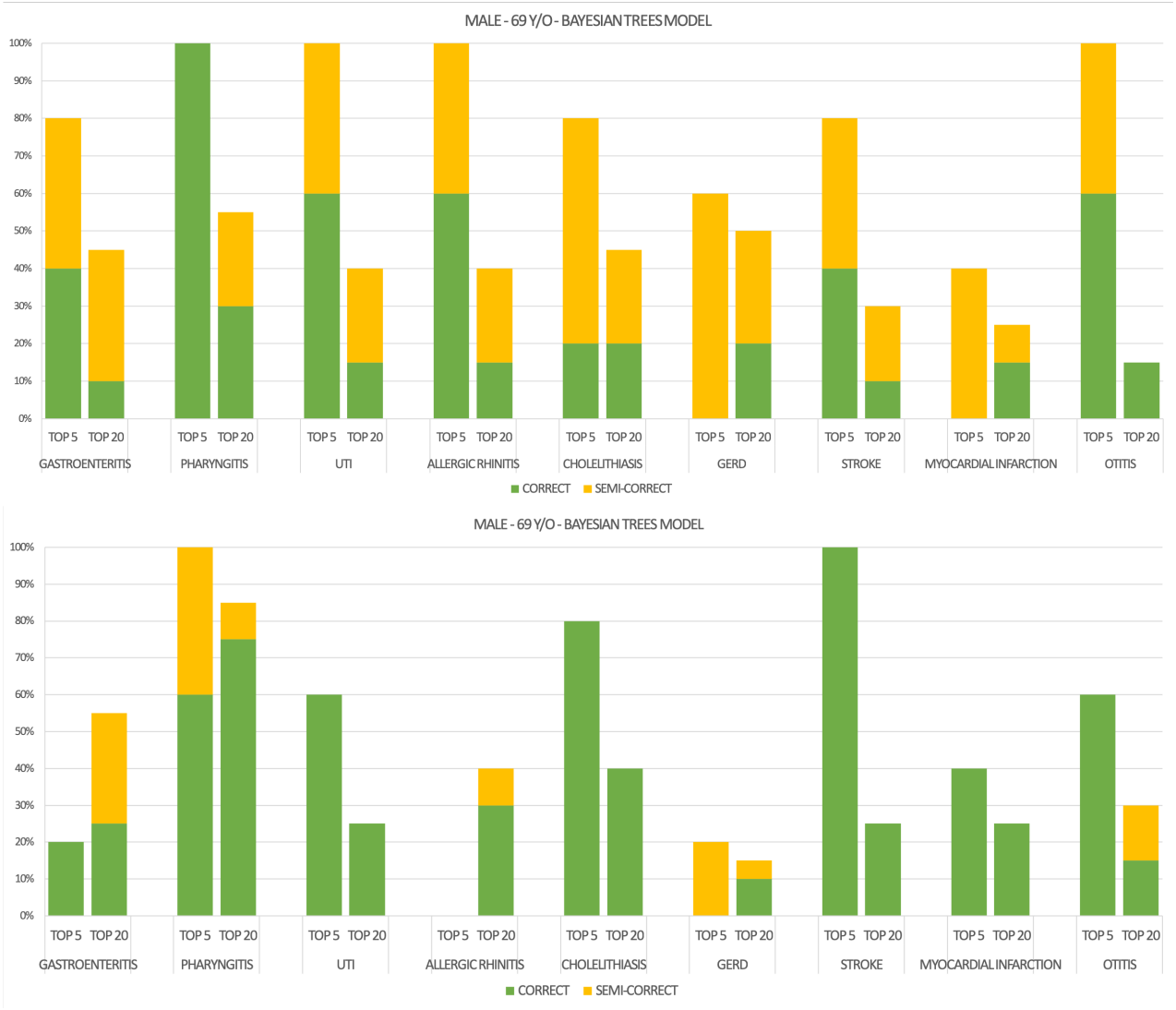


Figure 4.9: Exam recommendation results when using the Bayesian Trees array model on the 69-year old male test patient. the percentage of the recommended exams that are correct are in green and semi-correct are in yellow within the ranking's top 5 and top 20. The Chilean physician's criteria is at the top and at the bottom is the Japanese physician's criteria.

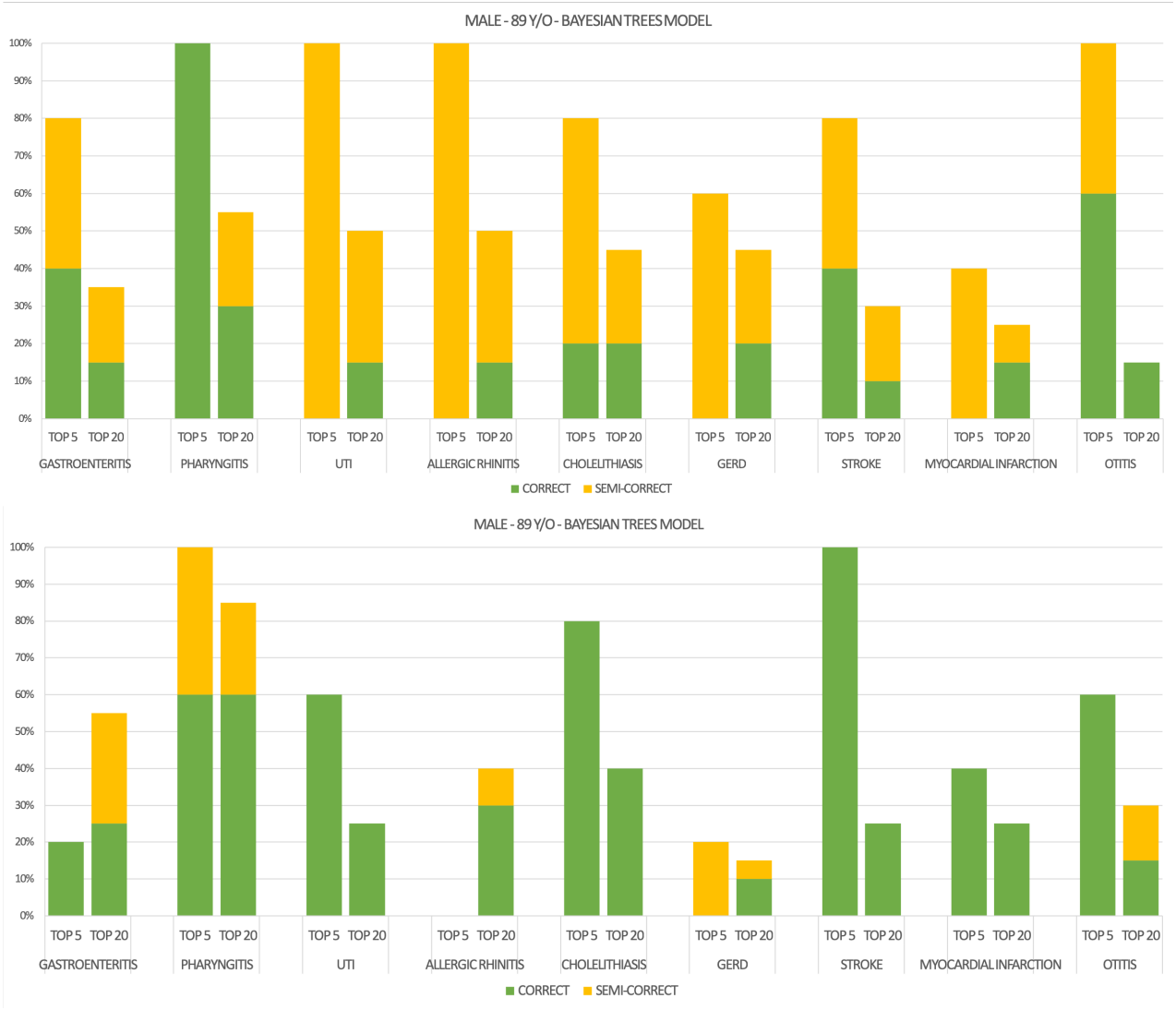


Figure 4.10: Exam recommendation results when using the Bayesian Trees array model on the 89-year old male test patient. The percentage of the recommended exams that are correct are in green and semi-correct are in yellow within the ranking's top 5 and top 20. The Chilean physician's criteria is at the top and at the bottom is the Japanese physician's criteria.

very different patient groups the results are almost identical, which is a little disappointing because it shows how patient information is much more irrelevant than it should be. As was discussed, the cause of this might be the small amount of age or sex-restricted exams and diseases that are in some way related to the chosen diagnoses along with the lack of modeled risk factors.

On the positive side, in the context of our limited preliminary evaluation, the results are quite good according to both medical experts. They show a significant percentage of relevant exams on the outcome of most of the tests, which is especially relevant considering the time restrictions of a hypothetical emergency room scenario.

Regarding the difference between both experts criteria, most of the relevant points were discussed in section 4.3.3. It is clear that the lack of strong differences between patient types manifested equally on both experts evaluation so the analysis does not change for different settings. However, both agreed that exams for male and female patients are virtually the same in the context of the emergency room, where the most relevant diseases are not typically sex-restricted, and they both agreed that the recommended exams for very young or very old patients are not *exactly* the same as for normal adults. We believe this at least shows that the approach used to classify exams is similar between both experts, even if their answers are quite different.

In the following chapter we will gather up more general conclusions about these results, the project itself and the medical point of view of the models' performance.

Chapter 5

Conclusions

In this chapter we draw some conclusions about the results obtained from our preliminary study and about the project in general. We include a more technical discussion but also focus on medically oriented conclusions. Finally, some brief words on future work and how the system could be improved and be properly tested on both technical and medical grounds.

First off, considering that the original list of diseases consists of about 4,000 names, it is possible to observe that in most cases specifying the presence of only a few symptoms can lead to good accuracy on both the pre-test diagnosis and the later exam recommendation. This, along with the fact that some of the exams serve for different diseases that share similar symptoms, shows that the symptom-disease links in the HSDN (after filtering and mapping their TF-IDF score to conditional probability values) are rather significant. Even if they only represent lexical co-occurrence of medical terms in literature, substantial information linking symptoms to diseases can be extracted from them if used properly.

Results appear to be good even for both algorithms. Considering our very limited frame of evaluation, it is a bit surprising that the more complex algorithm does not really seem to perform better. From a technical, but also medical point of view, this is good news since it might suggest that in our particular case it does not seem to be necessary to sacrifice accuracy for time. On top of that, the simple nature of the Bayesian Trees model allows it to perform very fast almost independently of the size of the knowledge base. This means that there is a lot of room for improving of the data itself without sacrificing too much speed, thus be-

coming an interesting default choice for potentially very different knowledge bases or domains.

On the other hand, the small observable differences on the results when changing patients' basic information shows that this extra layer of input data is not very relevant for the model's current state. It seems to be much more important to specify as much information about symptoms and signs as any other thing. In a real case scenario this is not always the case since a patient's basic information can tell a lot about the diseases it is at a higher risk of contracting, and which ones are more probable than others. Regarding this matter, we would like to have more disease and symptom restrictions in the database to account for these differences, but more importantly we would like to model risk factors in a smooth but robust manner that adequately reflects patients' pre-conditions.

Regarding the comparison between both experts criteria, we saw that there are various non-exclusive reasons behind their evident differences. These reasons include the clear contextual differences between Japanese and Chilean medical backgrounds, the natural discrepancies underlying medical judgment itself and other sources of error like misunderstanding of the processed data. Despite showing obvious differences for some cases, some others showed an interesting consistency. This exercise helps thus visualize at least to some small degree how some relevant diseases can manifest and be treated on the general population for very different countries, which acts accordingly to our intuition (Chapter 1). Also, both experts showed quite good results for the presented cases in the context of our evaluation, which is promising since the higher number of experts with a *good evaluation*, the closer we get to a good conclusion about the actual performance of the models.

Regarding the knowledge base a few things could also be said. First, the nature of medical data is rather heterogeneous among systems, patients, diseases, hospitals and many other factors. Categorization, *vectorization* or even simple processing of data is usually a very sensitive and slow process that requires a lot of manual revision. Data can be missing, wrong, incorrectly typed, duplicated, etc. If the data has problems, any system developed and put in production on top of it could be corrupted and potentially lead to *fatal* consequences. Other sources of knowledge like the ones obtained from internet are no different in this matter.

Usually not everything can be trusted completely so it is important to be careful with the numbers one is feeding to any algorithm. For this reason, we recommend that both computer and medical experts take a look at the data at different stages of the process as much as possible. In our case, even if the input from medical experts was very limited, an expert's revision of the data was fundamental and some key corrections in the structure of the database were made as a consequence. It is very likely that more changes would have been made had experts had unlimited availability.

We also recall on the usefulness of systems like the Unified Medical Language System that attempt to make dealing with medical information easier. Without a central source to connect and match different sources of data the task becomes immensely harder when the data comes from more than one different source. There are also many other attempts to organize medical information and huge improvements have been made. We predict that once medical data is more homogeneously organized and of easier access the machine learning techniques will be able to flourish more naturally, even on multi-classification problems like this one. An highly relevant example of an organization trying to structure medical data along a whole country is the Chilean *Centro Nacional en Sistemas de Información en Salud* (CENS) [54].

From a medical expert point of view, the results on the suggested exams are interesting because they not only show the best exams to perform, but also their expected outcome in some cases. This kind of visual aid in general is already very useful by itself since it can provide sometimes easily forgettable details about diagnoses and tests, thus improving the general efficiency of the decision making process for the emergency room.

Probably the most fundamental conclusion to be made is that Bayesian networks, or at least the *basic idea behind them* still appear to be a valid tool to be used as the ground for decision support systems, as long as the knowledge base is complete enough. But even if the knowledge base or data available to feed an artificially intelligent model is far from perfect it can still provide decision support as a visual aid at different stages of the medical procedure; not necessarily always for a final diagnosis, but also with middle steps like

orientation about exams. A model like the one developed tries to use the available data or knowledge as efficiently as possible to build a bridge between very simple input data, like age, sex and a few symptoms, and rather specific output data, like laboratory or image exams.

To finish, regarding the hypothesis stated in Chapter 1, we mentioned at the beginning of Chapter 4 that such a strong hypothesis could not be completely tested with the available resources, but rather a preliminary trial and the guidelines to do it could be stated. Even if the the hypothesis can not be officially confirmed nor denied, we think that with the work and analysis done, it can at least be pushed closer to the former. Trying to address a problem in the medical field in a very mathematical and prolific manner requires working with (or among) many experts even for smaller studies. Even if the evaluation becomes more difficult, a strong hypothesis can orient the work towards a more ambitious and clear goal and we believe it is worth taking the risk to study such a complex but important field (in a very serious and conscious way) despite not having all the resources available to return a strong conclusion. A preliminary study and its weaker conclusions could still be relevant and hopefully ground for future studies, related or not.

Regarding this, we believe that despite coming from non-strict settings, the study showed promising results. We also believe that it is possible to build an exam recommendation system based on Bayesian networks which could potentially improve the efficiency of the emergency room without requiring immense amounts of data and/or active input from several medical experts. Even the current version of the model and knowledge base could be an upgrade with respect to the default situation of an emergency room, as the one that was described in Chapter 1. Finally, we believe that this work represents a contribution because even though it does not lead to a scientifically backed conclusion, it indicates what else is required to achieve it and preliminary results indicate there is a high chance for them to be positive.

That being said, we believe that these results open the door to a whole set of new ideas on how the system can be improved:

First and most importantly, a more complete evaluation is required to draw conclusions

as strong as the hypothesis itself. We mentioned this at the beginning of Chapter 4, but we state it more formally here. A complete trial would require a few different medical experts from different relevant specialties including emergency room (hopefully with a lot of experience and research), to review different sources of official medical guidelines under very strict settings to make a 'laboratory-recreation' of the emergency room scenario, where many other variables and metrics are being measured. This, of course, would probably be very expensive and time-consuming but it would accomplish the goal of getting as close to the *truth* as possible.

Second, a better filtering of diseases and symptoms could be made with the help of more medical experts to reduce the number of diseases that are the same in an emergency context and also to be more specific with some symptoms that are too vague or simply missing. One important example that was not really addressed is the symptom (e.g. *Trauma*, which refers to accidents like broken ankles or even head injuries. These types of diagnoses are a big part of the emergency room and while they are mostly obvious at the time of arrival at the emergency room, having it actually implemented on the system would help to narrow hugely the number of possible diseases. In general, the symptoms part of the database is a little bit too small compared to the diseases part (contrary to what one could expect) and it does not really include important information like the specific body-part where the symptom is present. Making adjustments to the symptoms and the diseases part of the knowledge base seems to be a good first step towards improving the model.

Third, it would be interesting to incorporate in the model a way to measure diseases' risk level. Not only probability should be the only factor to order the output, but also their relevance in terms of prognosis and such. An idea would be to add a score from 1 to 5 (with help of a medical expert or something similar) to each diagnosis and weight the output probabilities with these numbers at inference time.

Fourth, it would be very positive to put additional information sources of findings, to make the knowledge base (after cleaning and restructuring) more complete and robust. What would be better, however, is to use more actual medical data to obtain information about

exams and findings in general related to diseases. More medical data from a real emergency room is more useful than spread medical information because it represents real life scenarios and important parameters could be extracted and introduced to the model.

Fifth, as mentioned multiple times, adding proper risk factors and additional disease restrictions is considered one of the top priorities for improvement of the model. The *soft approach evidence* model appears to be one way to do this [34]. We believe this is one of the main points to consider if the aim is to make results trustworthy and more personalized for different types of patients.

Sixth, to complement what was already stated about the evaluation, it would be wise to more extensively test with the BN2O model to completely understand the reasons for it performing similarly to the simpler Bayesian Trees model and not better. Perform tests with higher sampling sizes would be a good starting point to further evaluate the model and potentially confirm that the BN2O model is simply slightly sub-optimal with respect to the Bayesian Trees model.

Finally, and more on the usability side, since the real world scenario where the system is supposed to be used in are emergency rooms, it would be good to add a method to introduce symptoms one by one by asking the user for the next most probable symptom. The actual input requires the user to have knowledge on the entries of the database, but a practical application would, for example, first require input of one or two symptoms and then *ask* for the rest (just like a medical interview is) to complete the input data. This should not be so difficult if the most closely related symptoms within the database are being continuously suggested. Also, since the system is meant for use in Japanese hospitals, the data would need to be translated to the language for proper use. As for disease and symptoms, most of them have direct translations within the UMLS, but the rest of the findings would have to be manually translated.

In this work we discussed the theoretical grounds and later implemented a clinical decision support system oriented towards emergency rooms of Japanese hospitals. The goal

was to study different models that vary in speed and accuracy by evaluating them with help of at least a couple medical experts. To do this, we implemented two different models and algorithms based on Bayesian networks because of their extensive use and good results in CDSS literature and also based on the limited medical data. After a preliminary evaluation by the medical experts, results seem to be good in general for some of the most common and/or relevant emergency room diseases. In our particular case, the simplest, Bayesian Trees algorithm appeared to perform better in time and slightly better in accuracy. In the current state of the knowledge base, results does not seem to be very affected by differences in patient's basic information.

There is still a long way to go for clinical decision support systems since the era of Big Data is just beginning (especially for medical data). The next decades will most likely see a huge rise in projects involving medical data that aim to support, and even replace medical experts in some parts of their decision process. The economical and social improvements that properly used technologies on this line could have on some societies could be huge, but this is only possible if experts on both areas collaborate actively by sharing their knowledge and tools.

A project like the one developed is only a small example of what can be built with scattered sources of data and only limited input from medical experts. We believe that the system developed could potentially make at least some improvements in the emergency room process, even if in the form of visual aid or decision reinforcement.

Bibliography

[1] Stuart J Russell and Peter Norvig. *Artificial intelligence: a modern approach*. Malaysia; Pearson Education Limited,, 2016.

[2] Kevin Murphy. A brief introduction to graphical models and bayesian networks. 1998.

[3] Michael A Shwe, Blackford Middleton, DE Heckerman, Max Henrion, EJ Horvitz, HP Lehmann, and GF Cooper. Probabilistic diagnosis using a reformulation of the internist-1/qmr knowledge base. *Methods of information in Medicine*, 30(4):241–255, 1991.

[4] XueZhong Zhou, Jörg Menche, Albert-László Barabási, and Amitabh Sharma. Human symptoms–disease network. *Nature communications*, 5:4212, 2014.

[5] Olivier Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270, 2004.

[6] Randolph A Miller. Medical diagnostic decision support systems—past, present, and future: a threaded bibliography and brief commentary. *Journal of the American Medical Informatics Association*, 1(1):8–27, 1994.

[7] Scott DC Stern, Adam S Cifu, and Diane Altkorn. *Symptom to Diagnosis An Evidence Based Guide*. McGraw Hill Professional, 2014.

[8] Japan Healthcare Info. Japanese health insurance, 2011. URL `http://japanhealthinfo.com/japanese-healthcare-services/japanese-health-insurance/`.

[9] Michael J Ward, Heather Farley, Rahul K Khare, Erik Kulstad, Ryan L Mutter, Robert

Shesser, and Suzanne Stone-Griffith. Achieving efficiency in crowded emergency departments: a research agenda. *Academic Emergency Medicine*, 18(12):1303–1312, 2011.

[10] Ofir Ben-Assuli, Moshe Leshno, and Itamar Shabtai. Using electronic medical record systems for admission decisions in emergency departments: examining the crowdedness effect. *Journal of medical systems*, 36(6):3795–3803, 2012.

[11] Ivana Lapić and Dunja Rogić. Laboratory utilization in the emergency department–are the requested tests patient-oriented? *Signa Vitae*, 10(Suppl. 1):81–83, 2015.

[12] Robert W Derlet and John R Richards. Ten solutions for emergency department crowding. *Western Journal of Emergency Medicine*, 9(1):24, 2008.

[13] Jorge Quinteros, Nelson Baloian, Jose A Pino, Álvaro Riquelme, Sergio Peñafiel, Horacio Sanson, and Douglas Teoh. Diagnostic test suggestion via bayesian network of non-expert assisted knowledge base. In *Advanced Communication Technology (ICACT), 2018 20th International Conference on*, pages 340–346. IEEE, 2018.

[14] Jeremy Wyatt and David Spiegelhalter. Field trials of medical decision-aids: potential problems and solutions. In *Proceedings of the annual symposium on computer application in medical care*, page 3. American Medical Informatics Association, 1991.

[15] David Ferrucci, Eric Brown, Jennifer Chu-Carroll, James Fan, David Gondek, Aditya A Kalyanpur, Adam Lally, J William Murdock, Eric Nyberg, John Prager, et al. Building watson: An overview of the deepqa project. *AI magazine*, 31(3):59–79, 2010.

[16] RANDOLPH Miller, Fred E Masarie, and Jack D Myers. Quick medical reference (qmr) for diagnostic assistance. *MD computing: computers in medical practice*, 3(5):34–48, 1986.

[17] G Octo Barnett, James J Cimino, Jon A Hupp, and Edward P Hoffer. Dxplain: an evolving diagnostic decision-support system. *Jama*, 258(1):67–74, 1987.

[18] Elie Sanchez. *Solutions in composite fuzzy relation equations: application to medical diagnosis in Brouwerian logic*. Faculté de Médecine de Marseille, 1977.

[19] Bruce G Buchanan, Edward Hance Shortliffe, et al. *Rule-based expert systems*, volume 3. Addison-Wesley Reading, MA, 1984.

[20] Mu-Chun Su. Use of neural networks as medical diagnosis expert systems. *Computers in Biology and Medicine*, 24(6):419–429, 1994.

[21] Igor Kononenko. Inductive and bayesian learning in medical diagnosis. *Applied Artificial Intelligence an International Journal*, 7(4):317–337, 1993.

[22] Yves Coppieters and Alain Levêque. Ethics, privacy and the legal framework governing medical data: opportunities or threats for biomedical and public health research? *Archives of public health*, 71(1):15, 2013.

[23] Bruce Reiner. Strategies for medical data extraction and presentation part 1: Current limitations and deficiencies. *Journal of digital imaging*, 28(2):123–126, 2015.

[24] Jay Liebowitz. *The handbook of applied expert systems*. cRc Press, 1997.

[25] Judea Pearl. Bayesian networks: A model of self-activated memory for evidential reasoning. In *Proceedings of the 7th Conference of the Cognitive Science Society, 1985*, pages 329–334, 1985.

[26] Adam Zagorecki, Piotr Orzechowski, and Katarzyna Holownia. A system for automated general medical diagnosis using bayesian networks. In *MedInfo*, pages 461–465, 2013.

[27] David Heckerman, Dan Geiger, and David M Chickering. Learning bayesian networks: The combination of knowledge and statistical data. *Machine learning*, 20(3):197–243, 1995.

[28] Judea Pearl. Probabilistic reasoning in intelligent systems: Networks of plausible inference. 1988.

[29] Luis M de Campos, Juan M Fernández-Luna, and Juan F Huete. Bayesian networks and information retrieval: an introduction to the special issue. *Information processing & management*, 40(5):727–733, 2004.

[30] Ibsen Chivatá Cárdenas, Saad SH Al-jibouri, Johannes IM Halman, and Frits A van Tol. Capturing and integrating knowledge for managing risks in tunnel works. *Risk Analysis*, 33(1):92–108, 2013.

[31] Anthony C Constantinou, Norman E Fenton, and Martin Neil. pi-football: A bayesian network model for forecasting association football match outcomes. *Knowledge-Based Systems*, 36:322–339, 2012.

[32] K Wojtek Przytula and Don Thompson. Construction of bayesian networks for diagnostics. In *Aerospace Conference Proceedings, 2000 IEEE*, volume 5, pages 193–200. IEEE, 2000.

[33] Olivier Pourret, Patrick Naïm, and Bruce Marcot. *Bayesian networks: a practical guide to applications*, volume 73. John Wiley & Sons, 2008.

[34] Bruce D'Ambrosio. Symbolic probabilistic inference in large bn20 networks. In *Proceedings of the Tenth international conference on Uncertainty in artificial intelligence*, pages 128–135. Morgan Kaufmann Publishers Inc., 1994.

[35] Robert S Ledley and Lee B Lusted. Reasoning foundations of medical diagnosis. *Science*, 130(3366):9–21, 1959.

[36] Anthony K Akobeng. Understanding diagnostic tests 1: sensitivity, specificity and predictive values. *Acta paediatrica*, 96(3):338–341, 2007.

[37] Rajul Parikh, Annie Mathai, Shefali Parikh, G Chandra Sekhar, and Ravi Thomas. Understanding and using sensitivity, specificity and predictive values. *Indian journal of ophthalmology*, 56(1):45, 2008.

[38] Daniel Nikovski. Constructing bayesian networks for medical diagnosis from incomplete and partially correct statistics. *IEEE Transactions on Knowledge and Data Engineering*, 12(4):509–516, 2000.

[39] Max Henrion and Marek J Druzdzel. Qualitative propagation and scenario-based explanation of probabilistic reasoning. *arXiv preprint arXiv:1304.1082*, 2013.

[40] Sampath Srinivas. A generalization of the noisy-or model. In *Uncertainty in Artificial Intelligence, 1993*, pages 208–215. Elsevier, 1993.

[41] Irina Rish and Rina Dechter. On the impact of causal independence. Technical report, Technical report, Dept. Information and Computer Science, UCI, 1998.

[42] Max Henrion. Search-based methods to bound diagnostic probabilities in very large belief nets. In *Uncertainty Proceedings 1991*, pages 142–150. Elsevier, 1991.

[43] Umls metathesaurus - qmr (quick medical reference) - metadata. URL `https://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/QMR/metadata.html`.

[44] A database of sensitivity and specificity. URL `http://getthediagnosis.org/`.

[45] Timothy W. Jolis. Sensitivity and specificity. URL `http://www.sensitivityspecificity.com/`.

[46] Carolyn E Lipscomb. Medical subject headings (mesh). *Bulletin of the Medical Library Association*, 88(3):265, 2000.

[47] Karen Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21, 1972.

[48] World Health Organization. *International Classification of Functioning, Disability and Health: ICF*. World Health Organization, 2001.

[49] Donald AB Lindberg, Betsy L Humphreys, Alexa T McCray, et al. The unified medical language system. *IMIA Yearbook*, pages 41–51, 1993.

[50] Snomed ct to icd-10-cm map. URL `https://www.nlm.nih.gov/research/umls/mapping_projects/snomedct_to_icd10cm.html`.

[51] Essential evidence plus. URL `https://www.essentialevidenceplus.com/content/eee`.

[52] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967.

[53] Laso Guzmán, F JavierCordero Sánchez, Miguel Fuertes Martín, Aurelio Pastor Encinas, Isabel J Pérez Arellano, José Luis Ruiz, and Carlos MartínF Javier Laso Guzmán. *Diagnóstico diferencial en medicina interna*. Elsevier,, 2005.

[54] Cens – centro nacional en sistemas de información en salud. URL `http://cens.cl/wp/`.