TÉCNICO LISBOA

Multiple criteria decision analysis for biomarker prioritization:
Developing a socio-technical approach to assist researchers and clinicians in biomarker selection for validation and translation into clinical applications

Beatriz Ribeiro Norte

Thesis to obtain the Master of Science Degree in

Biomedical Engineering

Supervisors: Professor Mónica Duarte Correia de Oliveira
Doctor Deborah Penque

Examination Committee

Chairperson: Professor Paulo Rui Alves Fernandes
Supervisor: Professor Mónica Duarte Correia de Oliveira
Member of the Committee: Professor Klára Dimitrovová

October 2019

Declaration

I declare that this document is an original work of my own authorship and that it fulfills all the requirements of the Code of Conduct and Good Practices of the Universidade de Lisboa.

Preface

The work presented in this thesis was performed at Centro de Estudos de Gestão of Instituto Superior Técnico (Lisbon, Portugal), during the period February-October 2019, being supervised by Professor Mónica Duarte Correia de Oliveira and co-supervised by Doctor Deborah Penque.

Acknowledgments

I would like to thank my supervisors Professor Mónica Oliveira and Doctor Deborah Penque, as well as Professor Ana Vieira, for all their support and help throughout the development of my master thesis.

I would also like to acknowledge all the experts that gave up some of their time to participate in the several steps of my thesis.

I am grateful to my two sisters and all my friends for the support, good times, and for helping me stay positive throughout my life.

A very special word of gratitude goes to my boyfriend, Daniel, for all the positive energy, motivation and help keeping me focused on my objectives.

Lastly, the biggest thanks of all goes to my parents, specially my mother, for all the love, the help and the motivation, for supporting my choices, both in education and in my personal life, and for giving me the chance of having a great student life, one that I can be proud of. Thank you!

Abstract

Although thousands of biomarkers have already been discovered, not many have been applied into clinical practice, due to the high time and monetary resources necessary for their identification and validation. Due to its complexity and heterogeneity, and consequent lack of specific biomarkers, Chronic Obstructive Pulmonary Disease (COPD) is ideal to apply the model developed in this thesis.

With the purpose of evaluating and selecting the most promising COPD prognostic biomarkers among those found in literature, a multiple-criteria decision analysis (MCDA) model was developed, with focus on the structuring phase. To meet this goal, a socio-technical approach based on the MACBETH method was followed, including an assessment of areas of concern in the biomarker field, the definition of criteria for biomarker selection and respective descriptors of performance and reference levels (using literature, interviews and a Web-Delphi), testing for preference dependence and the design of the resultant value tree.

Results show that experts believe that the clinical relevance, clinical added value, quality of studies and test reliability are relevant dimensions, while the costs of development, the patient comfort, the easiness to measure and analyse a biomarker and the easiness to interpret the results are secondary, as one must be available to give up certain benefits if the biomarker significantly improves the patients' health/well-being. Although ten evaluation dimensions were considered relevant by experts, some dependencies were found, namely between the evaluation dimensions clinical relevance and clinical added value, leading to the grouping and reformulation of the dependent dimensions, resulting in independent evaluation criteria.

Overall, and despite some difficulties, the approach applied in this thesis worked well, resulting in a good structure for the MCDA model, with seven well defined and relevant evaluation criteria for the prioritization of COPD prognostic biomarkers. In the future, it would be interesting to complete the model, including the building, testing and validation phases of the model.

Keywords: Prioritization, biomarkers, MCDA, MACBETH, Delphi, COPD, preference dependence

Resumo

Apesar de milhares de biomarcadores terem já sido descobertos, não há muitos que tenham sido aplicados à prática clínica, devido à elevada quantidade de tempo e recursos monetários necessária para a sua identificação e validação. Devido à sua complexidade e heterogeneidade, e comsequente falta de biomarcadores específicos, a Doença Pulmonar Obstrutiva Crónica (DPOC) é um exemplo ideal para aplicar o model desenvolvido nesta tese.

Com o propósito de avaliar e selecionar os biomarcadores de prognóstico de DPOC mais promissores de entre todos os encontrados na literatura, um modelo multicritério de apoio à decisão foi desenvolvido, com foco na estruturação do modelo. Para ir de encontro a este objetivo, uma abordagem sócio-técnica baseada no método MACBETH foi seguida, incluindo o levantamento das áreas de preocupação no campo dos biomarcadores, a definição de critérios para a seleção de biomarcadores e respetivos descritores de performance e níveis de referêncoa (baseado em literatura, entrevistas e num Web-Delphi), teste de dependência de preferência e o desenho da árvore de valor resultante.

Os resultados mostram que os peritos acreditam que a relevância clínica, o valor clínico acrescentado, a qualidade dos estudos e a fiabilidade do teste são dimensões relevantes a ser consideradas aquando da seleção de um biomarcador, enquanto os custos de desenvolvimento, o conforto do paciente, a facilidade de medir e analisar um biomarcador e a facilidade de interpretar os resultados são secundários, uma vez que uma pessoa deve estar disponível para abdicar de certos benefícios se o biomarcador melhorar significativamente a saúde/bem-estar dos pacientes. Apesar de dez dimensões de avaliação terem sido consideradas relevantes pelos peritos, algumas dependências foram encontradas, nomeadamente entre as dimensões de avaliação relevância clínica e valor clínico acrescentado, o que levou ao agrupamento e reformulação das dimensões dependentes, resultando em critérios de avaliação independentes.

De um modo geral, e apesar de algumas dificuldades, a abordagem sócio-técnica aplicada nesta tese funcionou bem, resultando numa boa estrutura para o modelo multicritério de apoio à decisão, com sete critérios de avaliação bem definidos e relevantes para a prioritização de biomarcadores de prognóstico de DPOC. No futuro, seria interessante completar o modelo, incluindo as fases de construção, teste e validação do modelo.

Keywords: Prioritização, biomarcadores, modelo multicritério de apoio à decisão, MACBETH, Delphi, DPOC, dependência de preferência

Contents

List of Tables

List of Figures

Acronyms

AECOPD Acute Exacerbations of Chronic Obstructive Pulmonary Disease. 38, 39

AHP Analytic Hierarchy Process. 14, 15

AUC Area Under the Curve. 8, 43, 44, 46, 47, 50, 61

BBP Best Biomarker Practice. 1

COPD Chronic Obstructive Pulmonary Disease. 1–3, 6, 10, 11, 13, 17, 25–29, 31, 38, 39, 41, 55, 57, 63–65, 67, 68

COU Context of Use. 2, 37, 44, 46, 47, 50, 61, 68

ECM Extracelular Matrix. 26

ED Evaluation Dimension. 22–24, 30, 31, 38, 40–43, 46, 48, 49, 54, 56–60

FDA Food and Drug Administration. 6, 25, 27, 28, 37

FER Forced Expiratory Ratio. 27

FEV_1 Forced Expiratory Volume in 1 Second. 27, 28, 39

FPR False Positive Rate. 44

FPV Fundamental Point of View. 32

FVC Forced Vital Capacity. 27

GOLD Global Initiative for Chronic Obstructive Lung Disease. 27

HTA Health Technology Assessment. 14, 65

LR Likelihood Ratio. 8

MACBETH Measuring Attractiveness by a Categorical Based Evaluation Technique. 16, 17, 23, 28, 29, 32, 35, 63–65

MAVT Multi-Attribute Value Theory. 15

MCDA Multiple Criteria Decision Analysis. 1–3, 10, 11, 13–17, 30, 32, 38, 63–65, 67, 68

PV Point of View. 32

QALY Quality Adjusted Life Year. 7

ROC Receiver Operating Characteristics. 8, 44

ROS Reactive Oxygen Species. 26

TPR True Positive Rate. 44

WHO World Health Organization. 11

Chapter 1

Introduction

1.1 Motivation and Proposed Solution

Thousands of proteins have already been proven to be hallmarks of emerging disease, prognosis of a patient or response to treatment. The identification of protein biomarkers associated with a disease is essential for the improvement of personalized medicine based on blood tests, since they have a great impact in drug discovery and development, possibly leading to better disease outcomes, such as higher patient survival and lower healthcare costs, among others. However, even though several biomarkers have been discovered and presented in studies, there are not many that have been applied into clinical practice (only about one hundred and fifty out of thousands of identified biomarkers), due to the fact that the process of identification and validation of disease specific biomarkers is very time consuming and requires many resources, as well as to the fact that some studies that determine the clinical value of a biomarker are not reproducible and that some studies do not match regarding requirements for regulatory and marker approval [1].

In order to increase the number of clinically validated biomarkers, instead of increasing the number of studies that discover new ones, the Cost Action European project, CliniMark, aims to improve the quality and reproducibility of studies and to create the Best Biomarker Practice (BBP) guidelines, in order to establish a coherent biomarker development pipeline from discovery to clinical application, which shall provide guidance to [1]:

- Classify biomarkers considering their characteristics, predicted clinical use and phase of development;
- Select and validate appropriate research-grade biomarker detection tests;
- Select studies and biological samples that have been designed appropriately and that can be reproduced, to reliably validate biomarkers on a clinical level;
- Select and report on appropriate clinical data storage, biomarker data storage, data analysis protocols, privacy concerns, ethical issues, and statistical analysis methods.

To demonstrate this project, which is also being developed with the MEDI-VALUE research project, which develops collaborative approaches in Multiple Criteria Decision Analysis (MCDA) to evaluate medical devices, Chronic Obstructive Pulmonary Disease (COPD) will be used as an example. Due to its complexity and heterogeneity, as well as to its molecular and clinical characteristics, there is great difficulty to efficiently stratify the patients and to introduce personalized therapeutic approaches, which, together with the fact that the currently available clinical tools do not predict the progression and exacerbations of this disease efficiently [1], makes COPD a great candidate to demonstrate this project with,

due to its clear need for new drugs that can solve some of its major clinical problems: and it all starts with biomarker prioritization.

With all this in mind, an extensive literature survey about COPD biomarkers has been conducted, considering all literature published between 2016 and 2018, plus the biomarkers and articles mentioned in S. Ongay et al. [2], resulting in a list of about one hundred candidate biomarkers for COPD. There are several proteins in this list that showed interesting measurement values, with distinct statistical power, as well as interesting outcomes, with high relevance to COPD, when investigated by different research groups. Some proteins had different applications suggested by different authors, likely due to the objective of each study, and there were several proteins that showed potential for more than one outcome, such as diagnostic and prognostic, or Context of Use (COU). The same way, one specific prognosis, such as mortality or exacerbation, can have several biomarkers associated with distinct biological functions. However, it is very likely that not all selected biomarkers will be relevant in drug development in COPD.

For this reason, it is now necessary to evaluate all the biomarker candidates and select the most promising ones, to be further analysed and tested. MCDA methods and tools have been identified as a good option to reach this goal, by helping researchers in optimal decision making and involving the knowledge of different stakeholders (researchers and clinicians), while building an evaluation model. Therefore, the proposed solution to the problem above, and main goal of this thesis, is to create and test a MCDA approach, with focus on the structuring of the model, to help experts in COPD biomarker prioritization, but that can be adapted and used for any given disease, and not only COPD, which will be used solely as an example to apply the created model and to demonstrate and evaluate its results. Furthermore, by facilitating the biomarker prioritization process, and by selecting the most promising biomarkers, this approach will help reducing significantly both time and money resources associated with the process of identification and validation of disease specific biomarkers.

1.2 Document Structure

The remaining of the document is organized as follows:

- Chapter 2 presents all the fundamental concepts necessary to understand this thesis, including an overview of personalized medicine, biomarkers and COPD.
- Chapter 3 includes a literature review regarding MCDA and participatory methods for biomarker prioritization and regarding COPD.
- Chapter 4 describes the socio-technical methodology for biomarker prioritization to be used in this thesis.
- Chapter 5 presents the results of the application of the methodology to the case of COPD, followed by observations, drawn conclusions and discussion of results.
- Chapter 6 presents a brief discussion of the thesis as a whole.
- Chapter 7 includes the conclusion drawn from this master thesis, as well as outlines for future work.

Chapter 2

Context

In this chapter, the fundamental concepts to understand this thesis will be described.

First, an overview of personalized medicine will be presented, including its characteristics and advantages in clinical care, as well as its connection with biomarkers and biomarker prioritization. Secondly, this chapter will cover biomarkers: more specifically, their main characteristics, their cycle of discovery and development, their connection with drug development, economic evaluation associated with biomarkers and the key points associated with biomarker prioritization, namely why it is relevant and necessary and what would make a biomarker ideal. Lastly, COPD will be presented, as it will be used in this thesis as an example disease to demonstrate the application and results of the MCDA model that will be created.

2.1 Personalized Medicine

Personalized medicine, also known as precision medicine or individualized medicine, consists of treatment directed to the need of individual patients, considering biomarker, genetic, phenotypic or psychosocial characteristics that distinguish a certain patient from others with similar clinical presentations, classifying them according to their probable disease risk, prognosis and/or response to treatment [3][4]. Clinical phenotyping allows for patient classification into specific subgroups, based on observable characteristics and properties of an organism, while endotyping uses the presence of a biological mechanism to define a patient subgroup [3]. These subgroups provide information regarding prognostic and allow to determine more appropriate therapies for the patients [5].

The main goal of personalized medicine is to improve individual treatment, including clinical outcomes and minimization of side-effects for those patients likely to respond to a given treatment [5]. The fact that it enables risk assessments, earlier diagnosis and optimal treatments, leads to a better health care, as well as lower associated costs [6].

Over the last few decades, a lot of evidence has emerged, showing that a significant part of variability in drug response is genetically determined, depending on age, health status, nutrition, epigenetic factors, environmental exposure and concurrent therapy. Personalized medicine allows for the development of new agents and drugs specifically for patients that do not respond well, or as predicted, to medications [6].

As represented by the different colors in Figure 2.1, people are different and have different genetic characteristics and, as a consequence, can present different outcomes when subjected to the same drug, some benefiting from it, while others suffer adverse effects or no effect at all. However, when using personalized medicine, the drug that each patient is given takes into account their personal character-

istics. This information is obtained through several tests, which use biomarkers, and allow for patients to be separated in different groups according to their characteristics, and to receive the right medicine for them (either a different drug or a different dose). This way, each patient benefits from individualized treatment, leading to a higher benefit rate and less adverse effects [7].

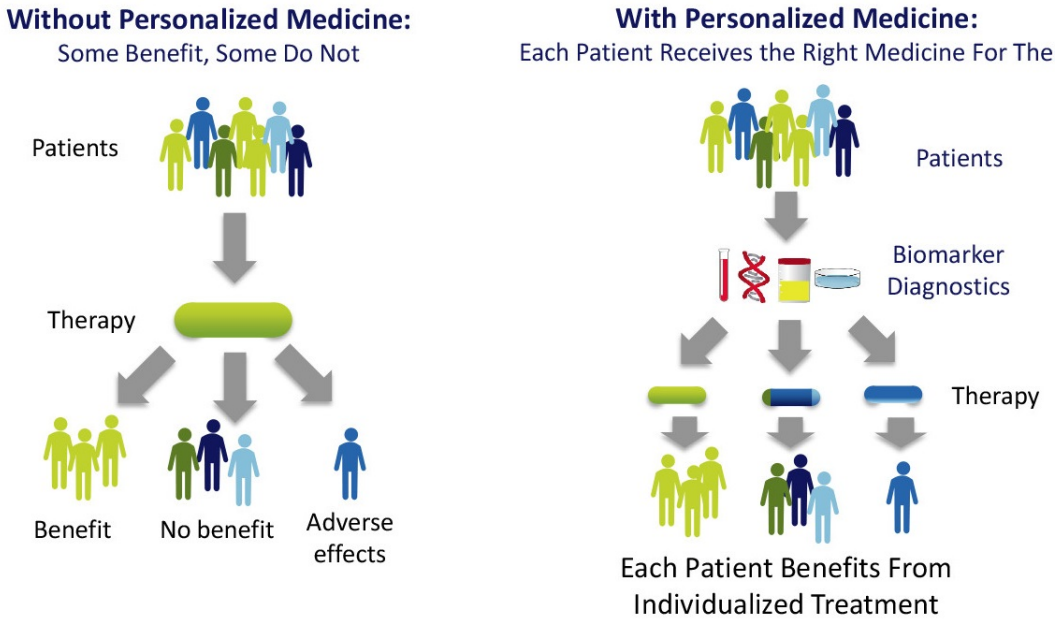

Figure 2.1: Treatment outcome with and without personalized medicine (source: [7]).

In sum, personalized medicine can offer: better medication selection, by predicting which patients are likely to respond successfully to a certain drug and which are not; safer dosing options, by predicting the optimal dose to use for each patient; and improvements in drug developments, by targeting specific population, with specific genetic characteristics, that can be helped by a certain medication, speeding up the clinical trial process and avoiding the exclusion of certain biomarkers or drugs for not being effective for the majority of the population, only helping (sometimes very successfully) individuals with specific characteristics. Also, by eliminating unnecessary treatments that would be ineffective or dangerous, personalized medicine leads to substantial drug cost savings for patients, as well as for the entire health care industry [6].

However, for personalized medicine to be efficiently applied in a certain disease, it is essential to have the best biomarkers of treatment response and prognosis in individual patients for that disease [8], hence the importance of biomarker prioritization, which will be further explored along the chapter. Only then will it be possible to reach the best outcomes, with the highest benefits, improving the lives of all those affected [8].

2.2 Biomarkers

According to the National Institutes of Health Biomarkers Definitions Working Group, a biomarker is a "characteristic that is objectively measured and evaluated as an indicator of normal biologic processes, pathogenic processes or pharmacologic responses to a therapeutic intervention" [9]. Recently, as a result of the advancement in personalized medicine, a biomarker can also be perceived as an indicator that enables the development of treatment interventions for specific patients, maximizing therapeutic benefits and minimizing the risk of treatment [10]. In general, a biomarker should be associated with the

biological mechanisms involved in a disease or its treatment and be able to correlate statistically with clinical outcomes [11].

2.2.1 Cycle of Biomarker Development

The cycle of biomarker discovery and development is represented in Figure 2.2. In the beginning of this cycle, there is an enormous number of candidate biomarkers and no analyzed samples. However, as we advance in the pipeline, the number of analyzed samples increases significantly, while the number of biomarker candidates decreases to a very low number, because there are few biomarkers that have all the necessary characteristics to be qualified or even accepted in a drug study. The main objectives and characteristics of each of the six phases of the biomarker discovery and development pipeline are described below.

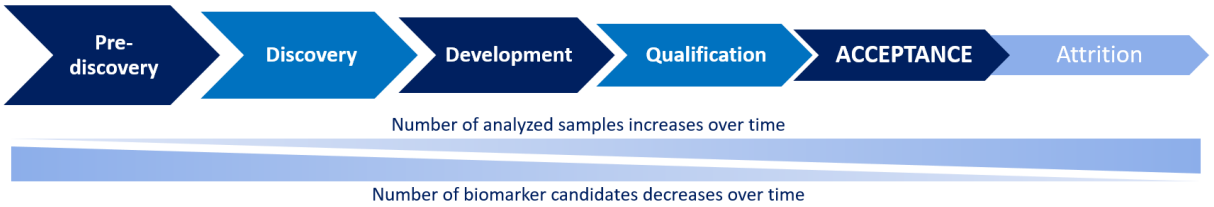

Figure 2.2: Biomarker discovery and development pipeline.

1. Pre-discovery: The first phase of the biomarker discovery and development pipeline is the pre-discovery, being characterized by the articulation of the clinical goals of the biomarker, the choice of the clinical endpoint of interest (e.g. mortality, exacerbation) and the development of a health economics framework, in order to assess the biomarker [10].

2. Discovery: The second phase, which usually lasts between 1 and 3 years, focus on the discovery of biomarker candidates, both on clinical and nonclinical studies, including DNA variations, transcriptome, proteome, phosphoproteome, metabolome, circulating cells, histology features, imaging properties and physiological measurements. This phase also involves the identification of biological samples (DNA, plasma, serum, etc.) and platform (targeted or untargeted) to be used, as well as independent cohorts for replication. Ideally, results obtained in the discovery phase should be replicated in multiple other cohorts, in order to ensure stability and generalizability of data [10][12].

3. Development: The development phase, which usually lasts between 3 and 5 years, is divided in two parts [10][12]:

- Analytical Validation: The first part is the analytical validation, being characterized by the development and utilization of clinically appropriate platforms and the evaluation of analytical performance characteristics following clinical guidelines. The results must show generalizability across different samples, as well as reproducibility and standardization of the assay;
- Clinical or Nonclinical Utility Studies: The second part consists on an evaluation of clinical/non-clinical utility. The results must show well-designed experiments, added value in research models and/or patients, such as improved health outcomes, and performance characteristics, indicating, for example, if the biomarker is cost-effective.

4. Qualification: The fourth phase, which usually lasts more than a year, is the phase where the biomarker is tried for qualification. For this to happen, it is necessary for there to be an agreement about intended context of use by multiple stakeholders, coordination of efforts toward data generation and strong evidence for stated context of use. If the biomarker is qualified, it will be allowed acceptance in future drug programs without the need of a new review [12].

5. Acceptance: The fifth phase is about acceptance in a drug program, which depends on the quality of evidence showing utility, the reproducibility in well-designed studies, the technical abilities of outside groups and the dissemination of data. It is relevant to notice that, in certain cases, a biomarker without qualification can be given regulatory acceptance in a drug program, but it is necessary to provide Food and Drug Administration (FDA) with evidence of analytical validity and clinical/nonclinical utility of the biomarker. In these cases, since the biomarker is not qualified, it cannot be used in other drug programs without re-review [12].

6. Attrition: The sixth and final phase is the attrition phase, which may occur if the performance of the biomarker was misjudged, which could happen due to new evidence of limited efficacy or data showing possible harm [12].

The whole process of bringing a biomarker from development to qualification requires a large investment, both in time and money. In fact, it is estimated to cost between 800 million and 1.7 billion dollars, requiring between 7 and 12 years. Therefore, the best long-term strategy to reach qualification involves industry, academic labs and disease foundations working together, and sharing goals, responsibilities and risks [11][12].

Nowadays, in the era of personalized medicine, many research efforts have been directed to the identification of novel biomarkers. However, among all the biomarkers that enter the biomarker discovery and development pipeline, only a small percentage reaches the end, being successfully translated from scientific discovery to clinical application. This fact results in a loss in health potential for both patients and society, as well as in wasted resources from public and private investors, used during research, development and evaluation of the biomarkers [13].

In this moment, and for the purpose of this thesis, we find ourselves almost in the end of the discovery phase of COPD biomarkers, preparing to enter the development one. However, before it is possible, it is necessary to narrow down the number of candidate biomarkers, in order to start the validation and translation into clinical applications process with the best possible candidates, so the investment, both in time and money, can be minimal, while achieving the best possible results.

2.2.2 Biomarkers in Drug Development

Biomarkers are used in all stages of drug development, being a valuable tool to create safe and efficacious drugs, by providing greater accuracy or more complete information regarding disease progression or drug performance [11]. They can be classified as one, or several, of the following categories, depending on their use and clinical context: diagnostic, prognostic, predictive, monitoring, response, safety or susceptibility biomarker [12].

In recent years, one of the main problems scientists have been facing in drug development is finding biomarkers that are predictive, or ”translate well”, from animal or in silico models to humans, because sometimes a biomarker can present significant values in an animal and not present them in a human. This lack of correlation may happen due to several reasons: the biomarker may give indications about the disease but may not reflect clinically important treatment effects; the therapy may impact the biomarker chosen, but the parameter may be irrelevant to the disease; and, finally, a drug can work in humans through many different pathways, which may not be reflected in animal models [11].

As for the economics, they have become a bigger and bigger part of drug development as time passed. In fact, the whole process of drug development has experienced a significant growth in cost, while the number of drugs submitted and approved per year has been decreasing. There are several causes for such a decline, including the complexity of new disease targets, diseases with multiple factors, more regulatory requirements and higher expectations related to drug safety. Inefficiencies in the pro-

cess, as well as marketing expenses, are also associated with the increase in drug development costs. The marketplace also has a significant impact on whether a new drug will be distributed to the patients, even if approved, because if a compound is not sufficiently different from other approved molecules, regarding safety, efficacy and/or cost, it is not likely that it will be supported by large insurers or that it will become an approved course of treatment in countries were health care systems are publicly funded [11]. Therefore, it is highly important to perform an economic evaluation of the biomarkers during the process of drug development.

2.2.3 Economic Evaluation of Biomarkers

In order to prioritize between competing innovations, and before making payment decisions regarding the development of new drugs and expensive interventions, decision makers need information regarding the health economic impact of interventions. Thus, economic studies need to be conducted [14][15].

To avoid elevated medical costs, it is natural that new biomarkers need to be examined with respect to their economic benefits, additionally to their clinical utility. However, biomarkers will always need to start by demonstrating their clinical usefulness, because a new biomarker will not make economic sense unless it is clinically useful, both in terms of patient care and clinical decisions [14].

From the moment the diagnostic accuracy and potential clinical usefulness is established, a biomarker may be subjected to several types of economic studies [14]:

- **Cost minimization:** Cost minimization studies are the simplest of the four, consisting on the determination of the intervention/test that, compared to others that produce the same outcome, is the cheapest (better relation price-outcome).
- **Cost-effectiveness analysis:** These studies have the purpose of determining the most efficient way to use a fixed set of resources, in order to obtain the best possible effect, which is usually a natural unit, such as a life year, but can also be numbers of clinical events (for example, number of strokes prevented).
- **Cost-benefit analysis:** In these studies, the costs of the test/intervention are compared with the costs of the benefit. However, assigning a monetary value to the benefit is not easy, requiring equating a monetary value to a year of life.
- **Cost-utility analysis:** Cost-utility studies are the most commonly used, and estimate the ratio between the cost of an intervention/test and its benefit in terms of years gained in full health, which is described using money per Quality Adjusted Life Year (QALY) gained. For a biomarker to be cost-effective, it must be associated with a maximum of USD 50.000 per QALY gained.

To determine the usefulness of a new laboratory test, including the usefulness of a biomarker, it is necessary to go through four different levels of studies/evidence that have to be examined before the biomarker can be adopted for clinical practice. The first one includes technical and analytic issues, from bias to imprecision and analytical measurement interval, among others, with the purpose of determining the reproducibility and accuracy of the results. The second one has the purpose of determining if the biologic factors that may affect a test result are relevant in that case: for example, if the blood concentrations of a biomarker depend on factors such as diet and exercise, with no relation to the disease, then it is likely that the biomarker will not be useful. If these first two levels are not cleared, then the biomarker makes no economic sense. The third level is all about diagnostic accuracy, including diagnostic sensitivity and specificity, likelihood ratios and predictive values. The fourth and final level regards usefulness/utility of the biomarker. Based on all the information gathered, the biomarker will

receive a grade from A (high certainty of substantial benefit) to D (zero or negative benefit). Thus, if the biomarker receives a mark lower than B, it is unlikely that it will make economic sense or that it will be recommended and adopted [14].

The only problem is that, in certain cases, it is very challenging to evaluate the cost effectiveness of biomarkers. For example, in the case of diagnostic biomarkers, it is difficult because diagnostics themselves do not have a direct influence over long-term outcomes, but rather impact the subsequent care process. Therefore, it is necessary to take under consideration the accuracy of the diagnostic test, the impact of the diagnostic on subsequent therapeutic decisions and the effectiveness of those therapies, which makes the whole process significantly more challenging [15].

All things considered, an economic evaluation is clearly critical in the development of a biomarker, and should be taken under consideration when selecting the most promising biomarkers for a certain disease, in order to avoid wasting both time and money with a biomarker that makes no economical sense and that, consequently, will likely have no future.

2.2.4 Added Value of a Biomarker and the Benefit-Cost-Risk Triangle

Throughout the years, the advances in proteomics, genomics and biotechnologies have led to the generation of multiple biomarkers that present clinical value in diagnosis, prognosis, prediction of risk and therapy response, among others [16].

However, questions about their usefulness have emerged, since not all these new biomarkers add value to the ones that already exist [16][17]. Therefore, before biomarker prioritization, it is crucial to determine if a set of new biomarkers has added value or not, which can be done through the building of two nested classification models, used to combine multiple biomarkers to predict an outcome: the first one being a partial model containing the existing biomarkers, while the second one is a full model where the new biomarkers are combined with the existing ones. This can be done using several different methods, including the Likelihood Ratio (LR) or the Wald Test, in order to determine the statistical significance of the new biomarker, or the area under the Receiver Operating Characteristics (ROC) curve, also known as Area Under the Curve (AUC), in order to compare the diagnostic performance (in terms of a diagnostic accuracy metric) of the two models [16].

After a biomarker has been considered to add value, it is also highly relevant to determine the benefits, costs and risks associated with it.

Biomarkers are of great help in healthcare, and can have several benefits, depending on their characteristics, including the ability to make a diagnosis or prognosis or to predict the outcome of a treatment, among others [18]. However, there are also costs and risks associated, that need to be taken under consideration.

The costs associated with the development of a new biomarker always include time and resources. Additionally, it can include costs associated with specific and unpredictable problems, such as laboratory errors. As mentioned in the previous section, it is essential to go through an economic evaluation of all biomarkers during the process of drug development, in order to assess the total costs associated with each one. As for risks, the simple fact that there is always an elevated cost associated with the development of a biomarker, which can end up being useless, having no relevant clinical application, is a very high risk. Also, there is always the risk that a drug, which uses a specific biomarker, may not be accepted by the public, either due to the price or due to the existence of similar products, with similar purposes and outcomes, more well known. Thus, for all this reasons, the benefit-cost-risk triangle is an issue that should be addressed in every biomarker development process, before selecting the most promising biomarkers, which should favor the benefit end of the triangle [18].

2.2.5 Biomarker Prioritization

When compared to standard non-biomarker guided approaches, a biomarker-guided approach leads to better health outcomes and a better experience for patients, either if the situation is ideal and the costs are lower, or even if the costs are slightly increased, in which case the benefits of the biomarker outweigh the costs. However, when biomarkers do not modify disease management, they are usually not very useful on a clinical level, due to the fact that they have little to no impact on patient-related health outcomes and tend not to be cost-effective [10]. Therefore, it is highly relevant to efficiently select the most promising biomarkers for a certain disease, hence the concept of biomarker prioritization, even though there are many challenges in biomarker prioritization and evaluation, including the large number of potential biomarkers, the distinct utilization contexts, the fact that they are used by different types of stakeholders, the variable levels of evidence, the little evidence and synthesis of evidence on biomarkers impacts, the uncertainty regarding impacts, biomarkers impacting one of several diseases, biomarkers potentially changing patient pathways, high costs of biomarker validation, absence of objective prioritization criteria and lack of consensus regarding the relevant benefits, risks and costs to be considered in biomarker evaluation.

An ideal biomarker, which would be the goal biomarker for any disease, is a biomarker that presents the following characteristics [10][19]:

- It should be safe, accurate, sensitive, not expensive and easy to measure;
- It should be visible before histopathological changes, but also present changes that give information after active damage (modifiable with effective therapy);
- It should present a correlation with the severity of the damage/disease;
- It should be easily accessible in the peripheral tissue, namely in the blood or urine, with minimal discomfort to the patient;
- It should be associated with a known mechanism, allowing full comprehension of the biomarker and its biomolecular characteristics;
- It should be able to indicate the location of the damage;
- Its results should be reproducible across age, gender and different racial and ethnic backgrounds;
- It should guide clinicians to intervene with more effective therapies for patients who need them, eliminating the use of ineffective or harmful ones.

Since there are several characteristics that a biomarker must have in order to be considered ideal, it is very likely that there is no such biomarker, meaning that it will probably be necessary to consider a group of biomarkers, instead of just one, to characterize a certain disease [19].

The number of biomarkers that enters the biomarker discovery and development pipeline during discovery, with the aim of reaching a clinical application, is too high compared to the resources available to do so. Consequently, it is essential to identify the most promising candidates, with the highest chance to succeed as a commercial product, which requires an early estimate of their potential clinical impact, commercial value, and cost, among others. However, on the one hand, the currently used methods for early biomarker evaluation do not provide much insight into clinical value. On the other hand, methods such as early economic modeling, used to assess clinical value, are too extensive to be used for biomarker prioritization and require more data, are computationally more complex and, consequently, require a lot of time and resources to implement. As a consequence, assessing and prioritizing multiple

biomarker candidates, each with multiple possible applications, is usually unfeasible using such methods [13].

The fact that this decision involves several different criteria, and that there is usually a great number of potential biomarkers after the discovery phase of the biomarker discovery and development pipeline, makes it very hard to select the most promising candidates simply by analysing and comparing all of them, one by one. Therefore, developing MCDA approaches presents high potential and interest for the prioritization of biomarkers, since these approaches take under consideration all of the necessary criteria and allow for the evaluation of multiple candidates.

2.3 Chronic Obstructive Pulmonary Disease

COPD is a complex and heterogeneous disease, that is characterized by persistent and progressive airflow limitation and associated with an abnormal chronic inflammatory response of the lungs due to noxious particles and gases, leading to an acceleration of structural changes and narrowing of airways and, consequently, decline in lung function, as represented in Figure 2.3 [20][21]. The airflow limitations result from physiological changes in the lungs due to inflammation, fibrosis and liminal exudates (by a pathogen) [20].

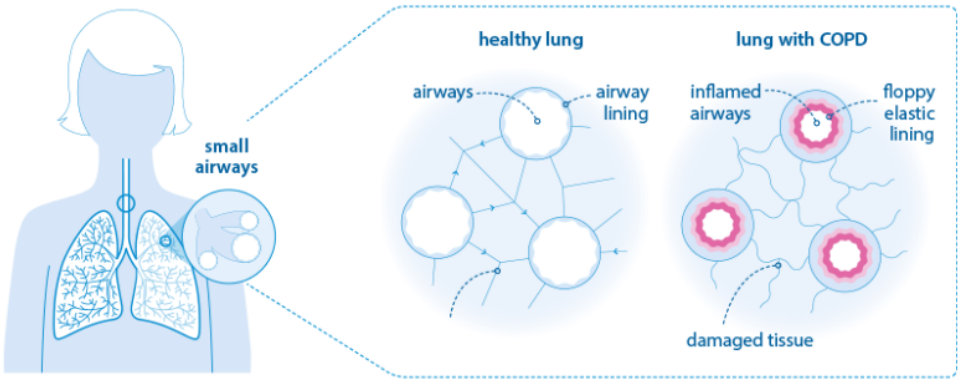

Figure 2.3: Comparison between an healthy lung and a lung with COPD (source: [22]).

COPD is a common, preventable and treatable disease, and even though it can also be caused by environmental factors, such as air pollution and biomass fuel exposure, its main cause is cigarette smoke: in fact, around 50% of smokers end up developing COPD, although the risk decreases by half after an individual quits smoking [5][23]. Apart from all this possible causes, host factors are also assumed to predispose individuals to develop COPD [5]. As for the major symptoms of COPD, they include chronic cough, excessive mucus production and dyspnea, which progress slowly in comparison to other lung diseases [20].

The fact that COPD is a heterogeneous disease means that pulmonary involvement usually varies widely, being frequently accompanied by extrapulmonary manifestations and several comorbidities [21]. Comorbidities, such as cardiac dysfunction, endocrine disorders, lung cancer and osteoporosis, as well as exacerbations, are associated with the severity of the disease [20][21][24]. These comorbidities are often associated with age, since clinical symptoms tend to manifest only after a prolonged exposure to the toxic substance, as well as with an unfavorable social environment, since they tend to arise in the most vulnerable segments of society [21].

As the disease progresses, pulmonary hypertension may occur due to hypoxic vasoconstriction. In severe COPD, patients occasionally suffer from dysfunction of respiratory skeletal muscles, which can result in weight loss, sepsis, fatigue and weakening, among others. Also, respiratory infections,

environment and pollution, result in sudden exacerbations in these patients, leading to an increase of the inflammation rate. Consequently, there will be a sudden increase in airway limitations and gas entrapment, which can cause systemic effects like cardiac failure and aging, in chronic patients [20].

It has been reported that COPD kills more than four million people per year, and 90% of those deaths occur in low and middle-income countries. According to World Health Organization (WHO), this disease was the fifth cause of death in 2002 and is expected to be the third leading cause of death by 2030 [20][24]. By association with its high prevalence, COPD generates significant social and healthcare costs [21], resulting in $44 billion healthcare expenses every year [24].

Due to the fact that COPD is a very complex and heterogeneous disease, making it very difficult to efficiently stratify patients and to introduce personalized therapeutic approaches, as well as the fact that the currently available tools do not predict the progression and exacerbation of this disease efficiently, it will be used as an example disease in this thesis for prioritization of biomarkers, since there is a clear need for new clinical approaches. Therefore, in the next chapter, a literature review will be made concerning COPD, from its characteristics to current clinical approaches, biomarkers and precision medicine associated with it, as well as concerning MCDA models and participatory methods, namely which will be used in this thesis and why.

Chapter 3

Literature Review

In this chapter, a literature review will me made, focused on MCDA and participatory methods, as well as on the application in COPD of all the concepts previously discussed in chapter 2.

First, an introduction to MCDA, the approach chosen for this thesis, will be made, followed by an overview of MCDA models in healthcare and, more specifically, in biomarker prioritization, which is the focus of this thesis. Secondly, an overview of participatory methods will be made, with special focus on the Delphi method, which will be used in this thesis. Finally, COPD will be further explored, namely regarding the use of biomarkers and the usefulness of personalized medicine in a complex and heterogeneous disease, such as this one.

Databases such as google scholar, pubmed, b-on and science direct were searched, using expressions such as biomarkers, prioritization, multicriteria decision models, MACBETH, participatory methods, Delphi, COPD and combinations among them.

3.1 Prioritization Approaches - Multiple Criteria Decision Analysis

Modelling decisions is not as straightforward as economic rationalists proclaim. Although there are some highly complex economic models, they are not necessarily cost-effective or realistic. Simply stated, a good model is defined as a model that reflects accurately the decision-maker(s) perceptions [25].

There are several decision making techniques available, but we chose to use Multiple Criteria Decision Analysis (MCDA) models for biomarker prioritization in this thesis, because they are not only some of the most commonly used approaches for priority-setting [26], but they can handle problems with multiple objectives, have an encompassing nature, are intuitive, have theoretically sound methods to balance benefits, costs and risks, allow for the use of different types of data, allow the involvement of different stakeholders, taking their preferences and values under consideration, have easily understandable outputs [27][28], and promote transparency, accountability and reasonableness in decision-making [29]. This technique also has the great advantage of capturing the knowledge from a decision, making it reusable for others who need to make the same decision, or a similar one [30]. Considering that the purpose of this thesis is to create a model to assist researchers and clinicians in the selection of the most promising biomarkers for validation and translation into clinical applications not only in COPD, but also in other diseases, the previously mentioned advantages of MCDA make it a very promising approach.

MCDA is a tool used for decision making, that uses a set of quantitative and qualitative approaches, that simultaneously and explicitly takes under consideration multidimensional and often conflicting factors, allowing for comparison of technologies, namely medical, by combining individual criteria into one overall assessment. This tool can make complex decision-making processes significantly simpler, and

has been successfully used since the 1960's to solve decision problems in many areas, namely in financial decision-making, environmental impact studies, resource allocation, budgeting and geographical information systems [31][32].

The fact that MCDA can consider factors that go beyond cost-effectiveness analysis, as well as the fact that it makes the decision process more rational, efficient and explicit, gives it potential to improve the quality of decision making and offers several possibilities that would be very interesting in Health Technology Assessment (HTA) [31][32]. Over the last 10 years, several different MCDA models have been created and applied in HTA, namely across Europe, each with its specific characteristics, including slightly different approaches, focus and complexities, after several groups of scientists started realizing some of the issues associated with the healthcare reimbursement decision-making process [31].

MCDA has the potential to overcome the challenges presented by traditional decision-making tools, in particular when the decision-making is complex and includes multiple criteria, multiple stakeholders and both quantitative and qualitative data [32].

3.1.1 Multiple Criteria Decision Analysis Models in Healthcare

Decision making in health care is of extreme importance, since these decisions are not only complex, but also involve uncertainties and have to consider the preferences and values of stakeholders. Even though a variety of methods have been proposed to support the decision-making process in healthcare, MCDA is one of the most frequently used methods [32][33]. In the healthcare domain, MCDA can be used to assess new health technologies and orphan drugs, improve the efficiency, rationality and legitimacy in resource allocation, support benefit risk assessment, develop universal coverage health benefit package, portfolio decision analysis, include stakeholder input and preferences in comparative effectiveness research, prioritize investment in public health interventions, prioritize patient's access to services and weigh the several endpoints in the assessment of efficiency and quality in healthcare [27][32][33].

In 2014, G. Adunlin et al. [32] conducted a study with two main goals: to systematically identify applications of MCDA in healthcare and, based on the identified bibliographical records, identify and report the publication trends of MCDA. English language studies, from 1980 to 2013, that used any MCDA technique in the healthcare area (instead of focusing on only one) and that involved the participation of decision makers, were considered for inclusion in this study: and a total of 66 citations were included, the majority conducted in the United States. As a result, it was determined that the number of publications reached its highest value in 2012, that cancer was the most researched healthcare topic (among a total of 60 interventions or disease areas), that the most covered area of application was diagnosis and treatment (among 14 different areas) and that the most commonly used MCDA technique was Analytic Hierarchy Process (AHP).

Among all the techniques used in the considered studies, AHP was the most commonly used, likely due to the fact that it is very flexible, helps capture both subjective and objective aspects of a decision, and a lot of software has been developed for this method [32]. However, this method has been very criticised by several authors over the years. According to L. Rietkott [34], one object of criticism is "the possibility of rank reversal in case of the introduction of an additional alternative to an existing decision problem". The fundamental scale has also received some negative criticism, namely by H. A. Donegan et al. [35], who claimed that the English verbal scale used in AHP presents some ambiguity. S. Karapetrovic et al. [36] have also shown in their study that the consistency ratio was above 10% in several cases, even though the judgements were made logically and not randomly. Lastly, according to C. Bana e Costa et al. [37], AHP violates the "Condition of Order Preservation", fundamental in decision-making, which states that, in a set of alternatives, both the order and intensity of preference must be

maintained. In sum, AHP has been severely criticised due to a missing adherence to axioms of utility theory [34] and not respecting the theoretical foundations of Multi-Attribute Value Theory (MAVT) [27].

According to A. Angelis and P. Kanavos [38], there are several MCDA approaches, which can be classified in three different groups: (a) value measurement methods, which includes MAVT, (b) outranking methods and (c) 'satisficing' and aspiration level methods. In this thesis, we wanted to use a value measurement method, due to the simplicity of the required value judgements and to the fact that they can be applied to multiple decision contexts. More specifically, we wanted to use a MAVT method, since they are comprehensive, robust and they are able to reduce motivational biases and ambiguity [38]. As a result, AHP, as well as all the other MCDA methods that do not respect the theoretical foundations of MAVT, has not been considered to be used in this thesis, even though it is one of the most frequently used and implemented MCDA methods [39].

Therefore, we analysed some of the most well known and most used MCDA methods that respect the theoretical foundations of MAVT, according to C. Bana e Costa [40], which are described below, in order to select the most suitable one for the objective of this thesis.

MCDA methods used for scoring:

- **Bisection Techniques:** This is a numerical approach, used for cardinal value measurement [40]. In this technique, the decision-maker is asked to determine, on the attribute scale, the value point that is exactly halfway of the maximum and minimum value of the scale, followed by the value point between the point previously discovered and the maximum, then the minimum, and so on, until the decision-maker is satisfied with the result [41].
- **Direct Rating:** This is a numerical approach, used for cardinal value measurement [40]. In this technique, the most important criterion is assigned 100 points (for example), and the least important 0 points. The importance of the remaining criteria is then determined comparatively to the most and least important criteria, in such a way that the intervals between them represent the intensity of preferences of the decision-maker [42].

MCDA methods used for weighting:

- **Trade-off:** This is a numerical approach, where two options, different in only two criteria, are compared at a time. One of the options has the best impact on the first criteria and the worst on the second, and the other option the exact opposite. This way, the decision-maker can determine the most important criterion of the two, by choosing his preferred option. After this, and in order to yield indifference between the two options, an adjustment of the impact level is made, consisting on worsening the best impact of the chosen option or improving the worst impact of the non-chosen option. This adjustments have to be performed for the n-1 pairs of options (where n is the number of criteria). If the value functions associated to the criteria are known, numerical values for the scaling constants can be derived through the resolution of a equation system [40].
- **Swing Weighting:** This is a numerical approach that is used to elicit relative criteria weights, which is done by giving a score of 0 to the attribute with the worst performance, 100 to the one with the best performance [38], and a value in between to the other attributes, according to their performance (the better the performance, the higher the value, as a percentage of the highest value swing) [43].
- **Point Allocation:** This is a numerical approach, where the decision maker is given a certain number of points, for example 100, and is asked to divide them among the criteria, describing

their weights directly. The more points a criterion receives, the more relative importance it has. Although this method is very easy, it has two main disadvantages: it becomes harder as the number of criteria increases and weights obtained are not very precise [44].

MCDA methods used for both scoring and weighting:

- **Measuring Attractiveness by a Categorical Based Evaluation Technique (MACBETH):** This is a non-numerical approach, that can be good for decision-makers that do not feel comfortable in directly scoring the options [40]. This technique considers pairwise attribute comparisons (on each criterion or among criteria). The qualitative judgements to compare them are made by the decision-makers, using a semantic scale to express the differences in attractiveness of those attributes (null, very weak, weak, moderate, strong, very strong and extreme), resulting in a value function [38][41].

After analysing all the options, while considering the characteristics of the problem presented in this thesis, we opted to use MACBETH as our multiple criteria decision method, since it can be used for both scoring and weighting, it is a non-numerical approach, which makes it easier for the decision-maker to express value judgements, and because we considered it the best option to answer the problem.

According to C. Bana e Costa et al. [45], MACBETH is a ”humanistic, interactive and constructive approach to the problem of how to build a quantitative model of values based on qualitative (verbal) difference judgements, that facilitates the path from ordinal to cardinal preference modelling, namely analysing judgmental inconsistency and offering suggestions to move the process forward”. The fact that this method allows quantitative models to be built based on qualitative judgements makes the whole process significantly simpler, which is a great advantage.

The MACBETH process, like any multicriteria analysis process, usually starts with an analysis of the context of the problem, followed by a discussion and definition of the exclusion and evaluation criteria, as well as the descriptors of performance (either quantitative or qualitative). The next step is the definition of the levels of performance and reference levels to be used for building the value function. Next, the MACBETH protocol for questioning the decision-makers is used, in order to obtain pairwise absolute judgements on differences of attractiveness between levels of performance or options of the criterion, using the semantic scale: null, very weak, weak, moderate, strong, very strong, extreme. If the answers are consistent, the M-MACBETH software will then suggest a quantitative value scale, through the resolution of a linear programming problem, which should be adjusted and validated with the decision-maker. In case the answers are not consistent, M-MACBETH will suggest an alternative to make them consistent. Following the creation of the value scale, the global value of each proposition is calculated, using the additive aggregation model. Finally, a robustness and sensitivity analysis of all the options considered to answer the problem in hand is performed. With all the information obtained, one of the options will be recommended as the best or most promising [45][46].

As MACBETH includes both technical and social components, it is a socio-technical approach [47]. Consequently, it requires the appropriate usage of participatory methods. Therefore, these methods will be presented and discussed further ahead in section 3.2.

When an additive model is applied, which is the case, interaction between criteria should not be allowed, this is, the performance of a criterion should not depend on the performance of other criteria. Therefore, to ensure that there are no preference dependencies between criteria, it is highly relevant to conduct a test on preference dependence [48]. This method will be presented and explored in section 3.3.

The MACBETH approach has already been applied in several healthcare related studies, namely for: the prioritization of community care programmes [49]; supporting planning of decisions in the long-term care sector [50]; hospital auditing [51]; building a population health index [52]; assisting in the diagnosis of Alzheimer’s disease [53]; and evaluating health and safety risks [54].

Although this method has not been applied in healthcare issues as many times as other MCDA methods, it has suffered very limited criticism and presents great potential [34].

3.1.2 Multiple Criteria Decision Analysis Models in Biomarker Prioritization

The number of biomarkers that enters the biomarker discovery and development pipeline, with the aim of reaching a clinical application, is too high compared to the resources available to do so (both monetary and time related). Consequently, it is essential to identify the most promising candidates, with the highest chance to succeed as a commercial product [13], in order to increase the percentage of qualified biomarkers, while reducing the costs. Therefore, it is highly helpful to use a tool, such as MCDA, that takes under consideration all of the necessary criteria (since, in this situation, factors other than cost are critical) [55] and allows the evaluation of multiple candidates, to help in the decision-making process of selecting the most promising biomarker candidates for a certain disease.

There are several examples in literature about the use of MCDA in healthcare. However, there are very few studies regarding the use of MCDA in biomarker prioritization, one example being the study by A. Miquel-Cases et al. [56], where MCDA is used, among other techniques, to select the most promising biomarkers for breast cancer, but with no particular focus on it.

As for the specific case of COPD, although there are articles where MCDA is used in association with it, namely in the study by K. Marsh et al. [57], where an evaluation of COPD treatments is made using MCDA, there is almost no information regarding biomarker prioritization in COPD using this method, even though there are some studies, such as the study by S. Ongay at al. [2], where biomarkers are prioritized by applying several criteria individually, including the number of individuals considered in the cohorts and statistically significant differences between COPD patients and healthy smokers.

Although there are not yet many examples of the application of MCDA in biomarker prioritization, specially in COPD, there are several examples of its application in healthcare and other areas, and it shows great promise. Therefore, to efficiently select the most promising biomarkers of a disease, a methodology using the MCDA method MACBETH is proposed and explained in the next chapter.

3.2 Participatory Methods

In order to select the criteria to be used in a multicriteria model, it is possible to use several methods, ranging from selection of the criteria based on literature review, to participatory processes. In this thesis, we will start by selecting criteria based on literature review and then we will be applying participatory processes, since the chance that the selected criteria will be more commonly understood, credible, technically useful and scientific and policy relevant is higher when using the later [58].

A participatory method is an approach that involves ”the public” relevant to the topic being evaluated in decision-making processes. The public can include citizens, stakeholders of a particular project, experts or even members of private industry and government. When applied early in a decision-making process, participatory methods allow the participants to share their perspectives on the issue being evaluated, being useful to achieve consensus when there are differences in opinion or conflicts among the participants. Using these methods allows all voices to be heard, leading to a more democratic decision-making process and to an improvement of the quality of the decisions [59].

However, there are many participatory methods available. Some of the better known and most used ones, based on N. Slocum [59], are presented in Table 3.1. When deciding which method to use, it is important to consider five elements: objectives, nature and scope of the issue, participants, time available and budget [59].

Table 3.1: Participatory methods and respective descriptions (adapted from: N. Slocum, 2003 [59]).

Participatory Methods	Description/Objective
Charrette	Intensive face-to-face process with the purpose of bringing people from different sub-groups of society into consensus in a short period of time. The main issue is divided in parts, each of them assigned to a certain sub-group of people, who periodically report back to the group. Feedback is then given. This process is repeated until consensus is reached.
Citizens Jury	This method is used to obtain informed citizen input into policy decisions, leading to a more democratic decision-making. The jury, composed by random citizens, is informed about the issue by experts, who present them with several perspectives. The jurors are often divided into subgroups to focus and deliberate on different aspects of the issue, with the goal of reaching a decision or providing recommendations in the form of a citizens' report.
Decision Conference	Process used when key players have to assess a socially controversial topic. Their questions and concerns are presented to a panel of experts and, based on their answers, a negotiation is initialed, assisted by an impartial facilitator, with the purpose of reaching consensus.
Delphi	This method consists on an iterative survey, where experts complete a questionnaire and are then presented with feedback/statistics on the responses given by all participants. Considering this information, they then fill the questionnaire again, being able to change their previous answers. The process is repeated as many times as necessary to reach consensus. This process can be conducted online or face-to-face, but the first one has the advantage of allowing the answers to be anonymous.
Expert Panel	In this process, a panel of experts synthesises diverse inputs, from testimonies to research reports (among others), producing a report that includes notes and recommendations regarding the topic under analysis, that can be used in the future.
Focus Group or Interviews	Planned discussion with a small group of stakeholders, conducted by a facilitator, in a permissive environment. The purpose of this process is to gather information about people's preferences and values, regarding a certain topic.
PAME (Participatory Assessment, Monitoring and Evaluation)	The objective of this method, which is usually conducted as part of another participatory method, is that stakeholders of a project can stop and reflect on the past, with the purpose of making decisions about the future. The participants share the responsibility for deciding what is to be evaluated, selecting the methods and data sources, carrying out the evaluation and, finally, analysing and evaluating the acquired information and presenting the results.
Planning Cells	In this method, randomly selected people work as public consults for a limited amount of time, with the purpose of reaching solutions for a certain planning or policy problem. The number of cells vary with the problem, but every cell has two process-escorts who are responsible for scheduling the information and moderating the plenary sessions. In each cell, participants acquire information about the problem, explore and discuss said information to reach solutions and evaluate those solutions in terms of desirable and undesirable consequences. Stakeholders, experts and interest groups can present their position to the participants of each cell. The results of this method are presented in a citizen report.
Scenarios Workshop	Scenarios are descriptions of potential futures focused on relationships between events and decisions points. They are useful when the past or present experience is unlikely to be a useful guide for the future, due to the problem being complex, the probability of a significant change being high, the time-horizon being long or the dominant trends maybe not being favourable.
World Café	Creative process that has the goal of facilitating the exchange of knowledge and ideas. In this process, a café ambiance is created, and the participants are divided in several groups, each group in one table, where they discuss a question or issue. At regular intervals, the participants move from table to table, except the table host, who always stays in the same table to summarise the previous conversations to the new participants. When the process is finished, the main ideas are summarised and follow-up possibilities are discussed.

During the structuring of the model, which is the focus of this thesis, we want to interview a few experts, followed by the gathering of the opinion of a great number of experts from different geographies in an easy and confidential way (an online method would be preferable), the time available is very reduced and there is no budget. Therefore, the interviews and the Delphi are the methods that will be used as they are the ones that answer better to the presented specifications. During the remaining phases, the Delphi method would be used as well, but we would also want to confirm our results and decisions, in order to reach consensus, and to validate the model, for which a decision conference would be the ideal choice. As a result, a collaborative value model will be applied in this thesis.

3.2.1 Collaborative Value Model

Multicriteria decision conferencing has proven to be effective, in several contexts, for small groups, in creating a collaborative environment where individual beliefs can be shared, common concerns identified, eventual conflicts managed and group agreement promoted when building a model. However, when the framework is to be extended to broader participatory contexts, a different design of the social process is needed, in order to capture the points of view of all participants. This problem can be solved using the Delphi method, since it elicits and analysis individual judgment knowledge from a very large and diverse group of individuals. The resulting model is called the collaborative value model [47].

The collaborative value model is a socio-technical model that combines multicriteria decision conferencing and the Delphi participatory process, with the goal of building widely informed evaluation models. This model allows for a unique collaborative learning, where participants are seen not only as experts providing individual expertise, but also as collective learners that construct shared judgemental knowledge, leading to widely informed and more acceptable evaluation models. It is built based on the existing collaborative knowledge acquisition methodology plus the Delphi method. After this, a decision conference takes place, where the knowledge obtained is analysed by a small group of key-players, with the purpose of collaboratively developing a widely informed multicriteria evaluation model [47]. An overview of the components that integrate the collaborative value model framework is presented in Figure 3.1.

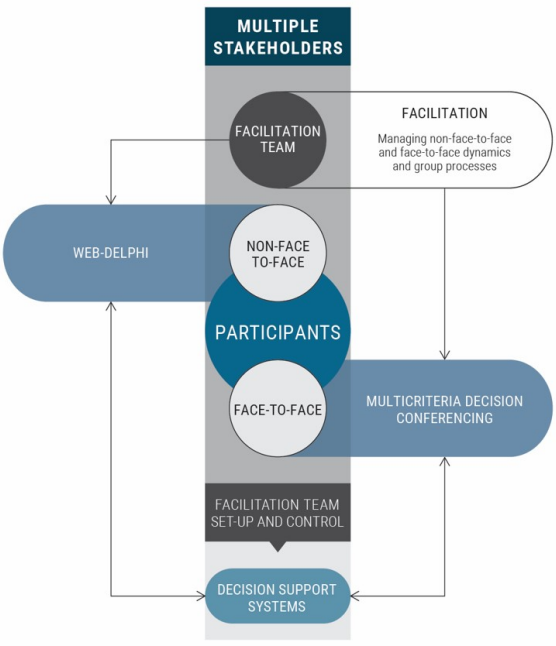

Figure 3.1: Components that integrate the Collaborative Value Modelling framework (source: A. C. Vieira et al., 2019 [47]).

3.2.2 Decision Conference

According to L. D. Phillips and C. A. Bana e Costa [60], a decision conference ”is a gathering of key players who wish to resolve important issues facing their organisation, assisted by an impartial facilitator who is a specialist in decision analysis and works as a process consultant, using a model of relevant data and judgements created on-the-spot to assist the group in thinking more clearly about the issues”. This process is not used to find the ”right answer” or the optimal solution, but to aid in thinking and group learning.

A decision conference is an iterative and interactive model-building process that is usually divided in four phases: exploration of the issues, structuring and building a model, exploring the model and, finally, agreeing about the way forward. Throughout these phases, new data can be introduced when needed and debate is encouraged, meaning that any issue can be handled as soon as it arises and different opinions can be easily expressed and discussed, resulting in gradual understanding of the issues and eventual consensus among the group [60].

Since decision conferences consist in a face-to-face process, they present several advantages for the development of multicriteria models: participating experts and stakeholders can discuss their points of view and concerns, manage conflicts and reach a shared understanding of the problems in an easier way. However, it has a significant downside: the model constructed using this process may be based in several perspectives about the problem, but the number of participants is small, leading to representativeness issues [47].

3.2.3 The Delphi Method

Decision conferences are a very effective socio-technical approach to create a collaborative environment with the purpose of identifying individual beliefs and common concerns, as well as managing eventual value conflicts and promoting agreement in group model building. However, this approach works better for small groups, decreasing its efficiency as the number of participants increases. In cases where the participatory contexts are broader, a different social process is required, in order to capture the points of view of all participants [47].

The Delphi method is a participatory process that enhances decision-making in a systematic way, characterized by its qualitative and structured technique to reach group agreement in an effective way, during a group communication process, in particular when the problem is complex. It presents several advantages, including ease of application, ensured anonymity, which allows participants not to fear group pressure when giving or changing their answer, and the ability to gather opinions and knowledge of a great number of individuals with several different backgrounds and geographic locations. Using this method, feedback of group opinion is given after every round and each participant can revise his/her answers and chose whether to maintain them or change them, after analysing the group statistics. Throughout the years, it has been used multiple times in several fields where group opinion is needed from individuals with different and varied views, namely in health, needs assessment, program planning, resource allocation and policy determination fields [58][61].

The management of the responses (and non-responses) is critical in studies that use the Delphi method. Therefore, an impartial facilitator is responsible for the management of the whole process. By using a web platform, the efficiency of both the facilitator and the process increases, making it easier for the participants to answer, improving the communication between rounds and simplifying the overview of the process and the analysis of the responses by the facilitator [58]. There are no specific guidelines regarding the response rate throughout the Delphi process. However, according to A. Freitas et al. [58], several authors find it necessary to have at least a 70% response rate in each round to maintain rigor.

For the purpose of this thesis, a web-based Delphi will be developed in order to assess the relevance of previously defined criteria, taking under consideration the opinions and views of experts and stakeholders. This method has specific rules regarding how to measure level of agreement and how to deal with differences in opinion [58].

In Figure 3.2, a flowchart of the decision rules to be adopted in this thesis in the web-based Delphi is presented. Using the example of criteria selection, for each criterion considered, each participant can evaluate its relevance using five different terms: Strongly Agree (SA) and Agree (A), indicating agreement, Strongly Disagree (SD) and Disagree (D), indicating disagreement or Neither Agree nor Disagree (NAD). Apart from this, participants can also add free-text comments in a specific space.

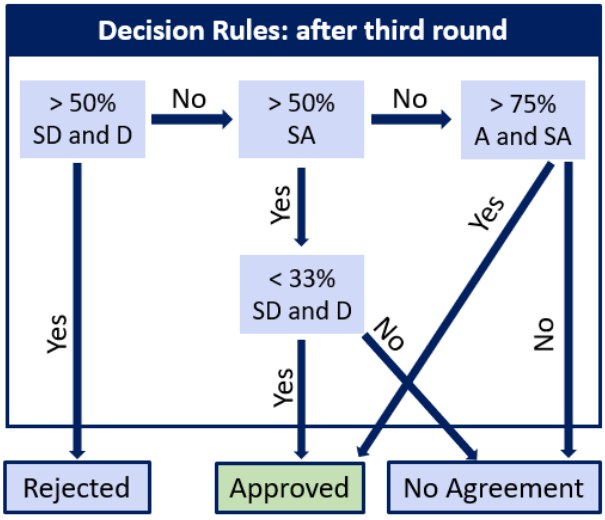

Figure 3.2: Flowchart of the decision rules to be adopted for criteria approval and rejection in the web-based Delphi, after the end of the third round (adapted from: A. Freitas et al., 2018 [58]). SD: Strongly disagree; D: Disagree; NAD: Neither agree nor disagree; A: Agree; SA: Strongly agree.

Different authors describe different ways to determine if consensus has been reached for a certain criterion: L. E. Miller [62] says it can be done if a certain percentage of the votes is within a determined range; F. L. Ulschak [63] suggests that consensus is achieved if 80% of participants' votes fall within two categories on a seven level scale; P. J. Green [64] suggests that consensus is achieved if a minimum of 70% of the participants rate three or higher in a four point Likert-type scale and the median needs to be equal or higher than 3.25.

In this thesis, we will be using a Likert-type scale with five qualitative scale levels (strongly agree, agree, neither agree nor disagree, disagree and strongly disagree). Therefore, we will consider that consensus is reached if a minimum of 75% of participants choose agree or strongly agree to rate a criterion.

When using the Delphi method, it is necessary to perform a succession of rounds to acquire the necessary knowledge (see Figure 3.3): the first has a divergent thinking nature, where the participants give their own judgement, while the remaining rounds have a convergent thinking nature, since the participants are invited to reconsider their previous judgements, after having access to anonymous statistics from the previous rounds. This promotes the identification of key issues and effective group thinking [47]. From the third round on, the goal is to achieve consensus or stability of the responses [65]. Therefore, three is the ideal number of rounds for a web-based Delphi.

In the first round, where divergent thinking is promoted, the participants are exposed to the proposed criteria for the first time, and answer according to their individual judgement [47][58]. In the second round, contrarily to the first one, convergent thinking is promoted [47], and the participants are presented

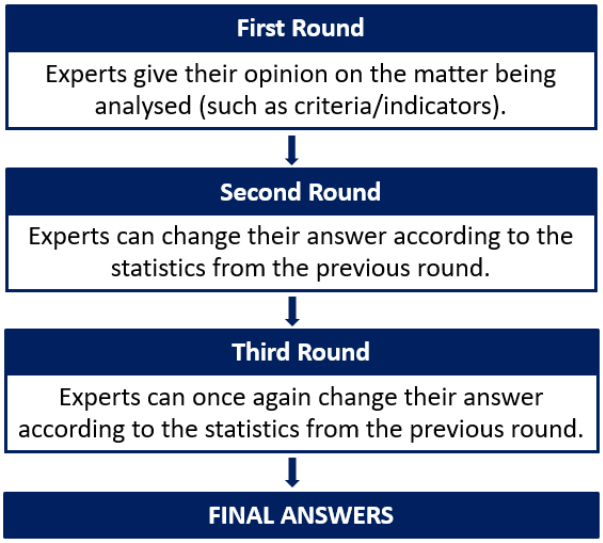

Figure 3.3: Flowchart presenting the main goals of each of the three rounds of the web-based Delphi.

with the results of the first round, including the rejected and approved criteria, and an anonymous aggregation of all participants answers for each criterion, which allows them to reconsider their answers, after analysing the group opinion. Every criterion, regardless of its percentage of approval/rejection on the previous round, is included in the second round for reevaluation, allowing the participants to maintain or change their previous answer after analysing the information provided, with the goal of reaching greater agreement. Finally, the same procedures applied in round two are applied in round three. However, after the third and last round, if no agreement is reached for a certain criterion, it will not be reevaluated in a new round, being, therefore, automatically excluded from the study for not presenting enough relevance to be accepted, according to the experts. Since every individual must participate in all three rounds, in order to provide their answer sequentially and to get feedback from the previous round, being able to alter their answer after knowing the group opinion, if an individual does not answer in one of the rounds, he/she is not invited to the next one, in order to maintain a continuous and stable participation [58]. The divergent-convergent thinking used in this method promotes understanding among the group, identification of key issues and effective group thinking [47].

In order to analyse the data obtained in Delphi, and to present the information regarding the collective judgements of the participants, there are two major statistics that can be used: measure of central tendency, which includes mean, median and mode, and level of dispersion, which includes standard deviation and inter-quartile range [61][66]. When the data obtained is ordinal, then non-parametric measures of central tendency are appropriate. In the other hand, when Delphi produces qualitative data, then methods such as content analysis (including graphical presentation of the data), and thematic analysis may be more appropriate [66].

3.3 Preference Dependence Test

When applying an additive model, one criteria cannot depend on another, and two criteria cannot interact, meaning that the choice of performance of one criterion should not depend on the performance of other criteria [48]. In fact, until an evaluation criterion is confirmed to be independent, it is called an Evaluation Dimension (ED).

Preference dependence can be discovered when the criteria are being defined, when scoring the alternatives, or when stakeholders say that they cannot make a judgement regarding their preference

for one criterion without knowing the performance of other criterion [48].

To test preference dependence between EDs, which must be done before starting the building of the model, the two reference levels of each descriptor (*good* (G) and *neutral* (N)) must be determined. For the purpose of illustrating the procedure, we will consider two EDs, a and b, and their respective levels of performance X_a and X_b, where: G_a and N_a are, respectively, the *good* and *neutral* levels of a, and G_b and N_b are, respectively, the *good* and *neutral* levels of b. Therefore, a set of global impacts to test for preference dependence between two EDs a and b would be: (N_a, N_b), (N_a, G_b), (G_a, N_b), (G_a, G_b) [67].

The relevant swings to consider for testing preference dependence between two EDs a and b are presented in Figure 3.4: (i) is the swing from an option that is on the *neutral* level on both EDs to an option that is on the *good* level on ED a and on the *neutral* level on ED b; (ii) is the swing from an option that is on the *neutral* level on ED a and on the *good* level on ED b to an option that is on the *good* level on both EDs; (iii) is the swing from an option that is on the *neutral* level on both EDs to an option that is on the *neutral* level on ED a and on the *good* level on ED b; and (iv) is the swing from an option that is on the *good* level on ED a and on the *neutral* level on ED b to an option that is on the *good* level on both EDs [67].

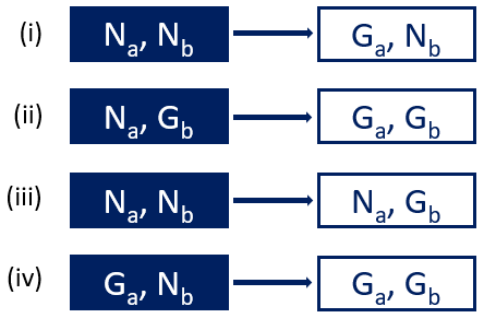

Figure 3.4: Swings for testing preference dependence between evaluation dimensions a and b, where G_x is the *good* level of ED X, and N_x is the *neutral* level of ED X (adapted from: C. A. Bana e Costa et al., 2017 [67]).

In other terms, the preference dependence test consists on verifying equations 3.1 and 3.2, which can be done using MACBETH (both are necessary because the difference of dependence is not symmetric) [67].

$$v(N_a, N_b) - v(G_a, N_b) = [v(N_a, G_b) - v(G_a, G_b)] \tag{3.1}$$

$$v(N_a, N_b) - v(N_a, G_b) = [v(G_a, N_b) - v(G_a, G_b)] \tag{3.2}$$

The decision-maker is asked to make qualitative judgments regarding the difference of attractiveness between the four swings presented in Figure 3.4. A M-MACBETH matrix is then filled with these judgements, which are made using the MACBETH semantic scale (null, very weak, weak, moderate, strong, very strong or extreme) [67].

The questions asked to the decision-maker to check if ED a is preference dependent on ED b are: (a) "What is the attractiveness of the swing/improvement from (N_a, N_b) to (G_a, N_b)?" or "What is the difference of attractiveness between the option (N_a, N_b) and the option (G_a, N_b)?"; (b) "What is the attractiveness of the swing/improvement from (N_a, G_b) to (G_a, G_b)?"

Considering that the only difference between swings (i) and (ii), contemplated in questions (a) and (b), is the level of ED b, if the two swings are judged differently, then ED a is preference dependent on

ED b.

Since preference dependence is not symmetric, it is also necessary to test if ED b is preference dependent on ED a, based on the following questions: (c) "What is the attractiveness of the swing/improvement from (N_a, N_b) to (N_a, G_b)?"; (d) "What is the attractiveness of the swing/improvement from (G_a, N_b) to (G_a, G_b)?"

In the case of swings (iii) and (iv), contemplated in questions (c) and (d), the level of ED a is fixed on its *good* and *neutral* reference level in each swing. If the two swings are judged differently, than ED b is preference dependent on ED a.

If the decision-maker shows difficulty in understanding the swings being assessed, there is the option of presenting them in a more graphical way, facilitating the visualization, as presented in Figure 3.5.

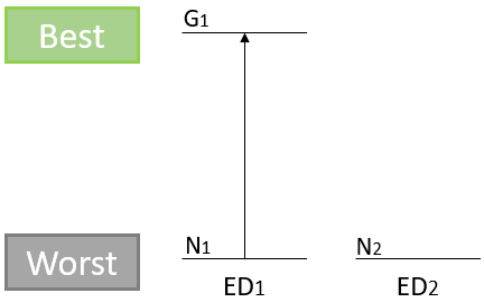

Figure 3.5: Graphical representation of the swing (N_a,N_b) $\rightarrow$ (G_a,N_b) to test preference dependence between evaluation dimensions a and b, where G_x is the *good* level of ED X, and N_x is the *neutral* level of ED X.

The outcome of the preference dependence test can be one of three: (1) the EDs are bidirectionally/bilaterally preference dependent; (2) the EDs are unilaterally preference dependent; (3) the ED are preference independent.

The first scenario occurs when the decision-maker gives different judgements to swings (i) and (ii), and then to swings (iii) and (iv). The second scenario occurs when either swings (i) and (ii) or swings (iii) and (iv) receive different judgements. Finally, the third scenario occurs when swings (i) and (ii), and then swings (iii) and (iv), receive equal judgments.

In case there is preference dependence between EDs, a redefinition is in order, taking under consideration the requirements of additive models. In order to address the problem, there are several methods that can be used, namely: (1) the Choquet Integral, which requires bilateral preference dependence; (2) the Multilinear Value Model, which requires weak difference independence; (3) the redefinition of dependent EDs by combining them into a single independent criterion [48][67]. If any EDs in this thesis happen to be preference dependent, either uni or bilaterally, the method to be applied to solve the problem is the third one, due to its simplicity in comparison to the remaining two. Therefore, if necessary after the test on preference dependence, the model will have to be restructured by merging and redefining the dependent evaluation dimensions, so that the final set of evaluation criteria meets the requirements of the additive model [48].

Despite some restrictions of the additive aggregation model, namely, according to T. C. Rodrigues et al. [67], loss of richness of interactions along the means-end chains when evaluating decision options, the fact that it is not suitable to be used in problems where there are value interactions that need to be explicitly addressed, and the fact that there are several applicability conditions that must be met, including non-redundancy, non-overlap, exhaustiveness and preference independence between criteria [27], the additive aggregation model is the most used of all multicriteria methods, and it is a simple [67], easy to construct, interpret and communicate to decision-makers [48][67] model, that allows not only to

order proposals in terms of their global attractiveness, but also to analyze relative differences of global attractiveness [46]. Therefore, this is the method that will be applied in this thesis, and will be further explored in chapter 3.

3.4 Personalized Medicine in COPD

Personalized medicine takes individual variability into account when making clinical decisions, leading to better treatment outcomes. For this reason, it is particularly relevant in heterogeneous diseases, such as COPD, where the original complexity is multiplied due to the heterogeneity of exacerbations [8].

However, due to the very fact that COPD has an heterogeneous profile, and also due to its molecular and clinical characteristics, it is very difficult to efficiently stratify the patients (separating them into subgroups based on a specific characteristic) and to introduce personalized therapeutic approaches [24]. Therefore, in recent years, several attempts have been made with the purpose of characterizing different phenotypes and endotypes to enable a more individualized approach to clinical care, treatment, and continued follow-up. For example, in the BIOMEPOC project, biomarkers were identified in blood in order to improve the characterization of patients [21]. Ultimately, understanding and quantifying the differences between patients, using biomarkers (for example), will be the key to bring personalized medicine to COPD patients [8].

In the particular case of this disease, personalized medicine will have to incorporate and target all relevant disease characteristics for the individual patient, including not only symptoms, information obtained through spirometry and exacerbation risk, but also outcomes such as depressive symptoms, hyperinflation, abnormal body composition and hypercapnia [5].

For it to have a role in COPD, personalized medicine will have to consider human beings as being composed of and operating within multiple systems that are self-adjusting and that interact with each other, where illness arises from the interaction between these systems. In the long run, personalized medicine will help guaranteeing a greater patient autonomy [5], but first, it is crucial to find better biomarkers for treatment response and prognosis in individual patients with COPD [8].

3.5 Biomarkers in COPD

The understanding of COPD pathogenesis has increased significantly over the past thirty years. However, the clinical tools currently available are not able to efficiently predict the progression and the exacerbations of the disease, and there are still no disease-modifiers for COPD, apart from smoking cessation and domiciliary oxygen therapy for hypoxemic patients, since drug therapies are only useful to relieve the symptoms of the patients, leading to a better quality of life, but do not offer an actual cure for COPD [10][24]. The problem is that, due to the fact that this disease is an heterogeneous and complex disease in terms of phenotypes and outcomes, early diagnosis and appropriate treatment are crucial to decrease the rate of mortality of this disease [24].

One major obstacle to new therapies is the fact that COPD has a heterogeneous pathogenesis, more specifically, the molecular processes responsible for driving the airflow limitation, characteristic of COPD, are thought to be highly variable. As a consequence, there is a clear need for new COPD therapeutics, and biomarkers have an essential role in this process. For this reason, and to try to accelerate drug discoveries, FDA has published guidelines for biomarker development. The success of this initiative has already been evidenced by the qualification of the protein biomarker fibrinogen [10]. However, in spite of this success, biomarker discovery, development and implementation are still immensely challenging, with the majority of biomarkers not making it beyond the discovery phase.

Biomarkers are very relevant in COPD, because they can be useful for diagnosis, patient characterization and stratification, disease severity quantification, prognosis determination, prediction and detection of a treatment response and investigation of the pathophysiology of the disease [3]. There are five main biological categories where the most promising COPD biomarker candidates can be included in [24]:

- **Extracelular Matrix (ECM) Remodelling:** The proteins that belong to the ECM remodelling category are mainly related to proteolytic activity, which, in the particular case of COPD, results in damage of the pulmonary tissues when in excess [24].
- **Oxidative Stress Response:** Oxidative stress arises when the production of Reactive Oxygen Species (ROS) overwhelms the intrinsic anti-oxidant defenses of the body, leading to an imbalance between oxidants and anti-oxidants [68]. It can be caused by ROS generated from cigarette smoke or environmental pollutants or by ROS generated during metabolic or inflammatory reactions. As a result of elevated values of oxidative stress in COPD, the patient can suffer from cell damage, remodeling of extracellular matrix and blood vessels, steroid resistance, cell necrosis, apoptosis, autophagy, unfolded protein response, cell proliferation, endothelial dysfunction, inactivation of antiproteases, premature cellular senescence, elevated mucus secretion, epigenetic changes and autoimmunity [20].
- **Lipid Metabolism:** In COPD, lipids act as inflammatory mediators. As a consequence, their levels increase as the oxidative stress levels increase [24].
- **Inflammation/immune Response:** The biomarkers that belong to the inflammation/immune response category are relevant to COPD because the severity of the disease is associated with the inflammation resultant from the activation of innate immune response, due to exposure to irritants, such as cigarette smoke or air pollutants, which are highly associated with this disease [24].
- **Vascular Tone Regulation:** Pulmonary vascular changes are associated with several mechanisms, namely chronic hypoxemia, inflammation and cigarette smoking. Therefore, it has been observed in patients with COPD, presenting a connection with shorter survival rates and adverse clinical outcomes [24].

According to E. Aydindogan et al. [24], a single biomarker does not appear to be able to effectively predict exacerbations or stratify COPD patients, since the disease presents a very complex molecular profile. Therefore, in the case of COPD, it would be ideal to have a biomarker panel, instead of only one biomarker, for efficient clinical implementation. Ideally, this panel would contain one protein biomarker from each of the five main biological categories presented above. However, in this thesis, the prioritization of COPD biomarkers will be done for individual biomarkers, without considering any connection they may have or possible grouping advantages, even though the biomarkers determined as the most promising could eventually work well together to characterize COPD.

In order to develop useful COPD biomarkers, it is essential to understand the processes involved in the progressive physiological deterioration that results from COPD, as well as to take under consideration and to focus on endotype biomarkers with specific clinical phenotypes, biomarkers that present a difference between early phases of COPD development and the phase where the disease is established, exacerbation subtype biomarkers and biomarkers to predict or measure drug effects [3].

Despite the fact that proteomics studies could use a wide range of biological samples, from blood to sputum, saliva fluids, bronchoalveolar lavage and lung tissues with relevance to COPD pathogenesis, the great majority of proteomics studies are based on plasma or serum biomarkers, due to their

assay reproducibility and the fact that these biological fluids are more accessible and easy to analyze, making them more practical to study [3][24]. As a consequence, biased proteomics approaches, namely immuno-based assays targeting specific circulating proteins that could be involved in COPD lung pathology, have been chosen over other approaches to clinically evaluate large cohort of patients in the process of COPD biomarker development [24]. Therefore, the type of the biological sample will not be relevant for the prioritization of COPD biomarkers.

3.5.1 Current Clinical COPD Approaches Using Biomarkers

Spirometry

The current standard respiratory function test for the detection of COPD is spirometry and, according to the guidelines, the current clinically approved biomarker for diagnosis, monitorization of disease progression and response to therapy is the Forced Expiratory Ratio (FER), which is given by the ratio between the Forced Expiratory Volume in 1 Second (FEV_1) and the Forced Vital Capacity (FVC) [23][24].

The lowest the value of FER, the highest the airway obstruction will be and, consequently, the likeliest it is for the patient to be diagnosed with COPD [23]. The threshold value of FER, recommended by Global Initiative for Chronic Obstructive Lung Disease (GOLD), is FER $\leq$ 0.7 [3]. However, this value has been challenged, because the normal value of FER is lower for women and decreases with age, meaning that the value of the threshold is too low for younger adults ($<$50 years of age) and too high for elderly individuals, which may lead to misclassification [3][23].

FEV_1 also gives information by itself, indicating the intensity of air-flow obstruction and being used to classify severity among patients: FEV_1 $>$80% indicates mild, 50-80% indicates moderate, 30-49% indicates severe and $<$30% indicates very severe condition. Also, the rate of fall of FEV_1 with age gives information about the progression of COPD [21][23]. However, since there are several highly relevant elements related to COPD that are not taken into account when measuring FEV_1, it does not provide much information on prognosis, either regarding future exacerbations, decline in lung function or appropriate patient stratification, and on appropriate clinical management and treatment [21][24]. Even so, the use of spirometry is useful because airway limitation causes gas trapping during expiration, leading to a hyperinflammatory reaction and, consequently, to hypoxemia and hypercapnia, which results on an increase of the severity of the disease. Therefore, the values provided by the spirometer will inform the doctors if the patient may end suffering from reduced ventilation, CO_2 retention or impairment in ventilator muscles, as a result of airway limitation [20].

Spirometry is a very useful procedure, even though diagnosis is frequently delayed, due to the presence of symptoms not being a reliable indicator of disease and the fact that the value of FVC tends to fall first, but with initial preservation of the FER. Consequently, it is more difficult to address risk factors, such as smoking, or to optimize treatment early enough. However, it has the advantages of being relatively quick to perform, practical, safe and very tolerable to most patients, presenting immediate results to the clinician [23].

3.5.2 Qualified COPD Biomarkers

Currently, Fibrinogen is the only COPD biomarker qualified by FDA [69].

Several studies have shown that plasma fibrinogen, when used with a clinical history of exacerbations, results in an increased ability to predict the occurrence of future events in COPD. This fact resulted in FDA qualifying fibrinogen as a biomarker to be used in clinical trials with exacerbations as outcome, in 2016 [3][69]. However, although it can predict the occurrence of exacerbations, it cannot predict their

nature, and it lacks in sensitivity and specificity to be used individually in clinical practice [3]. Therefore, it is highly necessary to qualify new biomarkers to complement the benefits of fibrinogen.

Although fibrinogen has been qualified for drug development by FDA, there are currently no validated COPD protein biomarkers widely used on a clinical level. The most commonly used tool for assessing the status of a patient is FEV_1, but it has no use for patient stratification and it does not give any information on possible future exacerbations. Therefore, there is a clear need for biomarkers to improve the care and management of COPD patients [24].

With this in mind, a socio-technical methodology based on MACBETH, which will be presented in chapter 3, will be applied to the case of COPD, with the intuit of structuring a model for selecting the most promising biomarkers among those found in literature, so they can be further researched, with the goal of being validated, qualified and, finally, successfully applied in drug development and clinical practice.

Chapter 4

Methodology

In this chapter, the proposed methodology will be presented, starting with an overview of the main steps of the socio-technical approach to be applied, followed by a detailed explanation of each of the steps pertaining the structuring of the model, which is the focus of this thesis, applied to the case of COPD biomarker prioritization.

4.1 Proposed Methodology

In this thesis, with the purpose of designing an approach that allows for the prioritization of COPD prognostic biomarkers, a MACBETH socio-technical approach will be used. MACBETH was the chosen approach because it can be used for both scoring and weighting, it is a non-numerical approach, which makes it easier for the decision-maker to express value judgements, and because we considered it the best option to answer the problem presented, among those considered.

The socio-technical approach to be used includes the technical elements of the MACBETH approach, but also the social elements of participatory methods (such as the Delphi process), resulting in an approach where scientific evidence is combined with the opinion of different experts, where there is better communication and a shared understanding of the issues and where a sense of common purpose is generated [27][60][70]. By using multiple sources of information (researchers and clinicians), the amount of information available to build the model is maximized, the potential impact of sources that rely on unreliable and inaccurate information is reduced and the model becomes more inclusive and representative, leading to more validity and credibility [70].

The methodology to be used was inspired by the EURO-HEALTHY project [71], which applied a MACBETH socio-technical approach to a health problem, using the collaborative value model, being adapted to the biomarker context of this thesis. Some adaptations were made in order to reach the most credible and valid model possible, namely the addition of the test on preference dependence.

As demonstrated in Figure 4.1, the MACBETH socio-technical approach to build a multicriteria model to prioritize biomarkers can be divided in three main phases: model structuring, model building and model testing and validation. These phases are then divided in technical and social approach: the technical component involves the use of MACBETH, while the social approach has the purpose of providing the necessary information to build the multicriteria model, using web-based Delphis, interviews and decision conferences, because it was decided that the model created in this thesis would be based on collaborative value modelling, since it elicits and analysis individual judgment knowledge from a very large and diverse group of individuals [47].

Before anything else, it is necessary to have a facilitation team, which is responsible for managing the

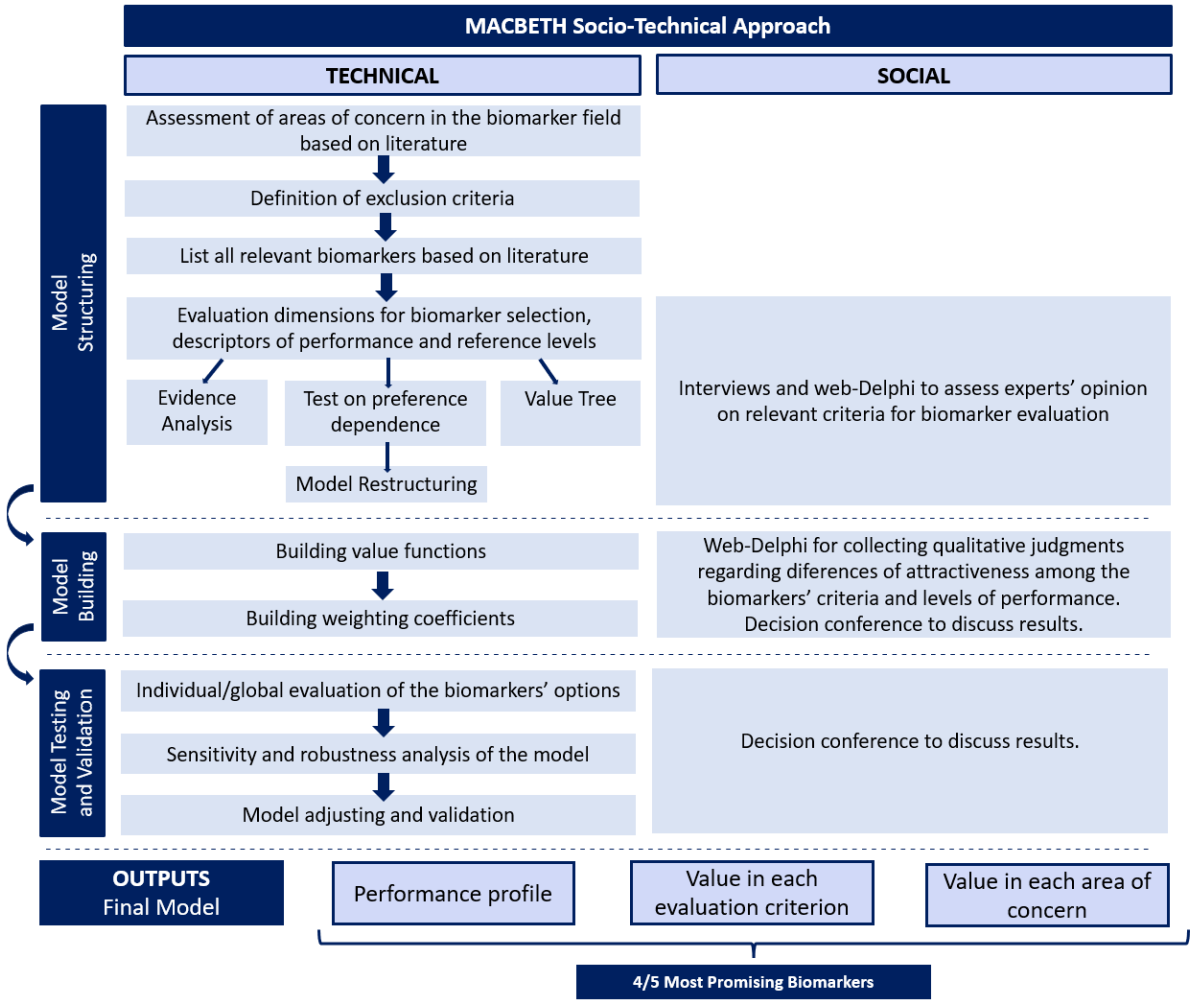

Figure 4.1: MACBETH socio-technical approach design for the prioritization of biomarkers (inspired in the methodology of the EURO-HEALTHY project [71] and adapted to the biomarker context).

dynamics of the whole process and for guaranteeing its success, as well as for conducting the activities during MCDA model building [47][58].

Initially, in the model structuring phase, we start by assessing the areas of concern in the biomarker field, in order to understand what is the problem, the key components of biomarkers' benefits, risks and costs and what needs to be done. After the areas of concern have been defined, the next step is the definition of exclusion criteria, which will allow for the automatic rejection of some of the biomarker options. This is followed by the creation of a list that includes all relevant biomarkers' options to be considered in the decision-making. After the list is complete with all the biomarkers approved after the application of the exclusion criteria, the evaluation criteria to be used in the decision-making process are determined, along with their descriptors of performance and reference levels, which should address the objectives and concerns that experts find fundamental to evaluate the biomarker options. Before the evaluation criteria are confirmed to be preference independent, they are called evaluation dimensions. Therefore, first of all, a list of possible EDs must be created based on literature, and the correspondent descriptors of performance and reference levels *good* and *neutral* must be defined. Secondly, this list will be presented to a few selected experts and an interview will be conducted with the purpose of changing, adding or eliminating any ED. Finally, in the form of a web-based Delphi, a larger group of experts will be presented with the updated list, where they can give individual qualitative judgements regarding the relevance of each proposed ED, in order to reach an agreement about the EDs to be used. After this,

a test on preference dependence must be conducted to verify if all EDs are independent, followed by a restructuring of the model in case dependencies are found, resulting in a final list of independent EDs, now called criteria. The chosen criteria are then assigned to one area of impact (benefits, costs or risks) and, finally, a value tree is designed, with the criteria divided according to their area of impact.

After the model structuring is finished, we can advance to the model building phase, which starts with the building of value functions and weighting coefficients, both resorting once again to the web-based Delphi social approach, in order to collect the necessary points of view from experts and stakeholders regarding differences of attractiveness between the biomarkers' criteria and between the levels of performance of those criteria. The web-Delphi results must then be further analysed and discussed in a decision conference, for the model to be built. Value functions are essential, as they are responsible for the conversion of performance into value on each criterion [47].

Before starting the last phase, it is necessary to acquire the necessary information for each biomarker being considered. This information will be used to attribute a level of performance to the biomarker in each evaluation criteria previously defined.

Finally, the model testing and validation phase is initiated. Firstly, using the respective value functions obtained previously, the several biomarker options are evaluated on each criterion, with the purpose of converting the performance of each option into a partial value score. Secondly, a global evaluation of the options is performed, in order to aggregate the value scores of each option across the evaluation criteria. After this, the requisiteness of the model is tested, by executing a robustness and sensitivity analysis of the results obtained [47]. During this phase it is necessary to conduct decision conferences to analyse and discuss the results and, in the end, for the model to be validated, it must be discussed and approved in a decision conference, by the experts involved.

After the model is validated, we obtain our outputs, which include performance profile, value in each evaluation criteria and value in each area of concern, from which we can obtain an ordered list of the biomarkers, according to their relevance. The four or five most relevant ones are then presented to the experts and stakeholders, allowing for the investigation team to continue their work, with the goal of validating and qualifying the selected biomarkers so that, in the future, they can be utilized in clinical care, in this specific case, of COPD.

By following all the steps mentioned along this section, the expected model can be obtained. However, the focus of this thesis will be the structuring of the model, as represented in Figure 4.2, because it was decided that it was more important to achieve a solid model structuring, than a poor complete model, as a result of the reduced time available.

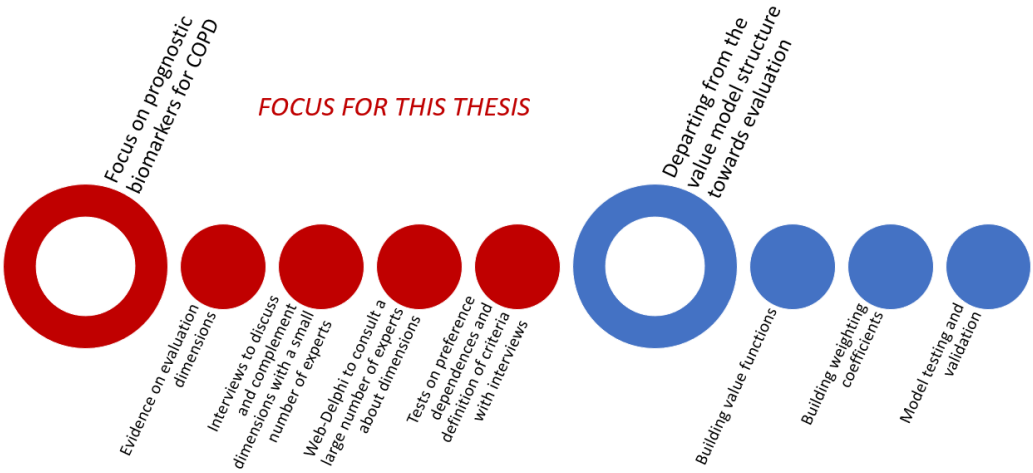

Figure 4.2: Methodology for reaching an MCDA model to prioritize biomarkers, with focus on the model structuring phase.

4.1.1 The MACBETH Method

MACBETH is a MCDA method that uses qualitative terms (non-numerical) for decision-makers to make judgements regarding differences in attractiveness in pairwise comparisons, with the purpose of determining scores for options and weights for criteria in a multicriteria model [34]. In order to apply this method, the software M-MACBETH is going to be used as a decision support system.

In this subsection, the steps and concepts that are relevant for the application of the MACBETH method will be explored.

Criteria

A Point of View (PV) consists of any aspect, from objectives to concerns, indicators, characteristics, attributes and restrictions considered relevant to the evaluation of proposals. The purpose of a criterion is to evaluate proposals in terms of a certain Fundamental Point of View (FPV), which can be a singular point of view, or be composed by several points of view [46]. In this thesis, we will be using the term dimensions (before criteria have been proven to be independent) and criteria (after the criteria have been proven to be independent).

During the development of a multicriteria model, the first thing to do, after assessing the problem and the areas of concern, is defining the criteria. There are two types of criteria: exclusion (or screening) criteria, which allow for an immediate rejection of non-valid options in the software, and evaluation criteria. Each criterion must meet the following properties [46]:

- **Isolable**, guaranteeing independence among the criteria;
- **Intelligible**, facilitating the understanding of the criterion;
- **Operational**, allowing for a good functioning of the additive model;
- **Consensual**, representing the concordance between the decision-maker and the facilitator regarding the defined criterion.

As for a family of criteria to be coherent, it must be [46][72]:

- **Complete (Exhaustive)**, including all the fundamental characteristics for the evaluation of the available options;
- **Non redundant**, not indicating the consequences of other criteria;
- **Concise**, considering only the hypothesis relevant to the model;
- **Consensual**, representing the concordance between the decision-maker and the facilitator;

To facilitate the structuring of the problem, a value tree is designed, which includes the goal and the criteria. For informative purposes, non-criteria nodes can be added, working as categories where the criteria can be divided into, but do not influence the decision [34].

Descriptors of Performance

After the definition of the criteria, descriptors of performance must be defined for each criterion. A descriptor of performance is a set of plausible levels of performance, ordered by preference, in terms of a criterion, that is essential to make a criterion operational. There are several types of descriptors, divided into three categories depending on their respective nature, and each descriptor is described by one of the elements of each category [46]:

- **Quantitative, qualitative or pictorial:** A quantitative descriptor uses numerical values, a qualitative one uses semantic and numerical expressions, and a pictorial one uses visual representations to describe the levels of performance;
- **Continuous or discrete:** A continuous descriptor allows for infinite levels of performance, while a discrete one is represented by finite levels of performance;
- **Direct, indirect or constructed:** The direct and indirect descriptors depend on whether the levels reflect, directly or not, the ends. A direct descriptor is associated with a natural criterion that reflects the ends, while an indirect one counts on a set of levels defined by a combination of several indicators related with the criterion. As for a constructed descriptor, it is a group of reference levels defined by an index or an holistic combination of several indicators, integrated in the criterion or sub-criterion pretended.

If it is possible to define direct descriptors, it is the most adequate option, since the more objective a descriptor is, the least ambiguous the criterion is, leading to a less controversial evaluation model [46]. Also, to avoid redundancy, each descriptor should be associated with only one criterion [73].

Reference Levels

When building an evaluation model, it is necessary to define two reference levels for each descriptor, which are required for benchmark analysis and the building of weighting coefficients. The reference levels to be used are *good* and *neutral*, representing the 0 and the 100 of the value scale, respectively: the *good* level expresses what is a good performance in each criterion, while the *neutral* level expresses a level of performance that is neither attractive nor unattractive, but is acceptable [70].

The identification of the *good* and *neutral* levels contributes to increase the intelligibility of the criteria and makes it possible to objectify the notion of intrinsic activity of each proposal, affecting it in one of the following categories[46]:

- **Very positive proposal**, if it is, at least, as attractive as a *good* fictitious proposal (for example, a biomarker with high clinical relevance).
- **Positive proposal**, if it is, at least, as attractive as a *neutral* fictitious proposal, but less attractive than a *good* fictitious proposal (for example, a biomarker with moderate/low clinical relevance).
- **Negative proposal**, if it is less attractive than a *neutral* fictitious proposal (for example, a biomarker with no relevance and prejudicial effects).

Value Scales

After defining the criteria and respective descriptors of performance and reference levels, the value scales are built, using pairwise comparisons. For this purpose, the facilitator asks the decision-makers if, considering a specific criterion, there is a difference of attractiveness between two options. If the answer is affirmative, the decision-makers are, first, asked which of the two options is the most attractive, providing ordinal preference information, and, second, are asked to give a qualitative judgement regarding the difference of attractiveness between the two options, providing cardinal preference information. To make the judgments, the decision-makers have to use the semantic scale: null, very weak, weak, moderate, strong, or extreme. With this information, it is possible to order the options according to their attractiveness. The questions are repeated for every pair of options, and with respect to every criterion [34]. The answers to the questions are inserted in matrices in the software M-MACBETH (one matrix for

each criterion) [46]. As each matrix is being filled, a consistency check takes place, in order to confirm if the new entries are consistent with the previously inserted judgements. In case of an inconsistency, the software suggests an alternative. When each matrix is completed and consistent, the qualitative judgements acquired are then used to generate values on an interval scale for each criterion, by means of linear programming, where the least attractive option is given a score of 0 and the most attractive one a score of 100 [34]. The obtained scales can and should be adjusted by the decision-makers, if necessary. After the decision-makers agree with the obtained value scale, it is validated.

Weighting of Evaluation Criteria

After the value scales are built, it is necessary to determine a weight scale, so the criteria can be assigned weights, which is done with a similar procedure as the one used to create value scales.

The determination of the weights must be done with reference to the impact levels of the criteria, with an intrinsic value, such as *good* and *neutral*, so they already have to be determined for each criterion. Otherwise, the weights will be arbitrary, like when the weights are directly determined, considering the intuitive notion of importance [46].

To order the weights of the criteria, the facilitator needs to ask the decision-makers a question such as the following: "Consider a fictitious proposal (N), *neutral* in every criteria. If it is possible to improve the impact from *neutral* to *good* in only one criterion, keeping all the other in the *neutral* level, in which criterion would this improvement be more attractive? And after that?" The repetition of this question until every criterion is chosen, leads directly to the ordering of the weights $k_j (j = 1, ..., n)$ [46]. Next, the decision-makers are asked to make a qualitative judgement, using once again the semantic categories used for the building of value scales (null, very weak, weak, moderate, strong, very strong and extreme), regarding the importance of the previously considered improvement on each criterion. Finally, the importance of improving from *neutral* to *good* is compared on two criteria at a time, using once again the semantic scale. The obtained judgements are inserted into a judgement matrix. Then, using linear programming, the weights are calculated, where 0 is the weight of the *neutral* option and 100 corresponds to the sum of weights [34]. Once again, the decision-makers can and should do the necessary adjustments to the obtained weights and, once they agree with them, the weights are validated.

Global Score

After the model has been built, it is time to test, evaluate and validate it, starting with the determination of the scores for each option, considering both criteria weights and values in each criterion.

Using the additive aggregation model, which is the most used of all multicriteria methods, as it allows not only to order proposals in terms of their global attractiveness, but also to analyze relative differences of global attractiveness, the global value, $V(p)$ of a proposal p, in a family of n criteria, is given by the generic expression [46]:

$$V(p) = \sum_{j=1}^{n} k_j . v_j(p) \tag{4.1}$$

where the parameters k_j are scaling factors or weights (determined previously) that allow to transform partial value units (v_j) into global value units (V). It is also necessary for the additive model to verify the following conditions [46]:

$$\sum_{j=1}^{n} k_j = 1, k_j > 0, \text{ for } j = (1, ..., n) \tag{4.2}$$

$$v_j(\text{good}_j) = 100, \forall j$$
$$v_j(\text{neutral}_j) = 0, \forall j \tag{4.3}$$

The determination of the global score of each proposal leads to the ordering of the proposals according to their attractiveness. The option with the highest global score is the most attractive and should, therefore, be chosen by the decision-makers [34]. The results and the model must then be analysed, discussed and validated in a decision conference.

Sensitivity and Robustness Analysis

After the model has been validated, a sensitivity and robustness analysis should be performed using M-MACBETH, in order to check the strength of the results [34].

The sensitivity analysis has the purpose of evaluating the impact that a change in the weight of a certain criterion has on the final result of the model [34]. As we can see in figure 4.3, the resultant graphics shows the variation of the overall score of each option (in black), as the weight of the criteria under analysis varies from 0 to 100%. The current weight of the criterion under analysis, is represented by the red line. M-MACBETH allows to determine the weight that causes an intersection between the lines of two options, this is, the weight that leads to a tie in the global score between two options [74].

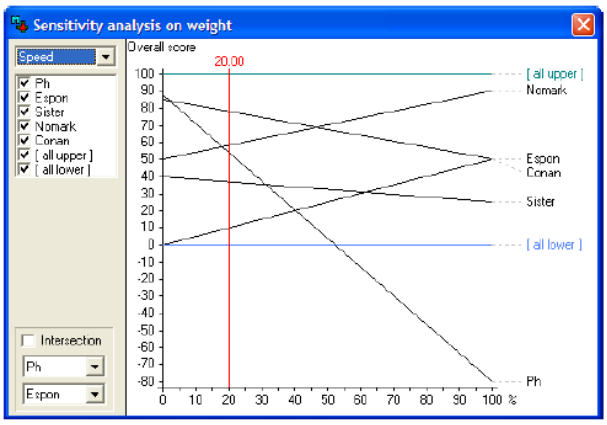

Figure 4.3: Example of a sensibility analysis graphic in the software M-MACBETH [74].

As for the robustness analysis, it allows to evaluate if a variation in the amounts of information, with different degrees of imprecision or uncertainty (such as a variation of $\pm 5\%$ in weight), makes a difference in the result of the best option. Therefore, the robustness analysis is very useful when the decision-making process is associated with uncertainty [34][74].

M-MACBETH organizes information in three types (ordinal, MACBETH and cardinal), and two sections (local and global). Ordinal information corresponds solemnly to ranking, not including information regarding differences of attractiveness, MACBETH information considers the semantic judgements made, but makes no distinction between possible numerical scales compatible with those judgements, and the cardinal information considers the specific scale validated by the decision-makers. As for local information, it includes specific information of each criterion, while global information regards criteria weights. By default, the analysis starts only with local and global ordinal information [74].

The results of this analysis are presented in the form of a matrix, as shown in Figure 4.4, where each entry (ij) relates the options according to the next four symbols [74]:

Δ (Dominance): option i is more attractive in at least one criterion and equally or more attractive in every other criterion, comparatively to j.

$+$ (Additive Dominance): option i is globally more attractive than option j;

$=$ (No dominance between options) - Both options are equally attractive in every criterion.

$?$ (Insufficient information) - The available information is not enough to determine the dominance between options.

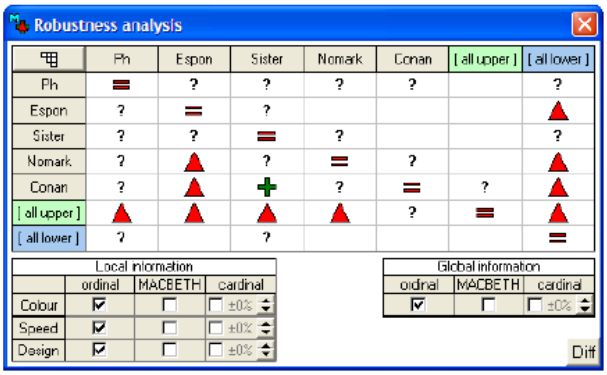

Figure 4.4: Example of a robustness analysis matrix in the software M-MACBETH [74].

4.2 Application in Biomarker Prioritization: Model Structuring

As explained before, the focus of this thesis lies on the first phase of the development of a multicriteria model: the structuring. Therefore, in this section, all the previously presented steps to structure a multicriteria model will be applied for the purpose of this thesis: prioritizing biomarkers.

There are two ways one could approach the problem in hand: through an alternative-focused thinking or a value-focused thinking. While in alternative-focused thinking the alternatives are identified before the objectives are specified, in value-focused thinking the exact opposite happens, as objectives are specified first, followed by the identification of alternatives [75]. In this thesis, we followed a bit of both approaches, as we already had a list of biomarkers alternatives before specifying the objectives (alternative-focused thinking), but new alternatives were added after the objectives have been specified (value-focused thinking).

The decision-maker throughout the entire process was Doctor Deborah Penque. Although several other experts had the opportunity to give their opinions, they were not responsible for making the final decisions. This way, during the entire process several reunions were conducted with the decision-maker who, based on her expert opinion, had to approve all steps of the methodology and all gathered information.

4.2.1 Areas of Concern, Needs in COPD and Context of Use

Based on evidence analysis and with input from the decision-maker, it became clear that the most relevant areas of concern in the biomarker field are benefits for the patient, the associated costs and the eventual risks. Therefore, each of the final evaluation criteria was included in one of these three categories.

As for the needs in COPD, Table 4.1 presents the current needs in function of the biomarker category. As a complement, the context of use is also associated with each need.

Among the seven different biomarker categories, there are different characteristics and objectives. Consequently, comparing biomarkers that belong to different categories is substantially more complex than comparing biomarkers among the same category. Therefore, in this thesis, in order to reduce the

Table 4.1: COPD needs and contexts of use associated with each FDA biomarker category.

Biomarker Categories	Needs in COPD	Context of Use in COPD
Susceptibility/ Risk	Evaluation of (non-synthomatic) individuals at risk for COPD (e.g., smokers/former-smokers, exposed population, aged population, population with genetic risk profile).	Prevention in population at a risk to develop COPD.
Diagnostic	Detect or confirm presence of COPD disease in suspected individuals. Identify patients with a subtype of COPD (e.g, emphysema, chronic bronchitis, ACO-asthma/COPD overlap) for therapeutic decision or clinical trial enrollment.	Improve current diagnostic performance and reduce harm from diagnostic error. Accurate diagnosis that can inform treatment decisions.
Prognostic	Identify likelihood of a clinical event , disease recurrence or progression in COPD patients (e.g., mortality, acute exacerbation, infection).	Clinical event prevention and management.
Pharmacodinamic/ Response	Identify biological response (as an end-point) in COPD patients exposed to a medical product.	Stratify different patient groups in terms of clinical response.
Predictive	Identify COPD patients that are more likely to experience a favorable or unfavorable effect from exposure to a medical product (e.g. corticosteroids).	Patient stratification in drug development or clinical trial enrollment.
Safety	Identify the likelihood of extension of toxicity as an adverse effect in COPD patients before or after exposure to a medical product.	Serially assessment to prevent harm and the early identification of events with a potential to affect the quality of healing actions.
Monitoring	Assess COPD status or COPD medical condition by serially measurements. Assess COPD patients for evidence of exposure to (or effect of) a medical product by serially measurements.	Serially assessment to reduce the risk of unnecessary harm associated with healthcare, delivery by achieving quality care and patient safety. Evaluation of toxicity: hazard identification and dose-response evaluation.

complexity of the model, we opted to focus in only one biomarker category, reducing, as a consequence, the number of candidate biomarkers.

In order to select the biomarker category and respective needs and COU in COPD, which consists on a brief description of the specific use of the biomarker in drug development [76], the following question was considered: ”Which biomarker category, and respective COU, has the largest number of biomarkers investigated, reflecting the greatest clinical needs in COPD, based on evidence from literature?”. After analysing the evidence associated with all the biomarkers previously selected, it was possible to conclude that biomarkers for prognosis of exacerbation and mortality are the most investigated category of biomarkers until this day. Therefore, in this thesis, we focused on prognosis of exacerbation and mortality as the central biomarker category, and its associated COU was to enrich clinical trial/further validation research for an event or population of interest.

4.2.2 Definition of Exclusion Criteria

Before defining the evaluation criteria, it was necessary to define the exclusion (or screening) criteria, which were used to automatically reject some biomarker options.

These criteria were established by the decision-maker and, for biomarker prioritization, they are:

- Biomarkers cannot be validated or in the qualification phase by FDA.
- Biomarkers cannot have a low amount of information associated.

- Biomarkers must be associated with studies that analyse more than one hundred individuals.

Only the biomarkers that were approved after the application of the exclusion criteria were considered in the model.

4.2.3 List of Biomarkers' Options

When this thesis was proposed, there was already a list of over one hundred biomarkers to be considered, obtained from literature published between 2016 and 2018. Despite the number being quite high already, some new biomarkers were added, based on the work of S. Ongay et al. [2], which contained the most relevant biomarkers found in literature until 2016, complementing, therefore, the initial list.

After all possible biomarker options were gathered, they were divided according to their category. Since we chose to work exclusively with prognostic biomarkers, only those were selected. Afterwards, the remaining options were filtered using the exclusion criteria previously defined. As a result, a final list of thirty-two biomarkers was obtained, which can be observed in Table 4.2, alongside with the correlation with COPD and Acute Exacerbations of Chronic Obstructive Pulmonary Disease (AECOPD), from which the most relevant criteria shall be determined in a future work, as this thesis is focused solely in the structuring of the MCDA model.

4.2.4 Definition of Evaluation Criteria, Descriptors of Performance and Reference Levels

For the purpose of defining the relevant criteria to be used in the MCDA model, it was necessary to go through four different phases, in order to obtain the most complete and relevant list possible, taking under consideration not only the literature, but also the opinion of several experts, both clinicians and researchers. The four phases were evidence analysis, interviews, web-based Delphi and test on preference dependence, which were then followed by a fifth phase, where the value tree was built. These five phases are further explored below.

For each considered criterion, it was also necessary to determine the descriptor and levels of performance, the reference levels *neutral* and *good* and the area of impact they belong to (either benefits, costs or risks).

As mentioned before, until an evaluation criterion is confirmed to be independent, it is called an evaluation dimension (ED). Therefore, until the fourth phase is concluded, and all criteria are confirmed to be independent, they will be referred to as EDs.

Phase I: Evidence Analysis

The first step consisted on an evidence analysis, including papers and other documents that either mention what an ideal biomarker should be like or mention necessary evaluation dimensions to have under consideration when selecting biomarkers, through which an initial list of possible EDs was assembled.

After the evaluation dimensions had been defined, it was necessary to define the descriptor and levels of performance and the reference levels *good* and *neutral* for each of them. This was done taking literature into consideration, namely examples of other descriptors of performance and quantitative information, but also common sense and logic.

In sum, the goal of this first phase was to create an initial list of evaluation dimensions and respective descriptors of performance and reference levels to be analysed and evaluated by experts during the remaining phases.

Table 4.2: List of potential COPD biomarkers to be evaluated.

Protein	Correlation with COPD	Reference
Apolipoprotein A-IV	Distinguishing AECOPD from the convalescent state.	[77]
Apolipoprotein C-II	Distinguishing AECOPD from the convalescent state.	[77]
Cystatin C	Hospital mortality risk secondary to COPD exacerbation.	[78]
Collagen type I (fragment) degraded by MMPs	BMI, airflow obstruction, dyspnea, exercise capacity (BODE) index; significant influence on time to exacerbation.	[79]
	Mortality	[80]
Collagen type III (fragment) degraded by ADAMTS	Mortality	[80]
Collagen type III (fragment)	FEV_1% predicted; dyspnea; severe exacerbation	[81]
Collagen type III (fragment) degraded by MMP	FEV_1% predicted; dyspnea; severe exacerbation	[81]
	Mortality	[80]
Collagen type III formation (Pro-peptide)	FEV_1% predicted; slower time of exacerbation; health-related quality of life.	[81]
Collagen type V (Pro-form)	BMI, airflow obstruction, dyspnea, exercise capacity (BODE) index.	[79]
Collagen type VI formation (Pro-peptide)	Mortality	[80]
	FEV_1% predicted; mortality; healthy-related quality of life.	[81]
Collagen Type VI Propetide (Pro-C6) / Collagen type I fragment (C6M) ratio	Higher in COPD vs. control; Higher in patients with previous exacerbation.	[81]
Collagen type IV (fragment)	Mortality	[80]
	BMI, airflow obstruction, dyspnea, exercise capacity (BODE) index; significant influence on time to exacerbation.	[79]
Collagen type VI (fragment) degraded by MMPs	FEV_1% predicted; dyspnea; health-related quality of life	[81]
	Mortality	[80]
Complement component C9	Distinguishing AECOPD from the convalescent state	[77]
Immunoglobulin (free light chains)	Weakly correlated with lung function; mortality in both COPD and A1ATD	[82]
C-terminal peptide of Vasopressin-neurophysin 2-copeptin	Future exacerbations and mortality	[83]
	Mortality risk in AECOPD	[84]
C-reactive protein	Systemic inflammation and COPD outcome	[2]
C-Reactive Protein degraded by MMPs	Mortality	[80]
Elastin (fragment) degraded by neutrophil elastase	FEV_1% predicted; dyspnea	[81]
Immunoglobulin A	Exacerbations	[85]
Interleukin-6	Systemic inflammation and COPD outcome	[2]
Lipopolysaccharide-binding protein	DistinguishinAECOPD from the convalescent state	[77]
Matrix metalloproteinase-9	Mortality	[86]
Matrix metalloproteinase-9 / Tissue inhibitor of metalloproteinases-1	Mortality	[86]
Tissue inhibitor of metalloproteinases-1	Mortality in non-COPD	[86]
Mid-regional pro-adrenomedullin	Early outcomes of COPD	[83]
N-terminal fragment of pro-BNP	All cause of mortality with or without exacerbation	[87]
	Increased hospital length of stay and need for intensive care	[88]
C-C motif chemokine 18	BODE index; higher risk of mortality and exacerbation episodes; prednisolone treatment respons; cardiovascular hospitalization.	[2]
Desmosine (Elastin cross linker)	Cardiovascular comorbidities; aortic stiffness; mortality in patients with COPD; no association with emphysema progression or lung decline.	[89]
Pentraxin 3	Correlated with infection at exacerbation	[90]
Surfactant Protein D	Higher risk for exacerbations and mortality, prednisolone treatment response	[2]
Chitinase like protein	Exacerbations	[91]
	Lung function; bronchodilator response; emphysema	[92]

Phase II: Interviews

Based on the list obtained in phase one, the second phase consisted on interviewing experts, both clinicians and researchers, so they could give their judgments regarding the relevance of the previously defined evaluation dimensions to biomarker prioritization and suggest alterations, if necessary, to improve the way an ED or its levels of performance and reference levels were described and attributed, respectively. Apart from this, experts could also suggest other EDs that they believed to be relevant for the decision-making situation on hand. Therefore, the goal of this second phase was to update the initial list of EDs according to the expert's input.

The choice of experts to participate in this phase is very important, as it defines the quality of the results. The experts must be specialized in an area of knowledge related to the the issue in hand, so both researchers and clinicians must be invited, in order to acquire input from both parts involved in the development of new drugs, which can be useful to detect if there are important EDs missing. Since the purpose of this second phase was only to rule out non relevant EDs, to add new ones that may be missing, and to make the necessary changes to the ones that are accepted, and not to make a final decision regarding which should be considered for biomarker selection, since there was still one other phase left, where the actual decisions would be made based on the opinion of a greater number of experts, we considered that having only two or three experts from each field (researchers and clinicians) participate in this second phase was enough to meet the goal.

The interviews were semi-structured, as the participants were to be asked predetermined questions, but some questions could arise spontaneously in a free-flowing conversation. The way a question or statement is formulated influences significantly this type of study, so the questions must be well thought out. The predetermined questions that were asked during the interview were:

1. "In the document sent to you, there were nine evaluation dimensions that, according to literature, seem to be relevant for the selection of a biomarker among many candidates, to be further investigated with the goal of reaching qualification. Do you consider them all to be relevant? If not, which do you believe to be irrelevant for this purpose?"
2. Do you consider the description of the presented evaluation dimensions to be correct and easily understandable, or do you believe it could be improved? How could it be improved?
3. Do you believe the reference levels were attributed correctly to the levels of performance, or should there be any change?
4. Do you propose any additional evaluation dimension to be considered for biomarker prioritization?

All four questions were open, so the experts could freely give their opinion and suggest new evaluation dimensions to be considered.

To simplify the interview and obtain as much information/opinions as possible, each evaluation dimension was analysed individual and sequentially. Therefore, the first three questions were asked sequentially for each ED, while the fourth question was only asked after all EDs had been covered, leading to the end of the interview.

In order to guarantee that every important aspect regarding each ED was covered, the interviews were done through a video call, and not using an online questionnaire, where the experts would have likely tended to be succinct, leading to a poorer gathering of information.

Phase III: Web-based Delphi

The third phase had the purpose of reaching a semi-final version of the list of EDs being built, based on the judgements of a greater number of experts, with the purpose of reaching a consensual, complete

and relevant group of EDs, since they are of great importance to make the best possible decision when selecting a new biomarker. For this, a web-based Delphi was used, using the platform Welphi and following the rules previously described in section 3.2.3.

In this phase, it is of great importance to include as many experts as possible, both researchers and clinicians, in order to achieve the most inclusive and representative model possible, resulting in more validity and credibility.

The participants were sent an invitation to participate in the Web-Delphi, introducing the problem, explaining the objective of the questionnaire and how it works. Periodically, the experts were sent reminders to encourage participation. In the end, an e-mail thanking the experts for their participation was sent.

The duration of the Web-Delphi is defined before the questionnaire begins, but a round can be extended in case of necessity (due to lack of answers). In the case of this thesis, we defined that each round would have the duration of one week and that twelve was the minimum number of answers acceptable for the first round: therefore, if after the stipulated time for the first round the number of answers was inferior to twelve, the round would be extended and a notification would be sent to all participants.

In a Delphi process, the description of the problem and of the process comes before the questions (in the invitation e-mail and/or in the welcoming page of the questionnaire), which allows the questions to be as simple as possible, making it more direct and easier for the participants to understand. Therefore, the question asked in this particular case was:

1. This evaluation dimension should be considered for the prioritization of prognostic biomarkers in COPD?

The question was then followed by the list of evaluation dimensions obtained in phase II, and the experts participating in the questionnaire were asked to give qualitative judgements regarding each of them, based on the following level scales: strongly agree, agree, neither agree nor disagree, disagree, strongly disagree. The participants were also able to leave comments regarding each of the EDs. As explained in section 3.2.3, when using a web-based Delphi, during the second and third rounds, the participants were encouraged to analyse the anonymous statistics from the previous round and change their answers if they so desired, with the purpose of reaching greater agreement.

After analysing the final answers (obtained after the last round), including not only the percentage of each level scale associated with each evaluation dimensions, but also the comments, and after applying the rules presented in 3.2 to accept or reject evaluation dimensions, an updated and semi-final list was obtained.

Phase IV: Test on Preference Dependence

When all evaluation dimensions had been defined, it was necessary to determine if they were preference independent on each other, in which case they could finally be called criteria, or not, in which case the eventual dependencies had to be resolved.

The first thing that had to be done was reflecting on which evaluation dimensions could be preference dependent on each other, so they could be tested. This process becomes easier by dividing the evaluation dimensions in clusters and reflecting on the probability of preference dependence among the elements of each cluster.

Secondly, it was necessary to define who would participate in this process: at least two individuals with expertise in the biomarker field should participate. The participants had to be interviewed at the

same time, so that answers could be discussed among them, and either face-to-face or via video call, so that doubts could be clarified immediately and more easily.

The interviews were structured, as all questions to be asked were predetermined. The evaluation dimensions were tested in pairs, and questions regarding the four swings presented in Figure 3.4 were be asked. Since we wanted to determine the difference of attractiveness between options of a swing, for each pair of evaluations dimensions a and b, the question asked for each of the swings were:

i. "What is the difference of attractiveness between the option (N_a, N_b) and the option (G_a, N_b)?"

ii. "What is the difference of attractiveness between the option (N_a, G_b) and the option (G_a, G_b)?"

iii. "What is the difference of attractiveness between the option (N_a, N_b) and the option (N_a, G_b)?"

iv. "What is the difference of attractiveness between the option (G_a, N_b) and the option (G_a, G_b)?"

The questions had to be answered using the MACBETH semantic scale: null, very weak, weak, moderate, strong, very strong, extreme.

After the interview, the answers must be analysed, pair by pair, and the result may be one of four: (1) if the participants give different judgments to questions (i) and (ii) and then to questions (iii) and (iv), then the EDs are preference dependent on each other; (2) if the participants give different judgements in questions (i) and (ii), but not in questions (iii) and (iv), then ED a is preference dependent on ED b; (3) if the participants give different judgements in questions (iii) and (iv), but not in questions (i) and (ii), then ED b is preference dependent on ED b: (4) if the answer is the same for questions (i) and (ii) and then for questions (iii) and (iv), then the EDs are preference independent on each other.

In case either scenario (1), (2) or (3) happens, it is necessary to resolve the dependencies. In this thesis, this was done by redefining the dependent EDs by combining them into a single independent criterion. After all dependencies were solved, a final list of relevant criteria for the prioritization of COPD prognostic biomarkers was obtained.

Phase V: Building a Value Tree

The last phase of the definition of evaluation criteria had the purpose of dividing the criteria into the respective areas of concern and, if it made sense, into clusters. Based on this, a value tree was built, using the program M-MACBETH.

The first branches of the tree consist on the areas of concern which, in the case of this thesis, are benefits, costs and risks, the second branches to the clusters, and the criteria associated with each of them are represented as branches that depart from each cluster or area of impact.

Chapter 5

Results

In this chapter, the obtained results will be presented, from the evaluation dimensions found in literature and the changes suffered after the interviews and Web-Delphi, to the final list of criteria and respective descriptors of performance and reference levels.

5.1 Definition of Criteria, Descriptors of Performance and Reference Levels

5.1.1 Phase I. Evidence analysis

The first phase of the process of defining criteria and respective descriptors of performance and reference levels involved a evidence analysis. This phase consisted on the definition of a list of evaluation dimensions based on literature, and subsequent formulation of their descriptors of performance and reference levels based on logic and examples and information found in literature. After this process was completed, Doctor Deborah Penque approved and complemented said list.

As a result, an initial list of evaluation dimensions and the descriptors of performance and reference levels of each of them were defined. The list of EDs obtained is presented on Table 5.1, while the descriptors of performance and reference levels can be observed in Table 5.2.

In Table 5.1, the evaluation dimensions are presented, including their description and the literary references that mentioned them. The EDs ”Expected Utility in Drug Development” and ”Ethical Issues” do not have references associated because these two were proposed by Doctor Deborah Penque as a complement to the the ones found in literature. Initially, there were more EDs (mentioned in literature) considered than the ones presented, including specificity of the results, sensitivity of the results and cost-effectiveness. However, due to the fact that the final criteria must be independent from each other, these EDs were not further considered: the specificity and sensitivity of results are necessary for calculating the AUC, meaning that the ”Clinical Added Value” would depend on both of them, while the cost-effectiveness would depend not only on the costs of validation and utility but also on the clinical relevance and clinical added value.

In Table 5.2, the descriptors of performance (including their type and the levels of performance) and the reference levels *good* and *neutral* associated with each evaluation dimension are presented. All of these were determined based on examples and information found in literature and logic. The quantitative descriptors of performance, associated with the EDs Clinical Added Value and Costs of Validation and Utility required a more through literature investigation in order to reach the values presented in Table 5.2:

Table 5.1: Evaluation dimensions descriptions after evidence analysis.

Evaluation Dimensions	Description	References
Clinical Added Value	New biomarker's added value in relation to the already qualified biomarkers, based on the difference of the AUC between the new biomarker and the already qualified ones.	[24]
Clinical Relevance	Clinical relevance, in terms of benefits to the patient resultant from the application of the biomarker on a clinical level.	[10][24][93][94]
Patient Comfort	Comfort of the patient, considering how invasive the procedure used to access the biomarker is (if it is accessible in the peripheral tissue, namely in blood or urine, the discomfort will be null/minimal; if not the discomfort may be greater [19]).	[19][56][93]
Quality of Evidence	Quantity of quality evidence available regarding the biomarker, considering the number of cohorts, statistical methods used, type of technology used, produced results and validation method.	[24][55][56]
Easiness to Measure, Analyse and Interpret	Easiness to measure and analyze the biomarker, considering the simplicity of the method used, the time needed for this purpose and the need for high-throughput, as well as the easiness to interpret the results.	[10][94]
Reproducibility of Results	Reproducibility of the results, indicating if, using the same methodology several times for a certain biomarker, the results are the same or not (the results are reproducible if they are the same).	[10][93][94]
Expected Utility in Drug Development	Potential to provide valuable information that may reduce uncertainty in regulatory decisions during drug development, allowing for a specific interpretation and application in medical product development ad regulatory review, within the stated COU [95].	—
Costs of Validation and Utility	Amount of money necessary for the validation of a biomarker and utility studies.	[10][56]
Ethical Issues	Ethical issues associated with the application of the biomarker in clinical practice, including psychological reactions (associated with factors such as risk of catastrophic reaction, no proven long-term treatments and risk of false positive) and social stigma [96][97].	—

- **Clinical Added Value:** The performance level of a biomarker is determined by calculating the AUC of the new biomarker and of the already qualified ones (used for the same COU), and comparing them by determining the mean of the difference between the AUC of the new biomarker and the AUC of each of the qualified biomarkers. If the difference is positive, then the new biomarker has added value; if it is negative, the new biomarker presents less value than the already qualified ones, meaning that it does not add value; if it is zero, the two scenarios have the same AUC, meaning that the new biomarker does not add value.

  An ROC curve plots True Positive Rate (TPR), given by equation 5.1, versus False Positive Rate (FPR), given by equation 5.2, which are equivalent, respectively, to sensitivity and 1 - sensitivity. The AUC measures the area under the ROC curve. Its value ranges from 0 to 1, where 0 corresponds to a model whose predictions are 100% wrong and 1 corresponds to a model whose predictions are 100% correct. Therefore, the higher the AUC, and the higher the difference between the AUC of the new biomarker and the already qualified ones, the better [98].

$$TPR = \frac{FalsePositives}{FalsePositives + TrueNegatives} \tag{5.1}$$

$$FPR = \frac{TruePositives}{TruePositives + FalseNegatives} \tag{5.2}$$

  In case the AUC of a biomarker is lower than 50%, it will be considered to be irrelevant. It should be higher than 70% in order to be a good enough biomarker and, ideally, it should be close to 90% (or higher) to be a very good biomarker. Therefore, the levels of performance presented in Table

Table 5.2: Descriptors of performance and reference levels *neutral* and *good* for each evaluation dimension after vidence analysis.

Evaluation Dimensions	Descriptors of Performance
Clinical Added Value	Quantitative, continuous and direct: • 25% • 20% • 15% **[GOOD]** • 10% • 5% • 0% **[NEUTRAL]**
Clinical Relevance	Qualitative, discrete and constructed: • Very relevant, presenting great clinical benefits to the patient. **[GOOD]** • Relevant, presenting significant clinical benefits to the patient. • Irrelevant, not presenting clinical benefits to the patient. **[NEUTRAL]**
Patient Comfort	Qualitative, discrete and constructed: • Non-invasive procedure, with no discomfort to the patient. **[GOOD]** • Semi-invasive procedure, with minimal discomfort to the patient. **[NEUTRAL]** • Semi-invasive procedure, with mild discomfort to the patient. • Invasive procedure, with mild/moderate discomfort to the patient. • Invasive procedure, with severe discomfort to the patient.
Quality of Evidence	Qualitative, discrete and constructed: • Many studies with evidence of high quality. **[GOOD]** • Some studies with evidence of high quality. • Few studies with evidence of high quality. **[NEUTRAL]** • Only studies with evidence of medium quality or lower.
Easiness to Measure, Analyse and Interpret	Qualitative, discrete and constructed: • Easy to measure and analyze the biomarker and easy to interpret the results. **[GOOD]** • Hard to measure and analyze the biomarker and easy to interpret the results. • Easy to measure and analyze the biomarker and hard to interpret the results. **[NEUTRAL]** • Hard to measure and analyze the biomarker and hard to interpret the results.
Reproducibility of Results	Qualitative, discrete and direct: • High reproducibility. **[GOOD]** • Medium reproducibility. **[NEUTRAL]** • Low reproducibility.
Expected Utility in Drug Development	Qualitative, discrete and constructed: • High potential to provide valuable information for drug development. **[GOOD]** • Moderate potential to provide valuable information for drug development. **[NEUTRAL]** • Low potential to provide valuable information for drug development.
Costs of Validation and Utility	Quantitative, continuous and direct: • 4M$ • 8M$ **[GOOD]** • 12M$ • 16M$ • 20M$ **[NEUTRAL]** • 24M$ • 28M$ • 32M$
Ethical Issues	Qualitative, discrete and constructed: • Presents no risk of psychological reactions and/or social stigma to the patient. **[GOOD]** • Presents low risk of psychological reactions and/or social stigma to the patient. • Presents moderate risk of psychological reactions and/or social stigma to the patient. **[NEUTRAL]** • Presents high risk of psychological reactions and/or social stigma to the patient.

5.2 represent the difference between the AUC of the new biomarker and the already qualified ones, assuming that the last ones hardly have an AUC lower than 65%.

- **Costs of Validation and Utility:** According to B. A. Hamilton [99], the costs associated with the development and validation of a biomarker vary according to its category. For the specific case of prognostic biomarkers, in which this thesis is focused on, the overall cost can range from 10 to 50 million dollars, where 40-60% correspond to validation and utility, 20-40% to developed assay and intended use and analytical validation, and the remaining 10-20% to identification and feasibility.

  This subsection of costs (validation and utility) can, therefore, vary from 40%*10M$=4M$ (minimum associated percentage times minimum overall cost) to 60%*50M$=30M$ (maximum associated percentage times maximum overall cost). Therefore, levels of performance ranged from 4M$ to 32M$, with increments of four, in order to include all possible values.

  The reference levels *neutral* and *good* where determined in the following way: the option closest to 60%*10M$=6M$ (maximum associated percentage times minimum overall cost), which is the best case scenario considering the values provided by B. A. Hamilton [99], was considered to be *good*, and the one closest to 40%*50M$=20M$ (minimum associated percentage times maximum overall cost), which is the worst case scenario, was considered to be *neutral*.

  The cost associated with each biomarker has to be estimated. The final value will depend on factors such as the method used, the number of patients to be tested and the number of times the biomarker will be measured for each patient.

5.1.2 Phase II. Interviews

With the initial evaluation dimensions well defined, the second phase of the process started. In this phase, four experts (two researchers and two clinicians, three male and one female, all members of CliniMark) were interviewed via video call, with the purpose of analysing the relevance of each of the previously defined evaluation dimensions, verifying the quality of their descriptions and descriptors of performance and the rightness of the reference levels attributed, suggesting alterations, if necessary, and suggesting new EDs believed to be relevant for the purpose at hand. Before the interviews, each expert was presented with a brief sum of the purpose of the thesis and, more specifically, of the interviews to be held, and were also asked to review a document presenting the chosen COU and the EDs previously defined, so that the interview would go smoothly, and it was easier for each of them to answer the questions asked.

In a general way, the four interviewed experts agreed with the EDs presented, their descriptions and respective descriptors of performance. However, there were several suggestions made and, in some cases, some differences in opinion between the experts. In order to go through every suggestion, each criterion is going to be analysed individually:

1. **Clinical Added Value:** All four experts found this evaluation dimension to be very important.

   It was suggested that the meaning of added value should be better explained. Therefore, the previous definition was rewritten, leading to the following: Clinical Added Value represents the new biomarker's added value (if it brings something new/different) in relation to the already qualified biomarkers, based on the difference of the Area Under the Curve (AUC) between the new biomarker and the already qualified ones.

   The four experts agreed with the method used (AUC) to determine the added value. One expert underlined that this method only works when biomarkers have already been investigated, meaning that it is not "new biomarker" friendly, which is problematic in the sense that a biomarker can

be very good and not have been investigated yet. However, since the idea in this case is to choose the most promising biomarkers from a list of biomarkers obtained in literature, this is not a problem. Another expert mentioned that sensitivity and specificity depend a bit on the test and on the disease: in some cases the benefit may be more by increasing the sensitivity, to avoid missing anybody who is ill, and in other cases one wants a high specificity to avoid false positives. It is always a compromise and should be decided on a case by case basis. It changes with the purpose for which we are selecting the biomarker. However, since we selected a specific COU, the purpose of the biomarkers under consideration is very similar. Taking this into account, and in order to simply things, the AUC and, consequently, the sensitivity and specificity, is a good enough choice to determine the added value of a biomarker.

Regarding the levels of performance and reference levels, one expert said that 10% should be considered a *good* reference level. In some cases, such as life and death situations, even 5% would be good. Therefore, the reference levels were changed. Also, two experts mentioned that an added value of 25% seemed too high. Therefore, that level of performance was eliminated.

2. **Clinical Relevance:** All four for experts found this evaluation dimension to be extremely important. It was suggested that specific benefits should be mentioned in the description, taking under consideration the chosen COU, to facilitate the determination of the level of performance for each biomarker. As a result, the previous definition was rewritten, leading to the following: Clinical relevance is the ability of a biomarker to improve patient well-being and outcome. The identification of the likelihood of a disease event such as disease recurrence or progression, acute exacerbation or infection, leads to changes in patient management, mortality and morbidity.

   As for the levels of performance, two experts suggested changing the term “irrelevant” to “not relevant”/”not relevant in this context”, as the term “irrelevant” is very negative and an irrelevant biomarker can be perceived as completely useless, which is generally not the case (it may be irrelevant in this context, but not in others). Two experts said that the levels of performance should be more specific, or that the first one should be eliminated, because choosing between very relevant and relevant is not easy. To respond to this suggestion, the levels of performance were rewritten, increasing their specificity. These two experts also mentioned that the benefits of a biomarker can vary from patient to patient so, even if a biomarker only presents benefits for a group of people with certain characteristics, it should still be considered, because it does not mean that a biomarker is not good, it only means that it is not a good generic biomarker, but rather a good specific one, which can be very useful for patient stratification (very important in drug development).

3. **Patient Comfort:** All four experts found this evaluation dimension to be relevant.

   One of the experts (clinician), mentioned that on a clinical level only four levels of performance are used: non-invasive procedure with no discomfort, semi-invasive with mild discomfort, invasive with moderate discomfort and invasive with severe discomfort. Therefore, the second level of performance, ”semi-invasive procedure, with minimal discomfort to the patient”, was eliminated, leaving only one level of performance referent to semi-invasive procedures, which became the new *neutral* level. Also, in the previous fourth level of performance, instead of mild/moderate discomfort only moderate discomfort was left.

4. **Level of Evidence (previous Quality of Evidence):** Two experts considered the evaluation dimension ”Quality of Evidence” to be relevant and two said that even though it may be relevant, it may also lead to biases, as people tend to choose biomarkers with a lot of evidence, not giving a chance to the ones that have not been very investigated yet. Also, three experts

said that high/medium/low quality of evidence should have a description associated and that some/many/few are not very clear terms.

In order to reduce the biases and to solve the problems presented, a suggestion made by one of the experts has been followed: instead of “Quality of Evidence”, this criterion should be called “Level of Evidence”, and each study should be classified according to its level of evidence.

As a result, the whole ED was redefined, including not only the definition (which can be checked in Table 5.3) but also the levels of performance and reference levels (which can be checked in Table 5.4).

5. **Quality of the Study** After changing the evaluation dimension ”Quality of Evidence” to ”Level of Evidence”, Doctor Deborah Penque suggested a new ED as a complement to the previous one: the Quality of the Study. According to the *Johns Hopkins Nursing Evidence-Based Practice* guide [100], evidence can not only be classified in levels (I, II, III and IV), but also using a quality guide.

   The high quality of a study is based on i) well-designed cohort, including adequate biospecimen repositories and sample size, ii) proper analytical and statistical methods and iii) accurate data analysis and interpretation to enable biomarker identification/validation meeting pre-specified performance criteria for a given clinical application context.

   Based on the recommendations presented on the description above, the descriptor of performance (qualitative, discrete and constructed), and the reference levels were also defined, and can be checked in Table 5.4.

6. **Easiness to Measure, Analyse and Interpret:** All four experts found this evaluation dimension to be very well defined.

   Two of them found it relevant, while the other two said that, in comparison to the benefit, this ED was not very significant. However, in case of a tie, it is always best to have a biomarker easy to measure, analyze and interpret. Therefore, they concluded that the easiness to measure, analyze and interpret had some relevance.

   Since all experts found this ED to be very well defined, there was no change on the description, descriptor of performance or reference levels.

7. **Test Reliability (previous Reproducibility of Results):** All four experts found “Reproducibility of Results” to be a relevant evaluation dimension. Additionally, two of them said that the reliability of the test should also be considered. However, reliability concerns the extent to which any experiment/test leads to the same results on repeated trials and, if the reliability of a test is high, then that test is accurate, reproducible and consistent from one trial to another [101].

   Since the reliability depends on the reproducibility, we opted to replace the second with the first, as it is more complete.

   As a result, the whole ED was redefined, including not only the definition (which can be checked in Table 5.3) but also the levels of performance and reference levels (which can be checked in Table 5.4).

8. **Potential Value to Address an Unmet Need in Drug Development (previous Expected Utility in Drug Development:** All four experts found this evaluation dimension to be relevant.

   Two of the experts thought this ED was well defined and that it was not very difficult to choose a level of performance since there were only three levels. However, the remaining two experts said that it was very subjective and should be better defined (because the potential depends on more than the biomarker), and that it would be hard to make a choice.

As a result, the whole ED was rewritten, including the name, with the purpose of reaching a less subjective ED, with better defined levels of performance. The new definition can be checked in Table 5.3 and the new levels of performance and reference levels in Table 5.4

9. **Costs of Development (previous Costs of Validation and Utility):** All four experts found this evaluation dimension to be very important.

   It was suggested that the ED name should be "Costs of Development", as the development phase includes both the analytical validation and the utility studies.

   It was also suggested that the description should be a bit clearer about the meaning of validation and utility studies. Therefore, the previous definition was rewritten, leading to the following: Amount of money necessary for the validation of a biomarker (where the generalizability across different samples and the reproducibility and standardization of the assay are determined) and utility studies (where the results must show performance characteristics, well-designed experiments and the added value in research models and/or patients) [12].

   Regarding the levels of performance, one expert said that the gaps between levels of performance should be bigger, suggesting: 5M$, 10M$,..., which was applied. Three experts found that the values presented and the way the reference levels were defined makes sense. However, they pointed out that the cost of development and, more specifically, validation and utility of a biomarker, can range even more than what was considered. The remaining expert did not agree with the proposed values because the costs can vary a lot, as there are many factors influencing them, including the context of business, the method used, the country the study is being held in, the fluid one wants to measure the biomarker in (spinal fluid will cost substantially more than blood/urine), the number of people to be involved in the study and even the depth of detail required. However, since the majority agreed with the proposed values and the cost will be estimated for each biomarker (where factors such as the country will be the same for all biomarkers), there will be no change in the considered values.

10. **Ethical Issues:** Three out of four experts found this evaluation dimension to be very relevant. However, the remaining one believes that it is not very relevant for biomarker selection. Since the majority found it to be relevant, it was not eliminated and shall be further evaluated and judged by a greater number of experts during Phase III.

As a result of the interviews, the initial list of evaluation dimensions and the descriptors of performance and reference levels were updated. The list of EDs obtained after processing the information gathered in the interviews is presented in Table 5.3, while the descriptors of performance and reference levels can be observed in Table 5.4.

5.1.3 Phase III. Web-based Delphi

After the interviews were done and the data collected was processed, the third phase of the definition of criteria and respective descriptors of performance and reference levels began.

Contrarily to the two previous phases, this phase was designed with the sole purpose of determining the relevance of the previously defined evaluation dimensions, based on the opinion of a great number of experts, instead of defining and improving the EDs, descriptors of performance and reference levels.

In this phase, regardless of the round, the participants were asked to evaluate the relevance of the ED obtained after the interviews, which were inserted in the web-based Delphi as indicators, each with a description associated containing not only the description of the ED in question, but also the descriptors

Table 5.3: Evaluation dimensions descriptions after interviews were made to experts.

Evaluation Dimensions	Description
Clinical Added Value	Clinical Added Value represents the new biomarker's added value (if it brings something new/different) in relation to the already qualified biomarkers, based on the difference of the AUC between the new biomarker and the already qualified ones.
Clinical Relevance	Clinical relevance is the ability of a biomarker to improve patient well-being and outcome. The identification of the likelihood of a disease event such as disease recurrence or progression, acute exacerbation or infection, leads to changes in patient management, mortality and morbidity.
Patient Comfort	Comfort of the patient, considering how invasive the procedure used to access the biomarker is (if it is accessible in the peripheral tissue, namely in blood or urine, the discomfort will be null/minimal; if not the discomfort may be greater [19].
Level of Evidence	Level of evidence (or hierarchy of evidence) are assigned to studies based on the methodological quality of their design, validity, and applicability to patient care. These decisions give the "grade (or strength) of recommendation". In Biomarker research, the level of evidence can be classified as (adapted from Kisser & Zechmeister-Koss [102]): • Highest (Level I) - Evidence from systematic review of Level II studies (well-designed randomized controlled trial (RTC)); • High (Level II) - Evidence from at least one well-designed RTC/prospective cohort study (Phase 2/3 explanatory study); • Moderate (Level III) - Evidence from non-randomized RCT/cohort study (Phase 1 explanatory study)/follow-up study; • Low (level IV) - Evidence from a case series / case control studies / historically controlled studies.
Quality of the Study	A high quality of a study is based on: i) well-designed cohort, including adequate biospecimen repositories and sample size; ii) proper analytical and statistical methods; iii) accurate data analysis and interpretation to enable biomarker identification/validation meeting pre-specified performance criteria for a given clinical application context
Easiness to Measure, Analyse and Interpret	Easiness to measure and analyze the biomarker, considering the simplicity of the method used, the time needed for this purpose and the need for high-throughput, as well as the easiness to interpret the results.
Test Reliability	Consistency or reproducibility of biomarker results across time (test-retest reliability), across items (internal consistency), and/or across researchers (inter-rater reliability) [103].
Potential Value to Address an Unmet Need in Drug Development	Potential value of a biomarker for a particular context of use (glsCOU) in drug development, likely to be recognized by the FDA as qualified biomarker in the near future. COU in drug development includes patient stratification, selection, trial enrichment, dose selection, response/efficacy assessments, safety assessments, among others.
Costs of Development	Amount of money necessary for the validation of a biomarker (where the generalizability across different samples and the reproducibility and standardization of the assay are determined) and utility studies (where the results must show performance characteristics, well-designed experiments and the added value in research models and/or patients) [12].
Ethical Issues	Ethical issues associated with the application of the biomarker in clinical practice, including psychological reactions (associated with factors such as risk of catastrophic reaction, no proven long-term treatments and risk of false positive) and social stigma [96][97].

Table 5.4: Descriptors of performance and reference levels *neutral* and *good* for each evaluation dimension after interviews were made to experts.

Evaluation Dimensions	Descriptors of Performance
Clinical Added Value	Quantitative, continuous and direct: • 20% • 15% • 10% **[GOOD]** • 5% • 0% **[NEUTRAL]**
Clinical Relevance	Qualitative, discrete and constructed: • Very relevant, showing ability to improve well-being and outcome to the generality of patients. **[GOOD]** • Relevant, showing ability to improve well-being and outcome to patients with specific characteristics. • Irrelevant, showing poor or no ability to improve patients' well-being and outcome. **[NEUTRAL]**
Patient Comfort	Qualitative, discrete and constructed: • Non-invasive procedure, with no discomfort to the patient. **[GOOD]** • Semi-invasive procedure, with mild discomfort to the patient. **[NEUTRAL]** • Invasive procedure, with mild/moderate discomfort to the patient. • Invasive procedure, with severe discomfort to the patient.
Level of Evidence	Qualitative, discrete and direct: • Highest Evidence Level (I). • High Evidence Level (II). **[GOOD]** • Moderate Evidence Level (III). **[NEUTRAL]** • Low Evidence Level (IV).
Quality of the Study	Qualitative, discrete and constructed: • High Quality: all recommendations are fully accomplished. **[GOOD]** • Medium Quality: one/two recommendations are not fully accomplished. **[NEUTRAL]** • Low Quality: no recommendation is accomplished.
Easiness to Measure, Analyse and Interpret	Qualitative, discrete and constructed: • Easy to measure and analyze the biomarker and easy to interpret the results. **[GOOD]** • Hard to measure and analyze the biomarker and easy to interpret the results. • Easy to measure and analyze the biomarker and hard to interpret the results. **[NEUTRAL]** • Hard to measure and analyze the biomarker and hard to interpret the results.
Test Reliability	Qualitative, discrete and direct: • High reliability. **[GOOD]** • Medium reliability. **[NEUTRAL]** • Low reliability.
Potential Value to Address an Unmet Need in Drug Development	Qualitative, discrete and constructed: • Potential high-value for a particular COU in drug development with high unmet need. **[GOOD]** • Potential high/moderate-value for a particular COU in drug development. • Potential low-value for a particular COU in drug development. **[NEUTRAL]**
Costs of Development	Quantitative, continuous and direct: • 5M$ **[GOOD]** • 10M$ • 15M$ • 20M$ **[NEUTRAL]** • 25M$ • 30M$
Ethical Issues	Qualitative, discrete and constructed: • Presents no risk of psychological reactions and/or social stigma to the patient. **[GOOD]** • Presents low risk of psychological reactions and/or social stigma to the patient. • Presents moderate risk of psychological reactions and/or social stigma to the patient. **[NEUTRAL]** • Presents high risk of psychological reactions and/or social stigma to the patient.

of performance and reference levels associated so that the experts participating could inform themselves before deciding on its relevance.

One hundred and eighteen experts (researchers and clinicians) were invited to participate in this web-based Delphi, all members of CliniMark.

Round One

In round one, experts only had to answer the question posed, using the scale levels available to evaluate the relevance of each evaluation dimension.

The round started on August 5, 2019, and lasted until August 23, 2019 (almost three weeks). For this phase, one hundred and eighteen experts (researchers and clinicians) were invited. Ideally, all participants would answer, but realistically, we defined that twelve was the minimum number of participants acceptable, as the number of participants tends to decreased from one round to the other, so it was essential to define a number high enough to try and guarantee that there were some participants left to answer in the third round. We considered that less than twelve participants was too little and too risky.

Initially, the round was planned to last one week. However, due to the low number of participants, it was necessary to extend the round twice: during the first week, there were only three answers, which led to an extension of one week; during the second week the number of participants increased to nine, which led to another extension of one week; during the final week, fifteen people answered the questionnaire, resulting in a total of 24 participants.

During the three weeks the round lasted, two reminders were sent per week to every expert that had not participated yet, in order to try and encourage them to participate. The reminders were clearly very effective, as almost all the were given in the days that a reminder had been sent.

After the third week, the round was closed, not only because we had reached a satisfactory number of participants, although not ideal, but also due to lack of time. However, if the round had lasted longer, lasting until September, it is likely that the number of participants would have increased significantly, as many of the contacted individuals were probably on vacation while the round was open.

In the end of the first round, there was a total of twenty-four participants, resulting in a 20% adhesion. Among the participants, 63% were women and 37% men, while 71% were researchers and 29% were clinicians, all from different geographic locations.

Indicador	Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree
1 - Clinical Added Value	63%	33%	4%		
2 - Clinical Relevance	75%	25%			
3 - Patient Comfort	29%	42%	25%	4%	
4 - Level of Evidence	58%	38%	4%		
5 - Quality of the Study	67%	33%			
6 - Easiness to Measure, Analyse and Interpret	33%	38%	17%	13%	
7 - Test Reliability	75%	25%			
8 - Potential Value to Address an Unmet Need in Drug Development	29%	46%	21%	4%	
9 - Costs of Development	8%	58%	21%	8%	4%
10 - Ethical Issues	38%	46%	17%		

Figure 5.1: Results of the first round of the Web-Delphi to determine the relevant evaluation dimensions for biomarker prioritization.

As expected, the results obtained in the first round, which can be observed in Figure 5.1, were not very consensual in some cases. For example, the evaluation dimension Costs of Development led to a great difference of opinions between experts, with some participants saying that they strongly agreed that it was a relevant evaluation dimension, while others disagreed/strongly disagreed. In the other hand, there were some evaluation dimensions, such as Clinical Relevance, Quality of the Study and Test Reliability where the opinion was consensual between experts, who either agreed or strongly agreed with their relevance for biomarker prioritization.

The evaluation dimensions that presented consensus among experts could have been approved in this first round. However, we opted to only approve or reject an evaluation dimension after the third round. This way, all evaluation dimensions were reevaluated in round two and three in order to try and reach consensus.

Round Two

In round two, participants were asked to review and analyse the statistics from the previous round and rethink their opinion, having the chance to alter their answer if they so desired.

The round started on August 24, 2019, and lasted until September 6, 2019. In this round, only the experts that participated in round one were invited, meaning that only twenty-four experts received an invitation. During the two weeks the round lasted, two reminders were sent per week to every expert that had not participated yet.

In the end of the round, there was a total of twenty-two answers, resulting in 92% adhesion. Among the participants, 59% were women and 41% men, while 77% were researchers and 23% were clinicians, all from different geographic locations.

Indicador	Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree
1 - Clinical Added Value	82%	18%			
2 - Clinical Relevance	82%	18%			
3 - Patient Comfort	27%	41%	27%	5%	
4 - Level of Evidence	68%	27%	5%		
5 - Quality of the Study	77%	23%			
6 - Easiness to Measure, Analyse and Interpret	27%	45%	18%	9%	
7 - Test Reliability	82%	18%			
8 - Potential Value to Address an Unmet Need in Drug Development	27%	59%	9%	5%	
9 - Costs of Development	5%	68%	18%	5%	5%
10 - Ethical Issues	36%	55%	9%		

Figure 5.2: Results of the second round of the Web-Delphi to determine the relevant evaluation dimensions for biomarker prioritization.

As expected, the results obtained in the second round, which can be observed in Figure 5.2, were more consensual than in round one, although in some cases there is still a clear difference in opinions. For example, in the case of the evaluation dimension Costs of Development, the percentage of participants answering *Strongly Agree*/*Agree* increased 7%, decreasing in both *Not Agree nor Disagree* and *Disagree*, meaning that the generic opinion started converging to the positive spectrum, although one expert still chose the option *Strongly Disagree* regarding the relevance of this evaluation dimension in

biomarker prioritization. In the other hand, in the case of the evaluation dimension Clinical Added Value, the opinion was unanimous that it is indeed relevant for biomarker prioritization, as all the answers were either *Strongly Agree* (great majority, with 82% of the votes) or *Agree*, contrarily to the previous round.

In the end of the round, four evaluation dimensions clearly stood out due to the fact that every single participant considered them to be relevant for biomarker prioritization: Clinical Added Value, Clinical Relevance, Quality of the Study and Test Reliability. However, three EDs may still be eliminated after the third round, due to lack of consensus: Patient Comfort, Easiness to Measure, Analyse and Interpret and Costs of Development.

Round Three

During the third and last round, participants were once again asked to review and analyse the statistics from the previous round and rethink their opinion, having the chance to alter their answer if they so desired, with the purpose of reaching more agreement.

The round started on September 7, 2019, and lasted until September 18, 2019. In this round, only the experts that participated in round two were invited, meaning that only twenty-two experts received an invitation. During the week and a half the round lasted, a total of four reminders were sent to the experts that had not participated yet.

In the end of the round, there was a total of twenty-one answers, resulting in 95% adhesion. Among the participants, 62% were women and 38% men, while 76% were researchers and 24% were clinicians, all from different geographic locations.

Indicador	Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree
1 - Clinical Added Value	86%	14%			
2 - Clinical Relevance	95%	5%			
3 - Patient Comfort	24%	52%	19%	5%	
4 - Level of Evidence	76%	24%			
5 - Quality of the Study	86%	14%			
6 - Easiness to Measure, Analyse and Interpret	29%	52%	14%	5%	
7 - Test Reliability	86%	14%			
8 - Potential Value to Address an Unmet Need in Drug Development	19%	71%	10%		
9 - Costs of Development	5%	76%	14%	5%	
10 - Ethical Issues	33%	57%	10%		

Figure 5.3: Results of the third and last round of the Web-Delphi to determine the relevant evaluation dimensions for biomarker prioritization.

As expected, the results obtained in the third and last round, which can be observed in Figure 5.3, were significantly more consensual than in the two previous rounds, as opinions tended to change from indifference and disagreement to agreement regarding the relevance of each evaluation dimension: for example, the Costs of Development had at least one expert choosing each scale level during the first two rounds, while in the third round *Strongly Disagree* had no votes; also, the percentage of agreeable votes started with 66% in round one and ended in 81% in round three. The improvement in consensus throughout the three rounds will be further analysed later.

Due to the qualitative nature of the scale levels applied in the process, it is not possible to determine

the mean, median and interquartile range of the opinions for each evaluation dimension. However, it is possible to determine the mode:

- Clinical Added Value, Clinical Relevance, Level of Evidence, Quality of the Study and Test Reliability have the scale level *Strongly Agree* as mode.
- All the remaining evaluation dimensions have the level scale *Agree* as mode.

After round three, the web-Delphi process was finished. Therefore, the rules of approval and rejection of evaluation dimensions presented in Figure 3.2 were finally applied to determine which evaluation dimensions were approved by the participants for the prioritization of prognostic biomarkers associated with COPD. According to those rules, an evaluation dimension will be approved if one of the two following scenarios happens:

1. The percentage of *Strongly Agree* votes is superior to 50% and the percentage of *Disagree* and *Strongly Disagree* votes is inferior to 33%.
2. The percentage of *Strongly Agree* votes, plus the percentage of *Agree* votes, is superior to 75%.

Table 5.5: Experts' opinions regarding evaluation dimensions for the prioritization of biomarkers in the web-Delphi, with focus on the percentage of votes for Strongly Agree (SA), Strongly Agree plus Agree (SA+A) and Strongly Disagree plus Disagree (SD+D).

Evaluation Dimension	SA (%)	SA+A (%)	SD+D (%)
Clinical Added Value	86	100	0
Clinical Relevance	95	100	0
Patient Comfort	24	76	5
Level of Evidence	76	100	0
Quality of the Study	86	100	0
Easiness to Measure, Analyse and Interpret	29	81	5
Test Reliability	86	100	0
Potential Value to Address an Unmet Need in Drug Development	19	90	0
Costs of Development	5	81	5
Ethical Issues	33	90	0

In Table 5.5 all the necessary percentages to evaluate each evaluation dimension are presented. Based on this values, the following conclusions can be drawn:

- Clinical Added Value, Clinical Relevance, Level of Evidence, Quality of the Study and Test Reliability had more than 50% of experts strongly agreeing with their relevance to prognostic biomarker prioritization in COPD, and none disagreeing, meaning that these five evaluation dimensions are automatically approved.
- Potential Value to Address an Unmet Need in Drug Development and Ethical Issues did not have more than 50% of experts strongly agreeing with their relevance to prognostic biomarker prioritization in COPD, but had 90% either agreeing or strongly agreeing, and none disagreeing. Therefore, they are also approved.
- Patient Comfort, Easiness to Measure, Analyse and Interpret and Costs of Development present some lack of consensus as the opinions given by experts range from *Strongly Agree* to *Disagree*. However, more than 75% of experts either agreed or strongly agreed with their relevance to prognostic biomarker prioritization in COPD, and only 5% disagreed. Therefore, even though the opinion is not consensual, the great majority of experts believed these evaluation dimensions to be relevant, leading to their approval.

As one cannot assume that the results of the Web-base Delphi are one hundred percent correct/trustworthy, since the participants do not interact with each other and no numerical judgements are obtained, Figure 5.3 was presented to Doctor Deborah Penque, who did not participate in the process, so she could decide if, based on the results and in her expert opinion, every evaluation dimension, specially the ones that raised more doubt, should actually be considered as relevant. In her opinion, since we are in an exploratory phase, we should accept all evaluation dimensions, including the three that raise some doubt, because even those presented a high percentage of acceptance.

As a result, ten out of ten proposed evaluation dimensions were considered to be relevant for the prioritization of prognostic biomarkers in COPD by the twenty-one experts participating during the entire web-based Delphi process, plus Doctor Deborah Penque.

Consensus

Being able to change or maintain their opinions after analysing the answers of the other participants, led to an increased agreement between participants throughout the three rounds. Three rounds proved to be enough to reach consensus between the participants.

Since the Delphi is an anonymous process, we did not have access to what which participant answered, meaning that it is not possible to analyse the evolution of opinion of each expert, or to verify if the opinion varied with the field of expertise and geographical location. This being said, it is possible to analyse the evolution of general consensus throughout the three rounds of the Web-Delphi. Table 5.6 presents the percentage of experts who chose either *Agree* or *Strongly Agree* as the scale level for each evaluation dimension in each round.

Table 5.6: Evolution of general consensus throughout the three rounds of the Web-Delphi, regarding the scale levels *Agree* and *Strongly Agree*.

Evaluation Dimension	$1^{st} \rightarrow 2^{nd} \rightarrow 3^{rd}$ round (%)
Clinical Added Value	96% → 100% → 100%
Clinical Relevance	100% → 100% → 100%
Patient Comfort	71% → 68% → 76%
Level of Evidence	96% → 95% → 100%
Quality of the Study	100% → 100% → 100%
Easiness to Measure, Analyse and Interpret	71% → 72% → 81%
Test Reliability	100% → 100% → 100%
Potential Value to Address an Unmet Need in Drug Development	75% → 86% → 90%
Cost of Development	66% → 73% → 81%
Ethical Issues	84% → 91% → 90%

The evaluation dimensions Clinical Relevance, Quality of the Study and Test Reliability were considered to be relevant for the prioritization of COPD biomarkers by all the experts, throughout all the three rounds.

The opinion of the experts was not unanimous initially regarding the evaluation dimensions Clinical Added Value and Level of evidence. However, in round three, consensus was reached, as every single expert considered these EDs to be relevant for biomarker prioritization.

The evaluation dimensions Ethical Issues and Potential Value to Address an Unmet Need in Drug Development did not have 100% of agreement in any round. However, the percentage of experts considering it to be relevant was high in every round, specially in the last one, where 90% of experts agreed with their relevance in biomarker prioritization.

The evaluation dimensions Easiness to Measure, Analyse and Interpret and Costs of Development did not reach consensus until the last round. 10% and 15% of experts, respectively, changed their

opinion to either *Strongly Agree* or *Agree* between round one and round three, leading to 81% of the votes being in one of this two level scales and guaranteeing agreement among participants.

The remaining evaluation dimension, Patient Comfort, suffered few changes throughout the three rounds. In fact, the difference between the first and last round is only 5%, meaning that probably only one expert changed his opinion. In the end, 76% of experts chose either the scale level *Strongly Agree* or *Agree* for this ED, which is enough to consider that there was consensus among participants and to accept the evaluation dimension, although there is some uncertainty associated, since the value is really close to the defined threshold (75%).

Throughout the three rounds of the web-based Delphi, no new evaluation dimension was suggested, but experts left a few comments regarding the relevance of the ones proposed. Based on this comments, a few points are worth mentioning:

- The ED Potential Value to Address an Unmet Need in Drug Development could be included in the ED Clinical Added Value, if the later was considered in a wider sense.
- In such a devastating disease as COPD, patient discomfort has to be accepted.
- Easiness to measure is not a priority, as there are many clinical tests involving complex procedures. However, the results should be easy to analyse and interpret.
- Ethics are important but open for discussion, as they depend on the study. They should be compatible with innovative research and development programs.
- Costs of development do matter, as some potential applications may be just too expensive to pursue. However, there are really expensive tests that are routinely used. It depends on the benefit.

According to experts, one as to be prepared to pay more for biomarker development, go through some invasive procedures, or have a harder time measuring, analysing and interpreting the results, if the clinical relevance/benefit is worth it.

Although the higher the number of participants the better, there was a solid number of participants throughout the three phases, with only three drop-outs out of twenty-four participants (13% drop-out rate), resulting in an inclusive and representative and, consequently, valid and credible model. The rigor was maintained throughout the whole web-Delphi process, as the response rate surpassed 90% in both round two and three. During the whole web-Delphi process, reminders were sent regularly to experts who had not completed the questionnaire yet and deadlines were extended in the first two rounds. This two measures helped increasing the number of participants during the first round and decreasing the drop-out rate in the second and third rounds, leading to a greater representativeness of all points of view.

The whole process was structured and transparent, which added validity to the results, and the anonymity of participants was assured. The facilitator had a fundamental role in the process, ensuring the participation of panelists and managing the group towards agreement. The use of a web platform increased the efficiency of the process, not only in the delivery of the questionnaire, making it easier for the participants to answer and for the facilitator to overview the process, but also after the rounds were completed, making it easier to analyse the answers.

5.1.4 Test on Preference Dependence

After the Web-based Delphi was complete, with ten evaluation dimensions approved, the fourth phase of the process began.

This phase started with a reflection regarding the possibility of some evaluations dimensions depending on each other.

To facilitate the process, two clusters were created: one regarding evaluation dimensions related to the clinical benefit of the biomarkers (which included Clinical Added Value, Clinical Relevance and Potential Value to Address an Unmet Need in Drug Development), and another related to the the studies and methodologies associated with each biomarker (which included Level of Evidence, Quality of the Study, Test Reliability and Easiness to Measure, Analyse and Interpret).

After pondering on the possibility of eventual dependencies among this evaluation dimensions, the following groups were considered to be possibly preference dependent:

- Clinical Added Value, Clinical Relevance and Potential Value, because these three EDs relate to the clinical benefit of the biomarker. Therefore, preference dependence must be tested for the three possible pairs.
- Quality of the Study and Level of Evidence, because these EDs are both connected with the study.
- Quality of the Study and Test Reliability, because for a study to be good and trustworthy, the tests performed must be reliable.

The Easiness to Measure, Analyze and Interpret was not considered to be possibly preference dependent on another ED, namely the Test Reliability or the Quality of the Study, because it has no direct connection with the quality or reliability of the study, but with the easiness to apply the methodology and analyze the results. The Costs of Development, Ethical Issues and Patient Comfort were not considered to have any connection in terms of preference with the remaining EDs.

After the evaluation dimensions to be tested were defined, two experts in the biomarker field were interviewed via video call, with the purpose of determining if there were preference dependencies among each pair: the possible results could be independent, unilaterally dependent or bilaterally dependent.

Before the interview, both experts were presented with a brief sum of the purpose of the interview, with a simple example to demonstrate what was expected, and with the list of approved evaluation dimensions and respective definitions, so that they would be as informed as possible when the interview began and it was easier for them to answer the questions asked. The results of the test can be observed in Table 5.7.

When comparing the EDs Clinical Relevance and Clinical Added Value, both experts stated clearly that there was some connection between the two, because it is not likely to exist a biomarker that has high relevance but no added value, or a biomarker that has no relevance but a high added value. As a consequence, when asked to give a judgement regarding the difference of attractiveness between two options (one question for each swing), both experts agreed that it would make a huge difference to go from a clinically irrelevant biomarker to a very relevant one, regardless of the added value, leading to a strong/very strong difference of attractiveness in swings one and two, and that it would make little difference to go from 0% added value to 10%, if the biomarker was clinically irrelevant, leading to a weak difference of attractiveness in swing three, while going from 0% added value to 10%, if the biomarker was very clinically relevant, made some difference, leading to a moderate difference of attractiveness in swing four. As a result, we can conclude that Clinical Relevance and Clinical Added Value are unilaterally preference dependent, because the judgements in swings one and two were the same, while in swings three and four they were different, meaning that the Clinical Added Value depends on the Clinical Relevance of a biomarker, but not the other way around.

When comparing the EDs Clinical Relevance and Potential Value to Address an Unmet Need in Drug Development, one expert stated that these two EDs would likely be independent, because it is possible to have a biomarker with high clinical relevance and no potential value to address an unmet

Table 5.7: Results of the test on preference dependence (judgements based on the MACBETH semantic scale).

Evaluation Dimensions	Swing 1	Swing 2	Swing 3	Swing 4	Result
Clinical Relevance ⇳ Clinical Added Value	Strong/ Very Strong	Strong/ Very Strong	Weak	Moderate	Unilateral Preference Dependence
Clinical Relevance ⇳ Potential Value to Address an Unmet Need in Drug Development	Weak	Weak	Very Strong	Very Strong	Preference Independence
Clinical Added Value ⇳ Potential Value to Address an Unmet Need in Drug Development	Weak	Weak	Strong	Very Strong	Preference Independence
Level of Evidence ⇳ Quality of the Study	Moderate/ Strong	Weak	Moderate	Moderate	Unilateral Preference Dependence
Quality of the Study ⇳ Test Reliability	Moderate	Weak	Moderate/ Strong	Moderate	Bilateral Preference Dependence

need in drug development, or to have a biomarker with no clinical relevance but still high potential value, although it is more likely for a biomarker to have high potential if it is clinically relevant. Furthermore, it was mentioned that if a biomarker has no potential value in drug development, because it does not meet its context of use, then it is no use and it would be a waste of time to further investigate it, while if it has high potential, its relevance is not very important. As a consequence, when asked to give their judgements regarding the four swings, both experts agreed that it would not make much difference to go from a clinically irrelevant biomarker to an very relevant one, regardless of its potential value in drug development, leading to a weak difference of attractiveness in swings one and two, and that it would make a very significant difference to go from a potential low value in drug development to a potential high/moderate value, regardless of the clinical relevance, leading to a very strong difference of attractiveness in swings three and four. As a result, we can conclude that the Clinical Relevance and Potential Value to Address an Unmet Need in Drug Development are preference independent, because the judgements in swings one and two, and then in swings three and four were the same.

Like in the previous case, when comparing the EDs Clinical Added Value and Potential Value to Address an Unmet Need in Drug Development, one expert stated that these two EDs would likely be independent, because it is possible to have a biomarker with high added value and no potential in drug development, or the opposite. It was also mentioned, once again, that if a biomarker has no potential value in drug development, then it is no use and it would be a waste of time to further investigate it, while if it has high potential, its added value is not the most important factor. As a consequence, when asked to give their judgements regarding the four swings, both experts agreed that it would not make much difference to go from a biomarker with 0% added value to one with 10% added value, regardless of its potential value in drug development, leading to a weak difference of attractiveness in swings one and two, and that it would make a very significant difference to go from a potential low value in drug development to a potential high/moderate value, regardless of the clinical added value, leading to a very strong difference of attractiveness in swings three and four. As a result, we can conclude that the Clinical Added Value and Potential Value to Address an Unmet Need in Drug Development are preference independent, because the judgements in swings one and two, and then in swings three and four were the same.

When comparing the EDs Level of Evidence and Quality of the Study, both experts stated that there was clearly some connection between the two, because if a study has high quality it can have moderate evidence and still be a very good study, however, if it has high evidence but only moderate or low quality one cannot trust the evidence. As a consequence, when asked to give their judgements regarding the four swings, both experts agreed that it would make a very significant difference to go from a biomarker associated with a study with moderate evidence to one with high evidence, if the study had moderate quality, leading to a moderate/strong difference of attractiveness in swing one, but that it would not make much difference if the study had high quality, leading to a weak difference of attractiveness in swing two. The experts also agreed that a change from moderate quality study to high quality study would make a significant difference regardless of the level of evidence, leading to a moderate difference of attractiveness in swings three and four. As a result, we can conclude that Level of Evidence and Quality of the Study are unilaterally preference dependent, because the judgements in swings one and two were the different, while in swings three and four they were the same, meaning that the the Level of Evidence depends on the Quality of the Study, but not the other way around.

Finally, when comparing the EDs Quality of the Study and Test Reliability, one expert explained that, although it is important to have a high quality study, it does not matter how good a study is if test is not reliable enough, because a medium test reliability is going to result in too many false positives or too many false negatives, leading to a lower accuracy, which cannot happen. As a consequence, when asked to give their judgements regarding the four swings, both experts agreed that it would make a significant difference to go from a biomarker associated with a moderate quality study to one with high quality, if the associated test had medium reliability, leading to a moderate difference of attractiveness in swing one, but that it would not make much difference if the test had high reliability, leading to a weak difference of attractiveness in swing two. The experts also agreed that a change from medium reliability test to high reliability test would make a significant to strong difference if the the quality of the study was moderate, leading to a moderate/strong difference of attractiveness in swing three, but that it would only make a significant difference if the quality of the study was high, leading to a moderate difference of attractiveness in swing four. As a result, we can conclude that Quality of the Study and Test Reliability are bilaterally preference dependent, because the judgements were different in swings one and two, and then in swings three and four, meaning that the the Quality of the Study depends on the Test Reliability and vice-versa.

Considering the results obtained, where three pairs of evaluation dimensions were found to be preference dependent, it was necessary to redefine them by combining them into single independent evaluation dimensions. Clinical Added Value and Clinical Relevance were combined in a single ED, named Clinical Benefit and Value, while Level of Evidence, Quality of the Study and Test Reliability were combined in a single ED, named Quality and Reliability of the Study. Regarding the new descriptors of performance, we opted to create constructed descriptors, in order to merge the descriptors of the previously dependent evaluation dimensions into a single descriptor. As a result, the five not preference dependent evaluation dimensions and the two newly created evaluation dimensions, may now be called evaluation criteria, as they are independent. The two new criteria resultant from the five preference dependent ones, including definition, descriptors of performance and reference levels, were approved by the two experts who participated in the interview to assess preference dependence. The definition of the seven resultant evaluation criteria can be observed in Table 5.8, and their respective descriptors of performance and reference levels in Table 5.9.

Table 5.8: Final list of evaluation criteria for biomarker prioritization and respective descriptions.

Evaluation Dimensions	Description
Clinical Benefit and Value	Represents how relevant the biomarker is in clinical practice, in terms of its ability to improve patient well-being and outcome and how much value it adds (if it brings something new/different) in relation to the already qualified biomarkers, based on the AUC between the new biomarker and the already qualified ones.
Patient Comfort	Comfort of the patient, considering how invasive the procedure used to access the biomarker is (if it is accessible in the peripheral tissue, namely in blood or urine, the discomfort will be null/minimal; if not the discomfort may be greater.
Quality and Reliability of the Study	A high quality and reliability of a study is based on: i) high methodological quality, validity and applicability to patient care; ii) well-designed cohort, including adequate biospecimen repositories and sample size; iii) proper analytical and statistical methods; iv) consistent and reproducible results across time (test-retest reliability), across items (internal consistency) and across researchers (inter-rate reliability); v) accurate data analysis and interpretation to enable biomarker identification/validation meeting pre-specified performance criteria for a given clinical application context
Easiness to Measure, Analyse and Interpret	Easiness to measure and analyze the biomarker, considering the simplicity of the method used, the time needed for this purpose and the need for high-throughput, as well as the easiness to interpret the results.
Potential Value to Address an Unmet Need in Drug Development	Potential value of a biomarker for a particular context of use (COU) in drug development, likely to be recognized by the FDA as qualified biomarker in the near future. COU in drug development includes patient stratification, selection, trial enrichment, dose selection, response/efficacy assessments, safety assessments, among others.
Costs of Validation and Utility	Amount of money necessary for the validation of a biomarker (where the generalizability across different samples and the reproducibility and standardization of the assay are determined) and utility studies (where the results must show performance characteristics, well-designed experiments and the added value in research models and/or patients).
Ethical Issues	Ethical issues associated with the application of the biomarker in clinical practice, including psychological reactions (associated with factors such as risk of catastrophic reaction, no proven long-term treatments and risk of false positive) and social stigma.

5.1.5 Value Tree

Considering the final list of evaluation criteria presented in Table 5.8, an initial division was made according to areas of concern: the Cost of Development was considered to be a cost, the Ethical Issues were considered to be a risk and the remaining criteria were considered to be benefits. Since there were five benefits, three clusters were created, after discussion with Doctor Deborah Penque:

- **Clinical Benefit**, which includes all criteria that present a clinical or pre-clinical (during drug development) benefit for the patient: Clinical Benefit and Value and Potential Value to Address an Unmet Need in Drug Development.

- **Quality of the Study, Analytical Method and Data**, which includes all study and methodology related criteria: Easiness to Measure, Analyse and Interpret and Quality and Reliability of the Study.

- **Patient care**, which includes the criteria directly related with the patient: Patient Comfort.

As a result, a value tree with the selected criteria divided by areas of concern and clusters was created, and is presented in Figure 5.4.

Table 5.9: Final list of evaluation criteria and respective descriptors of performance and reference levels *neutral* and *good*.

Evaluation Dimensions	Descriptors of Performance
Clinical Benefit and Value	Qualitative, discrete and constructed: • Very relevant, showing ability to improve well-being and outcome to the generality of patients and adding over 10% value to the already existing biomarkers. **[GOOD]** • Very relevant, showing ability to improve well-being and outcome to the generality of patients and adding 10% or less value to the already existing biomarkers. • Relevant, showing ability to improve well-being and outcome to patients with specific characteristics and adding 10% or less value to the already existing biomarkers. • Relevant, showing ability to improve well-being and outcome to patients with specific characteristics but adding no value to the already existing biomarkers. **[NEUTRAL]** • Irrelevant, showing poor or no ability to improve patient's well-being and outcome and adding no value to the already existing biomarkers.
Patient Comfort	Qualitative, discrete and constructed: • Non-invasive procedure, with no discomfort to the patient. **[GOOD]** • Semi-invasive procedure, with mild discomfort to the patient. **[NEUTRAL]** • Invasive procedure, with mild/moderate discomfort to the patient. • Invasive procedure, with severe discomfort to the patient.
Quality and Reliability of the Study	Qualitative, discrete and constructed: • High Quality: all recommendations are fully accomplished. **[GOOD]** • Medium Quality: three/four recommendations are fully accomplished. • Low Quality: one/two recommendations are fully accomplished. **[NEUTRAL]** • Very low quality: no recommendation is accomplished.
Easiness to Measure, Analyse and Interpret	Qualitative, discrete and constructed: • Easy to measure and analyze the biomarker and easy to interpret the results. **[GOOD]** • Hard to measure and analyze the biomarker and easy to interpret the results. • Easy to measure and analyze the biomarker and hard to interpret the results. **[NEUTRAL]** • Hard to measure and analyze the biomarker and hard to interpret the results.
Potential Value to Address an Unmet Need in Drug Development	Qualitative, discrete and constructed: • Potential high/moderate-value for a particular COU in drug development with high unmet need. **[GOOD]** • Potential high/moderate-value for a particular COU in drug development. **[NEUTRAL]** • Potential low-value for a particular COU in drug development.
Costs of Development	Quantitative, continuous and direct: • 5M$ **[GOOD]** • 10M$ • 15M$ • 20M$ **[NEUTRAL]** • 25M$ • 30M$
Ethical Issues	Qualitative, discrete and constructed: • Presents no risk of psychological reactions and/or social stigma to the patient. **[GOOD]** • Presents low risk of psychological reactions and/or social stigma to the patient. • Presents moderate risk of psychological reactions and/or social stigma to the patient. **[NEUTRAL]** • Presents high risk of psychological reactions and/or social stigma to the patient.

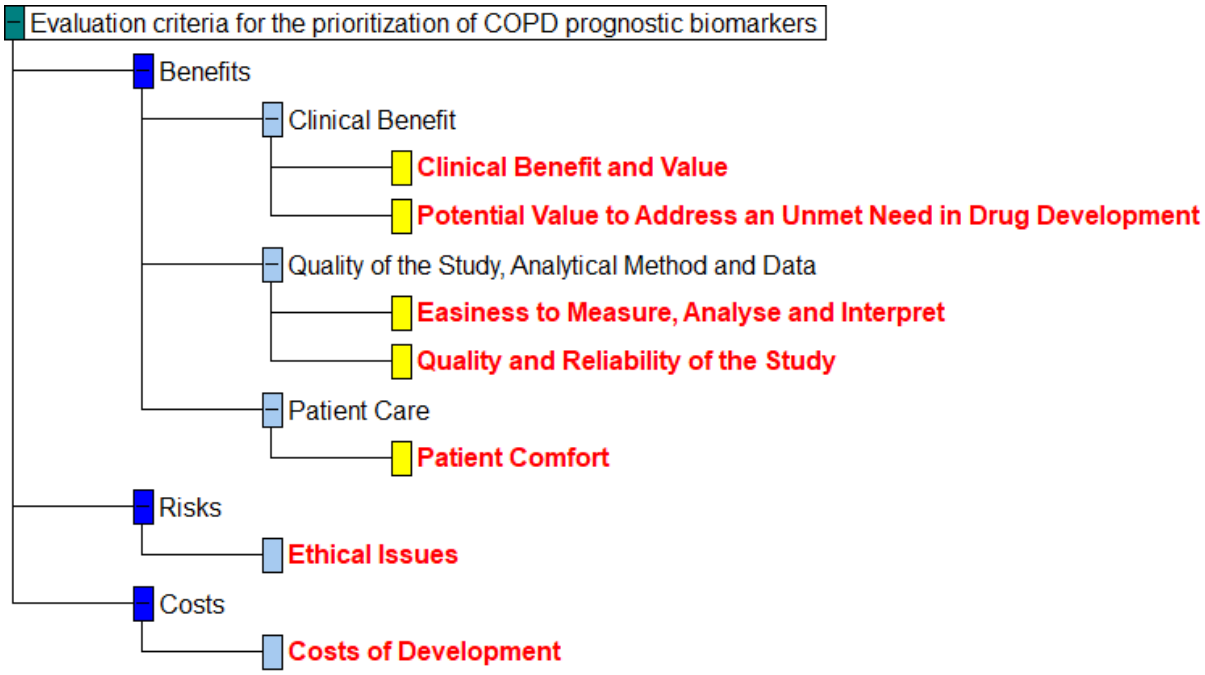

Figure 5.4: Organization of the criteria in a value tree using the software M-MACBETH. The dark blue nodes are the areas of concern, the light blue nodes with the terms not highlighted in red are the clusters and the evaluation criteria are the terms highlighted in red.

5.2 Discussion of Methodology and Results

The socio-technical approach used in this thesis for structuring the model, which included the technical elements of the MCDA model MACBETH, but also the social elements of participatory methods, was used for the purpose of combining scientific evidence with the opinion of different experts, and, consequently, reach greater validity and credibility of the model. It involved several steps: the first three steps, which did not involve a social part, were the assessment of areas of concern in the biomarker field, the definition of exclusion criteria and the creation of a list with all relevant COPD prognostic biomarkers found in literature; in the other hand, the fourth step included both a social and technical part, and had the purpose of determining the relevant evaluation criteria for the prioritization of COPD prognostic biomarkers, being divided in five phases (evidence analysis, interviews, web-based Delphi, test on preference dependence/restructuring of criteria and building of the value tree).

The use of said socio-technical methodology led to a greater acceptance, a more inclusive and representative model and, consequently, a more valid and credible model. Its greatest downside is the fact that it becomes substantially more complex when more than one biomarker category is included, which is the reason why we only focused on prognostic biomarkers in this thesis.

Contrarily to the solely technical phases, in the three phases that involved experts opinions, there were some difficulties noted. During the interviews, it was particularly difficult to explain to experts what criteria were and what they were used for, as all experts tended to be focused primarily on the benefit a biomarker could present, stating repeatedly that the benefit was the most important factor when selecting a biomarker. However, since the interviews were conducted through video-call, everything was explained in detail, including some examples, which helped the experts to understand the purpose of more than one criterion in a prioritization process. All experts found it substantially easier to understand what a descriptor of performance and reference level were, as well as to judge them. Although only four experts were interviewed, these interviews allowed for a thorough analysis of each of the proposed

criteria and respective descriptors of performance and reference levels, as experts suggested changes and mentioned relevant clinical or biological information that significantly helped to define the criteria with clarity and scientific rigor.

The web-based Delphi had several advantages, namely it: allowed for a far greater number of participants than a face-to-face decision conference would allow; reunited the opinions of individuals with different backgrounds and from different geographical locations, resulting in a very heterogeneous sample and diverse opinions, which adds value and credibility to the model; helped avoiding the opinion of domineering individuals from being the only heard; avoided group pressure (either due to social norms, customs, culture or standing within profession) for conformity; allowed for individuals to have access to all information before taking a position; and, by being anonymous, allowed people to freely give their opinion without fear of rejection. However, this method presented some disadvantages as well. The four that clearly stood out during the development of this thesis were: (i) it took a very long time to be completed; (ii) the e-mails sent to participants were perceived as spam by many experts (specially by those who were not familiar with the platform) and sometimes were directly sent to the spam instead of the participants' inbox, leading to less participation; (iii) without an incentive (monetary, for example) it was substantially harder to make individuals interested enough to participate. As a result, the percentage of experts that accepted to participate in the Web-Delphi conducted during this thesis was very low (only 20%), although this number also reflected the month the web-Delphi was developed in (August), in which a great percentage of the individuals were probably on vacation; (iv) the platform used did not allow for an analysis of the evolution of each participants' opinion throughout the rounds, nor for an evaluation of the the answers according to field of expertise, due to the fact that there was only access to a table with the percentage of answers in each scale level, and not individual answers (even coded, due to anonymity), which led to a poorer analysis of results.

During the test on preference dependence interviews, which were essential to guarantee the independence among criteria, it was particularly difficult for the two experts involved to understand the purpose of the test on preference dependence, the questions asked and what they should answer, even with the questions presented in a graphical way. However, since the interviews were once again conducted through a video-call, there was a chance to clarify all the doubts in real time, which made it easier for experts to understand the whole process. Initially, they also had some difficulty using the MACBETH semantic scale, as they were not used to give the type of judgments asked of them. However, the possibility of choosing two categories instead of one when evaluating the differences of attractiveness helped, as sometimes the experts were hesitant regarding which option to choose.

Despite the difficulties presented by the experts, who had never been subjected to such a way of thinking, after some explanation of the objective of the thesis and of each phase, and what was required of them, they were able to look at the problem in a new way and reacted satisfactorily to the new method, mentioning that this new model could be very helpful in the future.

In the end, seven evaluation criteria were considered relevant by the experts who participated in the several phases of this socio-technical process. However, it is likely that results would have been different with a different panel of experts and considering different contexts, with certain evaluation dimensions being considered more or less relevant, and some not being approved. For example, in the web-based Delphi, if there had been an even number of researchers and clinicians, or even a higher number of clinicians, the results could have been different. In fact, one of this web-Delphi's potential limitations is researcher bias, as the panel of experts that participated was composed by 76% researchers and only 24% clinicians.

Overall, the socio-technical methodology used in this thesis worked well, despite the few difficulties met throughout the process, resulting in a good structure for the MCDA model, with seven well defined and relevant evaluation criteria to be used for the prioritization of COPD prognostic biomarkers.

Chapter 6

Discussion

Although MCDA, and most specifically the MACBETH approach, has been applied in several healthcare related studies, there are very few studies regarding its use in biomarker prioritization, and almost no information regarding biomarker prioritization in general in COPD. Therefore, by applying the MACBETH socio-technical approach to the case of COPD, a completely new and original model has been developed, although there is still a long way to go until its completion.

By going through three different phases to define evaluation criteria and respective descriptors of performance and reference levels, based on three different methods, evidence analysis, interviews to four experts and web-based Delphi to twenty-two experts, instead of only one phase, we were able to guarantee that, according to the clinicians and researchers that participated in the study, the final list of evaluation criteria contained only relevant elements, adding validity and credibility to the model.

Furthermore, by testing for inter-dependencies in this thesis, we were able to guarantee that the final list of relevant evaluation criteria were independent of each other, which is an essential characteristic for evaluation criteria in a MCDA model. The application of this method in this thesis is particularly noteworthy due to its originality, since most literature on MCDA in HTA does not test or treat evaluation dimensions for inter-dependencies.

By only further analysing and testing the most promising biomarkers, which will likely have the best outcomes, the number of clinically validated and qualified biomarkers will tend to increase significantly. As a consequence, personalized medicine can be improved, by improving disease specific drugs using the qualified disease specific biomarkers, possibly leading to better disease outcomes, such as higher patient survival and lower health costs, among others. Therefore, the model designed and developed in this thesis comes as a great contribute in the drug development and clinical fields, as it is innovative and allows for the determination of the most promising disease specific biomarkers, when there are hundreds available, leading to the development of more efficient drugs and, consequently, better patient care.

Despite some difficulties met throughout the process, the socio-technical methodology used in this thesis worked well, resulting in a good structure for the MCDA model, with seven well defined and relevant evaluation criteria: clinical benefit and value, potential value to address an unmet need in drug development, easiness to measure, analyse and interpret, quality and reliability of the study, patient comfort, ethical issues and costs of development. The criteria defined will be very useful for CliniMark to prioritize COPD prognostic biomarkers, by facilitating the comparison of the different options and by reducing time and monetary resources.

Chapter 7

Conclusion and Future Work

7.1 Conclusion

In this master thesis, the structuring part of a MCDA model, more specifically a socio-technical MACBETH approach, was developed, for the purpose of evaluating and selecting the most promising COPD prognostic biomarkers among the hundreds found in literature.

Several steps were taken to reach a final list of valid and credible criteria, including evidence analysis, interviews to four experts, a web-based Delphi to several experts and, with the purpose of confirming the dependencies among them, a test on preference dependence. The evidence analysis allowed for an initial and very basic list of nine potentially relevant evaluation dimensions. The interviews were essential to reach a group of ten well-defined, reformulated and potentially relevant biomarkers and respective descriptors of performance and reference levels. The twenty two expert opinions gathered during the web-based Delphi were essential to confirm the relevance of the proposed evaluation dimensions. The opinions clearly converged throughout the three rounds, meaning the Delphi process was successful in promoting agreement. Finally, some dependencies were found during the test on preference dependence, which led to the grouping of some evaluation dimensions, resulting in a total of seven preference independent evaluation dimensions (when independent, called evaluation criteria). By including experts from different groups of expertise and geographical locations throughout the process, the model became more inclusive and representative, leading to more validity and credibility.

After the four mentioned phases, we reached a solid structuring of the model, and credible evaluation criteria to prioritize COPD prognostic biomarkers.

The main difficulties throughout the whole process involved selecting the criteria and building clear and non-subjective descriptors of performance, motivating the experts to participate in the Web-based Delphi, explaining to the experts what was being asked and what was the purpose of the test on preference dependence and the fact that a better development of methods would require more resources than the ones that were available.

In the future, when the model is completed, it will be possible to select the most promising COPD biomarkers, which will increase significantly the probability of the biomarkers' qualification success and reduce both the monetary and time resources necessary to increase the number of qualified biomarkers, which is currently too high. Furthermore, and more importantly, the number of qualified disease specific biomarkers will tend to increase, increasing, therefore, the quality and efficiency of patient care.

7.2 Future Work

In the future, it would be interesting to complete the model for COPD prognostic biomarkers, including the building, testing and validation of the model.

It would also be interesting to apply the developed MCDA model to a COU different from the one used in this thesis. It would be ideal to develop a new MCDA model that could compare and prioritize biomarkers with several different COUs, instead of the the single one considered during this thesis, for the purpose of simplicity. This would result in a more comprehensive model.

It would also be interesting to improve MCDA tools for stakeholder involvement and evaluation in the context and to develop a model with the participation of different combinations of participants, as well as a greater number of experts, from several fields, so that the model can be more robust, avoiding the variability associated with a group of decision-makers as small as the one used in this thesis.

Finally, the developed model could be applied to other biomarkers and, specially, diseases other than COPD, since the idea of developing such a model was to make it comprehensive, in such a way that it was not limited to the prioritization of biomarkers for only one disease.

Bibliography

[1] "CA16113 - CliniMARK: 'good biomarker practice' to increase the number of clinically validated biomarkers." European Cooperation in Science and Technology. Available at: https://www.cost.eu/actions/CA16113/#tabs—Name:overview (Accessed: 21-03-2019).

[2] S. Ongay, F. Klont, P. Horvatovich, R. Bischoff, and N. H. Ten Hacken, "Prioritization of COPD protein biomarkers, based on a systematic study of the literature," *Advances in Precision Medicine*, vol. 1, no. 1, p. 4, 2016.

[3] R. A. Stockley, D. M. Halpin, B. R. Celli, and D. Singh, "COPD Biomarkers and their Interpretation," *American Journal of Respiratory and Critical Care Medicine*, pp. 1–26, 2018.

[4] A. J. Vargas and C. C. Harris, "Biomarker development in the precision medicine era: Lung cancer as a case study," *Nature Reviews Cancer*, vol. 16, no. 8, pp. 525–537, 2016.

[5] E. F. Wouters, B. B. Wouters, I. M. Augustin, and F. M. Franssen, "Personalized medicine and chronic obstructive pulmonary disease," *Current Opinion in Pulmonary Medicine*, vol. 23, no. 3, pp. 241–246, 2017.

[6] F. R. Vogenberg, C. Isaacson Barash, and M. Pursel, "Personalized medicine: part 1: evolution and development into theranostics.," *P & T : a peer-reviewed journal for formulary management*, vol. 35, no. 10, pp. 560–576, 2010.

[7] "Value of Personalized Medicine - A New Treatment Paradigm." Pharmaceutical Research and Manufacturers of America. Available at: https://chartpack.phrma.org/personal-medicines-in-development-chartpack/a-new-treatment-paradigm/a-new-treatment-paradigm (Accessed: 02-04-2019).

[8] J. R. Hurst, "Precision Medicine in Chronic Obstructive Pulmonary Disease," *American Journal of Respiratory and Critical Care Medicine*, vol. 193, no. 6, pp. 594–596, 2016.

[9] K. Strimbu and J. A. Tavel, "Biomarkers In Risk Assessment: Validity And Validation," *Environmental Health*, vol. 5, no. 6, p. 144, 2001.

[10] Z. Hollander, M. L. DeMarco, D. D. Sin, M. Sadatsafavi, R. T. Ng, and B. M. McManus, "Biomarker Development in COPD," *Chest*, vol. 151, no. 2, pp. 455–467, 2016.

[11] M. R. Bleavins, C. Carini, M. J. Romet, and R. Rahbari, "Biomarkers," *Pharmaceutical Sciences Encyclopedia*, 2010.

[12] C. V. Gerlach, M. Derzi, S. K. Ramaiah, and V. S. Vaidya, "Industry Perspective on Biomarker Development and Qualification," *Clinical Pharmacology and Therapeutics*, vol. 103, no. 1, pp. 27–31, 2018.

[13] G. Graaf, D. Postmus, J. Westerink, and E. Buskens, "The early economic evaluation of novel biomarkers to accelerate their translation into clinical applications," *Cost Effectiveness and Resource Allocation*, vol. 16, no. 1, pp. 1–8, 2018.

[14] M. G. Scott, "When do new biomarkers make economic sense?," *Scandinavian Journal of Clinical and Laboratory Investigation*, vol. 70, no. SUPPL. 242, pp. 90–95, 2010.

[15] M. Oosterhoff, M. E. van der Maas, and L. M. Steuten, "A Systematic Review of Health Economic Evaluations of Diagnostic Biomarkers," *Applied Health Economics and Health Policy*, vol. 14, no. 1, pp. 51–65, 2016.

[16] W. Chen, F. W. Samuelson, B. D. Gallas, L. Kang, B. Sahiner, and N. Petrick, "On the assessment of the added value of new predictive biomarkers," *BMC Medical Research Methodology*, vol. 13, no. 1, 2013.

[17] N. R. Cook, "Quantifying the added value of new biomarkers: how and how not," *Diagnostic and Prognostic Research*, vol. 2, no. 1, pp. 1–7, 2018.

[18] R. Mayeux, "Biomarkers: Potential Uses and Limitations," *The Journal of the American Society for Experimental NeuroTherapeutics*, vol. 1, no. April, pp. 182–188, 2004.

[19] S. Robinson, R. Pool, and R. Giffin, *Emerging Safety Science: Workshop Summary*. The National Academies Press, 2008.

[20] T. Aggarwal, R. Wadhwa, N. Thapliyal, K. Sharma, V. Rani, and P. K. Maurya, "Oxidative, inflammatory, genetic, and epigenetic biomarkers associated with chronic obstructive pulmonary disorder," *Journal of Cellular Physiology*, vol. 234, no. 3, pp. 2067–2082, 2019.

[21] J. Gea, S. Pascual, A. Castro-Acosta, C. Hernández-Carcereny, R. Castelo, E. Márquez-Martín, C. Montón, A. Palou, R. Faner, L. I. Furlong, L. Seijo, F. Sanz, M. Torà, C. Vilaplana, C. Casadevall, J. L. López-Campos, E. Monsó, G. Peces-Barba, B. G. Cosío, A. Agustí, M. Admetlló, A. Agustí, C. Alvarez-Martínez, E. Barreiro, C. Casadevall, F. Casals, R. Castelo, A. Castro-Acosta, R. Córdova, B. G. Cosío, R. Faner, L. I. Furlong, M. García, J. Gea, J. G. González-García, C. Hernández-Carcereny, J. L. López-Campos, E. Márquez, E. Monsó, C. Montón, M. J. Ormaza, A. Palou, S. Pascual, G. Peces-Barba, P. Puigdevall, F. Sanz, L. Seijó, M. Torà, Y. Torralba, and C. Vilaplana, "The BIOMEPOC Project: Personalized Biomarkers and Clinical Profiles in Chronic Obstructive Pulmonary Disease," *Archivos de Bronconeumologia*, no. xx, 2018.

[22] "What is COPD?." British Lung Foundation. Available at: https://www.blf.org.uk/support-for-you/copd/what-is-copd (Accessed: 02-04-2019).

[23] D. P. Johns, J. A. E. Walters, and E. Haydn Walters, "Diagnosis and early detection of COPD using spirometry," *Journal of Thoracic Disease*, vol. 6, no. 11, pp. 1557–1569, 2014.

[24] E. Aydindogan, D. Penque, and J. Zoidakis, "Systematic review on recent potential biomarkers of chronic obstructive pulmonary disease," *Taylor & Francis Online*, 2018.

[25] R. J. McKenna, *Approaches to Decision Making*. 1996.

[26] P. K. Narayan and S. Narayan, "Using economic evidence to set healthcare priorities in low-income and lower-middle-income countries: a systematic review of methodological frameworks," vol. 1186, no. December 2007, pp. 1171–1186, 2008.

[27] M. D. Oliveira, I. Mataloto, and P. Kanavos, “Multi-criteria decision analysis for health technology assessment: addressing methodological challenges to improve the state of the art,” *The European Journal of Health Economics*, no. 0123456789, 2019.

[28] A. D. Montis and P. D. Toro, “Assessing the quality of different MCDA methods,” *Alternatives for environmental valuation*, no. June 2016, pp. 99–184, 2000.

[29] R. Baltussen, M. Paul Maria Jansen, L. Bijlmakers, J. Grutters, A. Kluytmans, R. P. Reuzel, M. Tummers, and G. J. v. der Wilt, “Value Assessment Frameworks for HTA Agencies: The Organization of Evidence-Informed Deliberative Processes,” *Value in Health*, vol. 20, no. 2, pp. 256–260, 2017.

[30] “Decision Making Techniques.” Decision Innovation. Available at: https://www.decision-making-solutions.com/decision_making_techniques.html (Accessed: 17-04-2019).

[31] A. Baran-Kooiker, M. Czech, and C. Kooiker, “Multi-Criteria Decision Analysis (MCDA) Models in Health Technology Assessment of Orphan Drugs—a Systematic Literature Review. Next Steps in Methodology Development?,” *Frontiers in Public Health*, vol. 6, no. October, 2018.

[32] G. Adunlin, V. Diaby, and H. Xiao, “Application of multicriteria decision analysis in health care: A systematic review and bibliometric analysis,” *Health Expectations*, vol. 18, no. 6, pp. 1894–1905, 2015.

[33] P. Thokala, N. Devlin, K. Marsh, R. Baltussen, M. Boysen, Z. Kalo, T. Longrenn, F. Mussen, S. Peacock, J. Watkins, and M. Ijzerman, “Multiple criteria decision analysis for health care decision making - An introduction: Report 1 of the ISPOR MCDA Emerging Good Practices Task Force,” *Value in Health*, vol. 19, no. 1, pp. 1–13, 2016.

[34] L. Rietkötter, “Ending the war in multi-criteria decision analysis: Taking the best from two worlds - The development and evaluation of guidelines for the use of MACBETH in multi-criteria group decision making for the assessment of new medical products,” 2014.

[35] H. A. Donegan, F. J. Dodd, and T. B. M. McMaster, “A New Approach to AHP Decision-Making,” vol. 2, no. May 2016, pp. 129–136, 2016.

[36] S. Karapetrovic and E. S. Rosenbloom, “Quality control approach to consistency paradoxes in AHP,” *European Journal of Operational Research*, vol. 119, no. 3, pp. 704–718, 1999.

[37] C. A. Bana e Costa and J. C. Vansnick, “A critical analysis of the eigenvalue method used to derive priorities in AHP,” *European Journal of Operational Research*, vol. 187, no. 3, pp. 1422–1428, 2008.

[38] A. Angelis and P. Kanavos, “Multiple Criteria Decision Analysis (MCDA) for evaluating new medicines in Health Technology Assessment and beyond: The Advance Value Framework,” *Social Science and Medicine*, vol. 188, pp. 137–156, 2017.

[39] M. R. Guarini, F. Battisti, and A. Chiovitti, “A methodology for the selection of multi-criteria decision analysis methods in real estate and land management processes,” *Sustainability (Switzerland)*, vol. 10, no. 2, 2018.

[40] C. Bana e Costa (2018). Multi-criteria Value Measurement, lecture notes. Retrieved from: https://fenix.tecnico.ulisboa.pt/downloadFile/1689468335616940/MAD%202018_2019%20Larry.pdf (Accessed: 06-05-2019).

[41] A. Zawodnik and M. Niewada, “Multiple Criteria Decision Analysis (MCDA) for Health Care Decision Making – overview of guidelines .,” pp. 1–12, 2018.

[42] M. Dabrowski, “The Simple Multi Attribute Rating Technique (SMART),” tech. rep., 2014.

[43] P. D. Site and F. Filippi, “Weighting methods in multi-attribute assessment of transport projects,” *European Transport Research Review*, vol. 1, no. 4, pp. 199–206, 2009.

[44] N. Hassan, Z. Kamal, A. S. Moniruzzaman, S. Zulkifli, and B. Yusop, *Weighting Methods and their Effects on Multi-Criteria Decision Making Model Outcomes in Water Resources Management*. Springer, 2015.

[45] C. Bana e Costa, J.-M. De Corte, and J.-C. Vansnick, “MACBETH. (Overview of MACBETH multicriteria decision analysis approach),” *International Journal of Information Technology and Decision Making*, vol. 11, no. January, pp. 359–387, 2003.

[46] C. A. Bana e Costa, J. A. Ferreira, and E. C. Correa, “Metodologia multicritério de apoio à avaliação de propostas em concursos públicos,” 2000.

[47] A. C. Vieira, M. D. Oliveira, and C. A. Bana e Costa, “Enhancing knowledge construction processes within multicriteria decision analysis: The Collaborative Value Modelling framework,” *Omega (United Kingdom)*, pp. 1–15, 2019.

[48] K. Marsh, M. Ijzerman, P. Thokala, R. Baltussen, M. Boysen, Z. Kaló, T. Lönngren, F. Mussen, S. Peacock, J. Watkins, and N. Devlin, “Multiple Criteria Decision Analysis for Health Care Decision Making - Emerging Good Practices: Report 2 of the ISPOR MCDA Emerging Good Practices Task Force,” *Value in Health*, vol. 19, no. 2, pp. 125–137, 2016.

[49] M. D. Oliveira, T. C. Rodrigues, C. A. Bana e Costa, and A. Brito de Sá, “Prioritizing health care interventions: A multicriteria resource allocation model to inform the choice of community care programmes,” in *Advanced Decision Making Methods Applied to Health Care*, vol. 173, pp. 141–154, 2012.

[50] T. Cardoso, M. D. Oliveira, A. Barbosa-Póvoa, and S. Nickel, “Moving towards an equitable long-term care network: A multi-objective and multi-period planning approach,” *Omega (United Kingdom)*, vol. 58, pp. 69–85, 2016.

[51] C. A. Bana E Costa, M. C. Carnero, and M. D. Oliveira, “A multi-criteria model for auditing a Predictive Maintenance Programme,” *European Journal of Operational Research*, vol. 217, no. 2, pp. 381–393, 2012.

[52] T. C. Rodrigues, “The MACBETH Approach to Health Value Measurement: Building a Population Health Index in Group Processes,” *Procedia Technology*, vol. 16, pp. 1361–1366, 2014.

[53] P. R. Pinheiro, A. K. A. De Castro, and M. C. D. Pinheiro, “A multicriteria model applied in the diagnosis of Alzheimer’s disease: A Bayesian network,” *Proceedings - 2008 IEEE 11th International Conference on Computational Science and Engineering, CSE 2008*, pp. 15–22, 2008.

[54] D. F. Lopes, M. D. Oliveira, and C. A. Bana E Costa, “Occupational health and safety: Designing and building with MACBETH a value risk-matrix for evaluating health and safety risks,” *Journal of Physics: Conference Series*, vol. 616, no. 1, 2015.

[55] FDA/CDER/OSP/OSPA, “Thoughts on Evidentiary Criteria for Biomarker Qualification : A “ Decision Science ” Perspective,” 2018.

[56] A. Miquel-Cases, P. C. Schouten, L. M. Steuten, V. P. Retèl, S. C. Linn, and W. H. van Harten, "(Very) Early technology assessment and translation of predictive biomarkers in breast cancer," *Cancer Treatment Reviews*, vol. 52, pp. 117–127, 2017.

[57] K. Marsh, E. Zaiser, P. Orfanos, S. Salverda, T. Wilcox, S. Sun, and S. Dixit, "Evaluation of COPD Treatments: A Multicriteria Decision Analysis of Aclidinium and Tiotropium in the United States," *Value in Health*, vol. 20, no. 1, pp. 132–140, 2017.

[58] A. Freitas, P. Santana, M. D. Oliveira, R. Almendra, J. C. Bana e Costa, and C. A. Bana e Costa, "Indicators for evaluating European population health: a Delphi selection process," *BMC Public Health*, vol. 18, no. 1, pp. 1–20, 2018.

[59] N. Slocum, *Participatory methods toolkit: A practitioner' s manual*. 2003.

[60] L. D. Phillips and C. A. Bana e Costa, "Transparent prioritisation, budgeting and resource allocation with multi-criteria decision analysis and decision conferencing," *Annals of Operations Research*, vol. 154, no. 1, pp. 51–68, 2007.

[61] C.-c. Hsu and B. A. Sandford, "The Delphi Technique: Making Sense of Consensus," *Practical Assessment, Research and Evaluation*, vol. 12, no. 10, pp. 1531–7714, 2007.

[62] L. E. Miller, "Determining what could/should be: The Delphi technique and its application," 2006.

[63] F. L. Ulschak, *Human resource development: The theory and practice of need assessment*. Reston Publishing Company, Inc., 1983.

[64] P. J. Green, "The content of a college-level outdoor leadership course," tech. rep., 1982.

[65] J. W. Murry and J. O. Hammons, "Delphi: A Versatile Methodology for Conducting Qualitative Research," *The Review of Higher Education*, vol. 18, no. 4, pp. 423–436, 2017.

[66] I. Belton, A. Macdonald, G. Wright, and I. Hamlin, "Improving the practical application of the Delphi method in group-based judgment : A six-step prescription for a well-founded and defensible process," *Technological Forecasting & Social Change*, vol. 147, no. July, pp. 72–82, 2019.

[67] T. C. Rodrigues, G. Montibeller, M. D. Oliveira, and C. A. Bana e Costa, "Modelling multicriteria value interactions with Reasoning Maps," *European Journal of Operational Research*, vol. 258, no. 3, pp. 1054–1071, 2017.

[68] C. Cook and L. Petrucelli, "Oxidative stress," *Parkinson's Disease, Second Edition*, no. October 2014, pp. 559–582, 2012.

[69] "List of Qualified Biomarkers." U.S. Food and Drug Administration. Available at: https://www.fda.gov/Drugs/DevelopmentApprovalProcess/DrugDevelopmentToolsQualificationProgram/BiomarkerQualificationProgram/ucm535383.htm (Accessed: 08-04-2019).

[70] T. C. Rodrigues, M. D. Oliveira, C. A. Bana e Costa, and P. Santana, "The MACBETH approach to health value measurement : a multicriteria model for building a value-based population health index," 2014.

[71] "WP6 Decision Support for Multicriteria Modelling of the Population Health Index and Evaluation, Foresight and Selection of Policies". Euro-Healthy. Available at: http://www.euro-healthy.eu/research/decision-multicriteria-modelling (Accessed: 16-09-2019).

[72] C. A. Bana e Costa and E. Beinat, “Model-structuring in public decision-aiding,” *The London School of Economics and Political Science*, 2005.

[73] C. A. Bana e Costa and E. Beinat, “Estruturação de Modelos de Análise Multicritério de Problemas de Decisão Pública,” *Centre of Management Studies of IST (CEG-IST)*, no. 3, pp. 1–30, 2010.

[74] C. A. Bana e Costa, J. De Corte, J. Vansnick, J. C. Bana e Costa, M. P. Chagas, E. C. Correa, I. M. Joao, D. Lopes, F. M. Lopes, J. C. Lourenco, R. Sanchez-Lopez, R. Sobrinho, R. Lavoie, and T. Rodrigues, “M-MACBETH (beta): User’s Guide,” 2017.

[75] O. G. León, “Value-Focused Thinking versus Alternative-Focused Thinking: Effects on Generation of Objectives,” *Organizational Behavior and Human Decision Processes*, vol. 80, no. 3, pp. 213–227, 1999.

[76] “Context of Use”. U.S. Food and Drug Administration. Available at: https://www.fda.gov/drugs/cder-biomarker-qualification-program/context-use (Accessed: 29-07-2019).

[77] J. M. Leung, V. Chen, Z. Hollander, D. Dai, S. J. Tebbutt, S. D. Aaron, K. L. Vandemheen, S. I. Rennard, J. M. FitzGerald, P. G. Woodruff, S. C. Lazarus, J. E. Connett, H. O. Coxson, B. Miller, C. Borchers, B. M. McManus, R. T. Ng, and D. D. Sin, “COPD exacerbation biomarkers validated using multiple reaction monitoring mass spectrometry,” *PLoS ONE*, vol. 11, no. 8, pp. 1–12, 2016.

[78] G. Hu, Y. Wu, Y. Zhou, Y. Yu, W. Liang, and P. Ran, “Cystatin C as a predictor of in-hospital mortality after exacerbation of COPD,” *Respiratory Care*, vol. 61, no. 7, pp. 950–957, 2016.

[79] D. M. Schumann, D. Leeming, E. Papakonstantinou, F. Blasi, K. Kostikas, W. Boersma, R. Louis, B. Milenkovic, J. Aerts, J. M. Sand, E. F. Wouters, G. Rohde, C. Prat, A. Torres, T. Welte, M. Tamm, M. Karsdal, and D. Stolz, “Collagen Degradation and Formation Are Elevated in Exacerbated COPD Compared With Stable Disease,” *Chest*, vol. 154, no. 4, pp. 798–807, 2018.

[80] J. M. Sand, D. J. Leeming, I. Byrjalsen, A. R. Bihlet, P. Lange, R. Tal-Singer, B. E. Miller, M. A. Karsdal, and J. Vestbo, “High levels of biomarkers of collagen remodeling are associated with increased mortality in COPD - results from the ECLIPSE study,” *Respiratory Research*, vol. 17, no. 1, pp. 1–12, 2016.

[81] D. Stolz, D. J. Leeming, J. H. E. Kristensen, M. A. Karsdal, W. Boersma, R. Louis, B. Milenkovic, K. Kostikas, F. Blasi, J. Aerts, J. M. Sand, E. F. Wouters, G. Rohde, C. Prat, A. Torres, T. Welte, M. Roth, E. Papakonstantinou, and M. Tamm, “Systemic Biomarkers of Collagen and Elastin Turnover Are Associated With Clinically Relevant Outcomes in COPD,” *Chest*, vol. 151, no. 1, pp. 47–59, 2017.

[82] J. A. Hampson, R. A. Stockley, and A. M. Turner, “Free light chains: Potential biomarker and predictor of mortality in alpha-1-antitrypsin deficiency and usual COPD,” *Respiratory Research*, vol. 17, no. 1, pp. 1–9, 2016.

[83] M. Dres, P. Hausfater, F. Foissac, M. Bernard, L. M. Joly, M. Sebbane, A. L. Philippon, C. Gil-Jardiné, J. Schmidt, M. Maignan, J. M. Treluyer, and N. Roche, “Mid-regional pro-adrenomedullin and copeptin to predict short-term prognosis of COPD exacerbations: A multicenter prospective blinded study,” *International Journal of COPD*, vol. 12, pp. 1047–1056, 2017.

[84] J. A. Winther, J. Brynildsen, A. D. Høiseth, H. Strand, I. Følling, G. Christensen, S. Nygård, H. Røsjø, and T. Omland, "Prognostic and diagnostic significance of copeptin in acute exacerbation of chronic obstructive pulmonary disease and acute heart failure: Data from the ACE 2 study," *Respiratory Research*, vol. 18, no. 1, pp. 1–10, 2017.

[85] N. Putcha, G. G. Paul, A. Azar, R. A. Wise, W. K. O'Neal, M. T. Dransfield, P. G. Woodruff, J. L. Curtis, A. P. Comellas, M. B. Drummond, A. A. Lambert, L. M. Paulin, A. Fawzy, R. E. Kanner, R. Paine, M. L. K. Han, F. J. Martinez, R. P. Bowler, R. G. Barr, and N. N. Hansel, "Lower serum IgA is associated with COPD exacerbation risk in SPIROMICS," *PLoS ONE*, vol. 13, no. 4, pp. 1–10, 2018.

[86] R. Linder, E. Rönmark, J. Pourazar, A. F. Behndig, A. Blomberg, and A. Lindberg, "Proteolytic biomarkers are related to prognosis in COPD- report from a population-based cohort," *Respiratory Research*, vol. 19, no. 1, pp. 5–11, 2018.

[87] R. Pavasini, G. Tavazzi, S. Biscaglia, F. Guerra, A. Pecoraro, F. Zaraket, F. Gallo, G. Spitaleri, M. Contoli, R. Ferrari, and G. Campo, "Amino terminal pro brain natriuretic peptide predicts all-cause mortality in patients with chronic obstructive pulmonary disease: Systematic review and meta-analysis," *Chronic Respiratory Disease*, vol. 14, no. 2, pp. 117–126, 2017.

[88] M. Adrish, V. B. Nannaka, E. J. Cano, B. Bajantri, and G. Diaz-Fuentes, "Significance of NT-pro-BNP in acute exacerbation of COPD patients without underlying left ventricular dysfunction," *International Journal of COPD*, vol. 12, pp. 1183–1189, 2017.

[89] R. A. Rabinovich, B. E. Miller, K. Wrobel, K. Ranjit, M. C. Williams, E. Drost, L. D. Edwards, D. A. Lomas, S. I. Rennard, A. Agustí, R. Tal-Singer, J. Vestbo, E. F. Wouters, M. John, E. J. Van Beek, J. T. Murchison, C. E. Bolton, W. Macnee, and J. T. Huang, "Circulating desmosine levels do not predict emphysema progression but are associated with cardiovascular risk and mortality in COPD," *European Respiratory Journal*, vol. 47, no. 5, pp. 1365–1373, 2016.

[90] S. J. Thulborn, M. Dilpazir, K. Haldar, V. Mistry, C. E. Brightling, M. R. Barer, and M. Bafadhel, "Investigating the role of pentraxin 3 as a biomarker for bacterial infection in subjects with COPD," *International Journal of COPD*, vol. 12, pp. 1199–1205, 2017.

[91] X. Tong, D. Wang, S. Liu, Y. Ma, Z. Li, P. Tian, and H. Fan, "The YKL-40 protein is a potential biomarker for COPD: A meta-analysis and systematic review," *International Journal of COPD*, vol. 13, pp. 409–418, 2018.

[92] J. Wang, H. Lv, Z. Luo, S. Mou, J. Liu, C. Liu, S. Deng, Y. Jiang, J. Lin, C. Wu, X. Liu, J. He, and D. Jiang, "Plasma YKL-40 and NGAL are useful in distinguishing ACO from asthma and COPD," *Respiratory Research*, vol. 19, no. 1, pp. 1–10, 2018.

[93] D. D. Sin and S. F. Man, "Biomarkers in COPD: Are we there yet?," *Chest*, vol. 133, no. 6, pp. 1296–1298, 2008.

[94] R. A. Stockley, "Biomarkers in COPD: time for a deep breath," *Thorax*, vol. 62, no. 8, pp. 656–657, 2007.

[95] "About Biomarker and Qualification". U.S. Food and Drug Administration. Available at: https://www.fda.gov/drugs/cder-biomarker-qualification-program/about-biomarkers-and-qualification#How_can_qualified_biomarkers_improve_the_drug_development_process (Accessed: 08-07-2019).

[96] S. Gauthier, “Ethical issues in the use of new diagnostic biomarkers and future combination therapies for AD.”

[97] M. Jurjako, L. Malatesti, and I. A. Brazil, “Some Ethical Considerations About the Use of Biomarkers for the Classification of Adult Antisocial Individuals,” *International Journal of Forensic Mental Health*, pp. 1–15, 2018.

[98] “Classification: ROC Curve and AUC”. Machine Learning Crash Course. Available at: https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc (Accessed: 02-07-2019).

[99] A. Hamilton, “Cost Drivers in the Development and Validation of Biomarkers Used in Drug Development,” 2018.

[100] S. L. Dearholt and D. Dang, *The Johns Hopkins Nursing Evidence-Based Practice Model and Guidelines*, vol. 39. 2009.

[101] E. G. Carmines and R. A. Zeller, *Reliability and Validity Assessment*. Sage Publications, Inc., 1979.

[102] I. Zechmeister-Koss and A. Kisser, “Procedural Guidance for the Systematic Evaluation of Biomarker Tests,” 2014.

[103] A. Bruton, J. H. Conway, and S. T. Holgate, “Reliability: What is it, and how is it measured?,” vol. 86, no. 2, pp. 94–99, 2000.