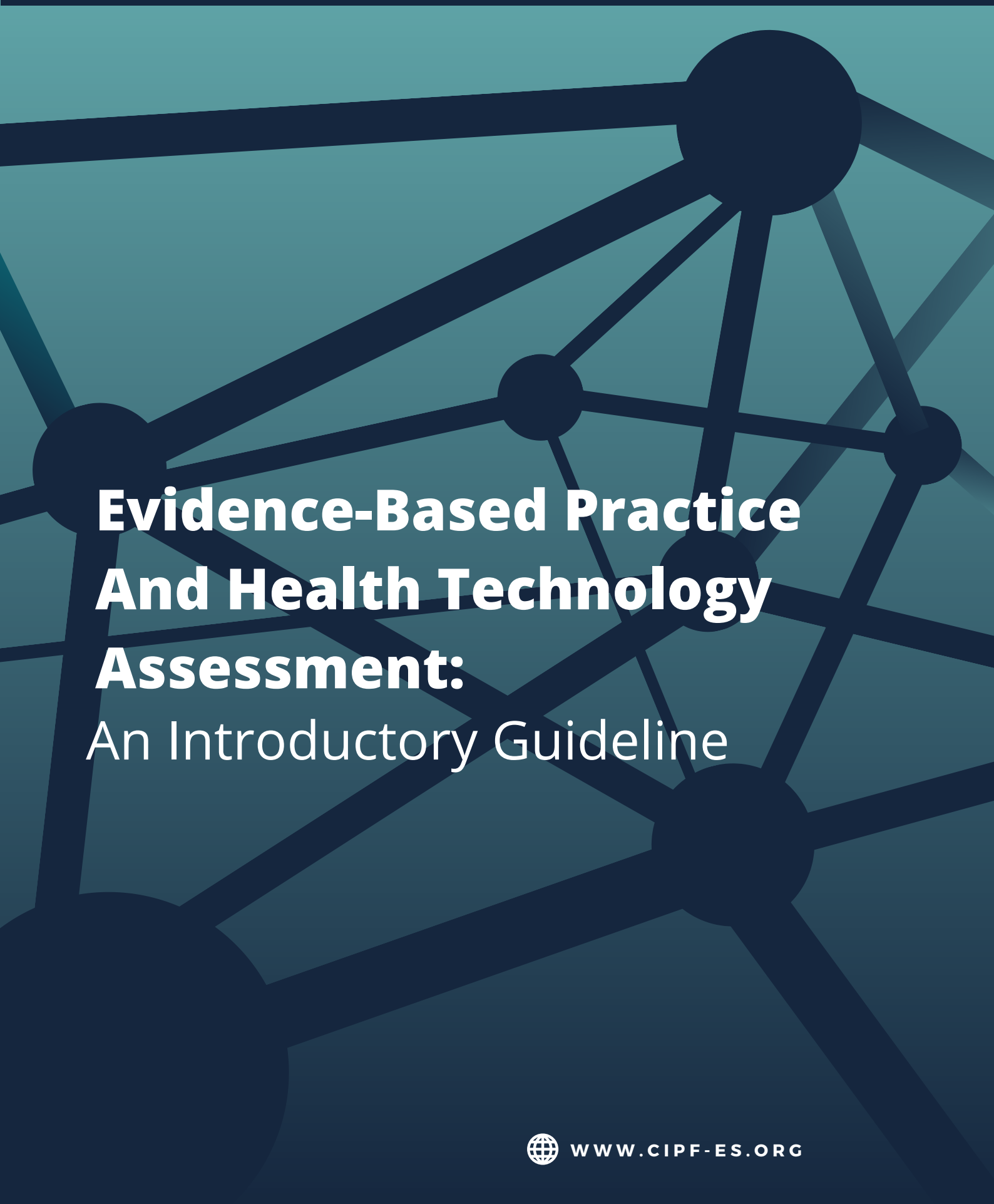
Editors:
Fernanda Stumpf Tonin
Fernando Fernandez-Llimos



Evidence-Based Practice And Health Technology Assessment:

An Introductory Guideline

WWW.CIPF-ES.ORG

Editors
Fernanda Stumpf Tonin
Fernando Fernandez-Llimos

Evidence-Based Practice And Health Technology Assessment:

An Introductory Guideline

1st Edition

2021

cipf

WWW.CIPF-ES.ORG

Evidence-Based Practice and Health Technology Assessment: An Introductory Guideline

2021

Editors

Fernanda Stumpf Tonin: PhD in Pharmacy (University of Lisbon, Portugal) and PhD in Pharmaceutical Sciences (Federal University of Paraná, Brazil). Visiting Research at University Technology of Sydney (Australia). HEOR consultant and scientific adviser. Member of the Editorial-board of Pharmacy Practice. Vice-chair of the Pharmacy Practice Special Interest Group (SIG) of International Pharmaceutical Federation (FIP).

Fernando Fernandez-Llimos: PhD in Pharmacy (University of Granada, Spain). Professor at the Faculty of Pharmacy (University of Porto, Portugal). Researcher at the HIS-EHR: Health Information Systems & Electronic Health Records of the CINTESIS – Center for Health Technology and Services Research (U.Porto). Editor-in-Chief of Pharmacy Practice.

Evidence-Based Practice and Health Tecnology Assessment: An Introductory Guideline.
Editors: Fernanda Stumpf Tonin, Fernando Fernandez-Llimos. CIPF, 2021

ISBN 978-84-124195-0-4

Published by: CIPF
www.cipf-es.org
CIPF Headquarters
Rua das Regateiras, 55
36800 Redondela, Spain

Cover and graphic project: Cesar Casas

Authors

Alexandre F. Cobre: PhD candidate in Pharmaceutical Sciences (Federal University of Paraná, Brazil). Research interests: epidemiological studies, bioinformatics, and statistical methods. MSc in Pharmaceutical Sciences (Federal University of Paraná, Brazil). Bachelor of Pharmacy (Lúrio University, Nampula, Mozambique).

Aline A. Garabeli: PhD candidate in Pharmaceutical Sciences (Federal University of Paraná, Brazil). Research interests: pharmacist-led interventions and evidence-based healthcare. MSc in Pharmaceutical Sciences (Federal University of Paraná, Brazil). Bachelor of Pharmacy (State University of Ponta Grossa, Brazil).

Aline F. Bonetti: PhD candidate in Pharmaceutical Sciences and substitute professor in Pharmacy (Federal University of Paraná, Brazil). Clinical Pharmacy and Hospital Pharmacy specialist (Clinics Hospital, Curitiba, Brazil). Member of the Permanent Education group of the Federal Pharmacy Council (Brazil).

Beatriz Böger: PhD in Pharmaceutical Sciences (Federal University of Paraná, Brazil with internship period at University of Trier, Germany). Specialist in Biology, Genetics and Forensic Toxicology. HEOR consultant with interests on evidence synthesis and budget impact analyses in oncology.

Daniela O. de Melo: PhD in Pharmaceutical Sciences (University of São Paulo, Brazil). Professor at the Federal University of São Paulo, Brazil. Coordinator of the Chronic Disease and Informed Decision (CHRONIDE) and of the Center for Health Technology Assessment Unifesp D (NATS UNIFESP D).

Danielly Chierrito: PhD in Pharmaceutical Sciences (State University of Maringá, Brazil). HEOR consultant and postdoctoral researcher with interests on evidence synthesis and evaluation of biological activities of medicinal plants and cell cultures.

Mariana M. Fachi: PhD and postdoctoral researcher in Pharmaceutical Sciences (Federal University of Paraná, Brazil). Clinical Pharmacy Specialist (Hospital Pequeno Príncipe, Curitiba, Brazil). Leader of HTA and HEOR office with interests on biostatistics, epidemiological studies' design, evidence synthesis and pharmacoeconomics.

Ignacio Aznar-Lou: PhD in Pharmacy (University of Barcelona, Spain). Postdoctoral visiting researcher at Southern Denmark University (Denmark) and at University of Lisbon (Portugal). Researcher at Health Technology Assessment in Primary Care and Mental Health (PRISMA group) at Sant Joan de Déu Institut de Recerca (Barcelona, Spain).

Letícia P. Leonart: PhD in Pharmaceutical Sciences (Federal University of Paraná, Brazil). HEOR consultant and postdoctoral researcher in Health Technology Assessment (Federal University of Paraná, Brazil), with interests on research methods for evidence synthesis and pharmacoeconomics in rare diseases.

Patricia M. Aguiar: PhD in Pharmaceutical Sciences (University of São Paulo, Brazil). Clinical Professor in the Faculty of Pharmaceutical Sciences at the University of São Paulo, Brazil.

Roberto Pontarolo: PhD in Sciences-Biochemistry (Federal University of Paraná, Brazil). Professor at the Pharmacy department and Leader of the Research Center of Biopharmacy (CEB) and Group of Studies in Health Technology Assessment (GEATS) from the Federal University of Paraná, Brazil. Member of the advisor board of the Brazilian Regulatory Agency (ANVISA) and of the National Institute of Science and Technology for Pharmaceutical Innovation (INCT-IF).

Tácio de M. Lima: PhD in Pharmaceutical Sciences (University of São Paulo, Brazil with internship at Coimbra University, Portugal). Hospital Pharmacy specialist (Fluminense Federal University, Rio de Janeiro, Brazil). Professor at the Department of Pharmaceutical Sciences (Federal Rural University of Rio de Janeiro, Brazil). Member of the National Forum on University Pharmacies (FNFU).

Vasco M. Pontinha: PhD candidate in Pharmacoeconomics and Health Outcomes program, Department of Pharmacotherapy and Outcomes Science (Virginia Commonwealth University, United States of America). MBA from Catolica Lisbon School of Business and Economics (Lisbon, Portugal).

Table of contents

EVIDENCE-BASED PRACTICE: MAIN CONCEPTS AND DEFINITIONS

MARIANA M. FACHI
DANIELLY CHIERRITO

Abstract

Healthcare professionals are constantly involved with scientific information and should be able to carefully select resources and keep updated on new literature and new tools to address a variety of healthcare decisions. Evidence-based practice is a key-element and indicator of high-quality patient care. The evidence' strength depends, among others, on the study design and rigor of the methods used. Epidemiological studies are the main source of information, broadly classified into primary or secondary sources. Studies can also be divided into observational or interventional studies, according to the researcher's role. Observational studies are usually classified into descriptive studies (e.g. case or series report), or analytical studies (e.g. cross-sectional studies, ecological studies, case-control and cohort studies). Among the experimental or interventional studies, the randomized clinical trials stand out, as these are gold-standard models to evaluate the effects of a health technology in a given clinical setting. Secondary studies (e.g. narrative review, scope review and systematic review) synthesize information from primary studies, aiming at reducing the selection bias. The overall scientific evidence is classified into hierarchical levels according to the credibility (quality) of the information. This chapter provides an overview of the main concepts of evidence-based healthcare aiming at enabling healthcare professionals, students, and researchers to search, synthetize and critical analyze different epidemiological studies and clinical evidence.

Keywords: evidence-based health; study design; epidemiological studies

MARIANA M. FACHI
DANIELLY CHIERRITO

Introduction

The scientific literature actively contributes to the dissemination of a large number of information, constantly redefining knowledge and making it available in different formats. High quality scientific evidence is an important tool to support decision-making by healthcare professionals and other players. These decisions range from the definition of the best therapeutic approach to be delivered to a patient to discussions on the incorporation of health technologies (e.g. drugs, vaccines, medical procedures, services) in a given scenario. In this sense, evidence-based practice is defined as a multidisciplinary approach that seeks to combine the best available evidence with professional's experience and clinical practice, and patient's needs and values. This approach uses tools from clinical epidemiology, statistics, information technology and scientific methodology to synthetize data on a given topic, also assisting in the interpretation of clinical evidence [1, 2].

Nonetheless, conflicting information on the definition of scientific evidence, how to access and interpret epidemiological studies, and how to translate the results into practice still exist. Additionally, in the past decades, healthcare professionals have faced a large volume of published information, which makes almost impossible to be constantly updated.

Both primary and secondary studies, often classified into hierarchical levels according to the quality of evidence generated, are the main sources of scientific evidence. These studies can also be classified according to their methodological designs. Primary studies are broadly divided into observational studies - descriptive studies (case reports and series) or analytical studies (such as ecological, cross-sectional, case-control and cohort studies), and experimental

trials. Secondary studies address non-systematic reviews, such as narrative and scope reviews, and systematic reviews with or without meta-analyses [3, 4].

This chapter addresses the main concepts on evidence-based practice, including the classification of primary and secondary studies, their hierarchical levels, and degrees of recommendation.

Definition of 'information'

The concept of 'information' is used in almost every scientific discipline within its own different context. In the healthcare, 'information' may be defined as the provision of unbiased, evidence-based, and critically evaluated data and experiences [1, 2]. The access to the most relevant, updated, user specific and objective information is paramount to make appropriate decisions (e.g. prescription, dispensing and use of drugs), and to inform, underpin, or shape scientific research [3]. Information is available in different forms, both printed and electronic, and they vary according to the needs of end-users.

The term 'evidence' (from the Latin evidentĭa.ae: visibility, clarity, transparency, proof, foundation) is defined as 'anything that does not give rise to doubt, generally used in science to refer to elements that support or refute a theory, hypothesis or idea' [5]. It can be used as a synonym for "proof", that is, everything that is used to define the veracity of a situation [1, 2].

History of evidence-based practice

The recent advances in communication technology and internet access allowed a faster and broader dissemination of information worldwide. More than 1,000,000 biomedical journal articles are published annually, most of them with online access, which turns impossible to healthcare professionals to keep up

to date to the most reliable evidence. This 'paradox of the information' reflects the apparent contradiction that the more information we have access to, the more difficult is to use it.

The definition of the term 'information literacy' has been articulated by a range of models and terminology developed by both academics and librarians. According to the American Library Association (ALA), 'information literacy' is the ability to recognize when information is needed, then to locate and evaluate the appropriate information, and finally to use it effectively and responsibly. Because information now comes in many different formats, both printed and electronically, and its quality varies enormously, healthcare professionals need to develop the cognitive and transferable skills to be able to work efficiently with information. This includes an understanding, amongst others, of the resources available, how to find information and evaluate results, how to work with or exploit results, ethical and responsible use, and how to communicate or share findings [1, 6].

In this context, in the early 1990s the North American physician David L. Sackett firstly defined the concept of 'evidence-based medicine' - later named 'evidence-based practice' as 'the conscientious, explicit, and judicious use of the best and most up-to-date evidence for decision-making about the care of individual patients' [1, 7, 8]. This concepts also considers the individual clinical expertise and patients' values and choices to guide decision-making in the healthcare field.

Evidence-based practices are mainly grounded on scientific research, using different tools from clinical epidemiology, statistics, information technology and scientific methodology. The ultimate goal of this approach is to present information able to support decision-making process by health

professionals and other players minimizing unsafe or risky practices and maximizing benefits to patients (i.e. better health outcomes).

In this context, the critical thinking is a key element for evidence-based practice, whose practice can be summarized in five crucial steps: 1. formulate an answerable research question based on the healthcare professional needs; 2. detect the best evidence for answering the research question; 3. to critically assess the retrieved evidence and its value; 4. apply the findings into clinical practice / decision-making; 5. assess the performance [9, 10].

Over the years, the evidence-based practice that was initially directed to decisions at the individual clinical level (focus on the patient), as stated by David L. Sackett, has now expanded to public health scenarios and development of healthcare policies, clinical guidelines, and further healthcare protocols both at national and international levels. In this context, the process known as Health Technology Assessment (HTA) emerges in order to support decisions on the safe use and financing of technologies in a given setting [8, 10].

Epidemiological studies: advantages and limitations

Primary and secondary epidemiological studies are the main source of information for generating evidence, as they include the surveillance, analyses, and experimental factors (physical, biological, social, cultural, or behavioral) that can influence on patients' health. The selection of the most appropriate design for an epidemiological study, the interpretation of the obtained results and the use of the final evidence in practice are important steps that involve the knowledge of concepts in epidemiology. Each epidemiological design has advantages and disadvantages, and researchers must take into account all potential sources of bias (error) and confusion and try to minimize them. The

basic elements of epidemiological studies are the population to be studied (sample); the intervention, exposure or condition under study (called the 'main variable'), presence or absence of comparators (controls); time horizon of the study (period in which it occurs); the results to be measured (outcomes) and possible confusing variables [4, 11, 12].

Initially, primary studies can be classified as observational (i.e. the researcher is limited to observe the evolution of the study variables) or interventional, also called experimental, in which the main variable under study is introduced by the researcher [13-15]. Secondary studies, on the other hand, gather or summarize information from primary studies [11].

Furthermore, studies can also be classified according to the presence or absence of a control group (which allows comparison between the groups under study), or also according to the temporal nature of the study design (which may be retrospective or prospective). Retrospective studies are those in which all data are collected from the past (they already occurred before the study started) by recording that moment or asking participants to remember these data, therefore they are more susceptible to bias. In prospective studies, the exposure may have already occurred, but the outcome has not yet occurred, that is, the study is conducted in the "present", following the participant over time, collecting process data for later analysis in the "future" [4, 17]. A summary of studies' classification is depicted in Figure 1.

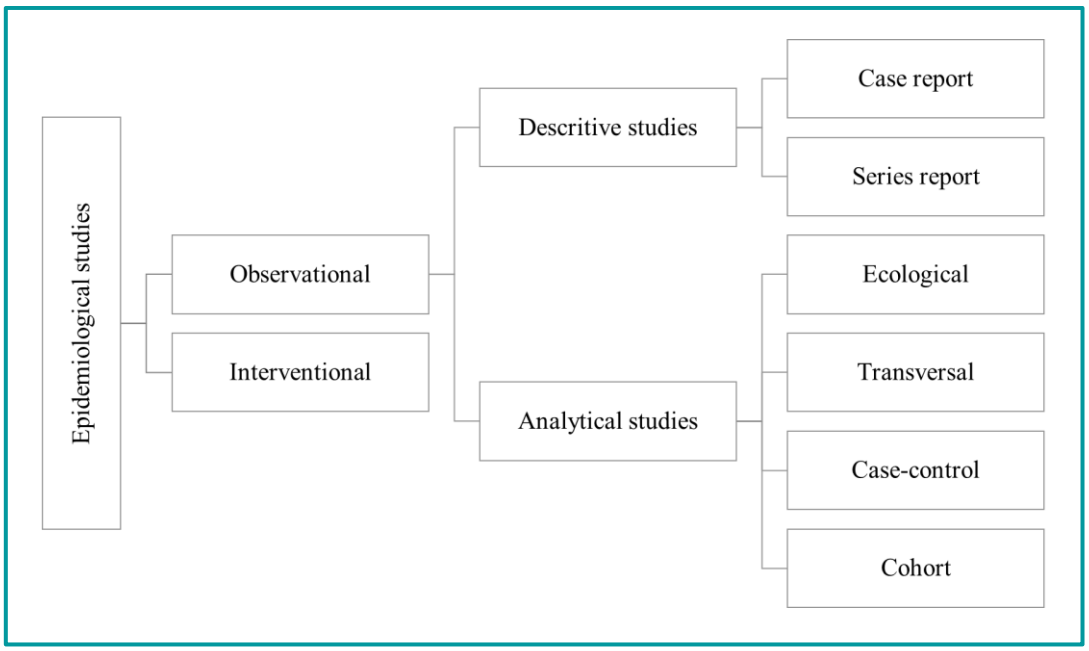


Figure 1. Classification of the study designs

Primary studies

Observational studies

Observational studies, also called real-world studies, are carried out to assess the course of a disease or the relationship between risk factors (exposures) and outcomes, thus making it possible to evaluate the prevalence, natural history, etiology or risk groups of a disease or health-related conditions [12, 18]. Observational studies are classified into descriptive - which are limited to describing the occurrence of a condition in a population (e.g. case reports, series of cases), and analytical studies - which address the relationships between health status and other variables, aiming at evaluating potential associations between exposure and a health-related disease or condition [4, 19]. The latter include cross-sectional (or sectional or prevalence), ecological (correlation), case-control (case-reference) and cohort (longitudinal or follow-up) studies [11, 19].

- **Case reports and case series**

Case reports and case series are detailed retrospective, qualitative descriptions of clinical cases, including information about the signs, symptoms, and other characteristics of the patient. They can also address the therapeutic procedures used. The main difference between case reports and case series consists in the number of reported cases (i.e. a case report usually includes less than three cases) [21]. Both designs do not have a comparator group. When compared to other study designs, they occupy the lowest position in the hierarchically pyramid of evidence as they are prone to researchers' bias and may not be representative or generalizable to the population [4, 19].

Nonetheless, this type of design is usually simple, fast, and less expensive to perform. Case reports or series are important to discuss the pathophysiological mechanisms of a disease or health-related condition – especially rare or orphan diseases, including in-depth analyzes or experimental investigations of a person or group in a real-world environment. This kind of study can initially identify a new condition or adverse healthcare event, contributing to the acquisition of additional knowledge and fostering the conduction of further studies [4, 19-21].

- **Ecological studies**

One of the most basic observational study is an ecological study. This study design compares clusters of people, usually grouped based on their geographical location or temporal associations. Ecological studies assign one exposure level for each distinct group and can provide a rough estimation of prevalence of disease within a population. They are usually retrospective and used in public health research especially when data is unavailable at the individual level or when large-scale comparisons are needed to study the effect

of population-level exposures on a disease condition [4, 19]. The results of these studies can ground future investigations of individuals' behavior (i.e. starting point for hypothesis generation). However, one should be aware about the different confounding factors that can be associated with ecological studies. There are inherent potential weaknesses with this approach, including loss of data resolution and potential misclassification. Quantification of both the number of cases and the total population can be difficult, leading to error or bias. Lastly, due to the limited amount of data available, it is difficult to control for other factors that may mask or falsely suggest a relationship between the exposure and the outcome. [12, 17].

- **Cross-sectional study**

Cross-sectional studies are also called prevalence studies because one of the main measures available is study population prevalence (i.e. the proportion of individuals with the disease at a given point in time vs. the total number of surveyed individuals at that time). They can also be used for network analysis of associations and may represent a first stage of a cohort study or clinical trial. Cross-sectional studies are retrospective studies that provide an overview of the characteristics of the study subjects at a single point in time (i.e. there is no follow-up; exposures and disease outcomes are assessed at the same time). A common cross-sectional study type is the diagnostic accuracy study [19].

Cross-sectional study samples are selected based on their exposure status, without regard for their outcome status. Outcome status is obtained after participants are enrolled. Ideally, a wider distribution of exposure will allow for a higher likelihood of finding an association between the exposure and outcome if one exists. Measures of risk for the exposure-outcome relationship that can be

calculated in cross-sectional study design are odds ratio, prevalence odds ratio, prevalence ratio, and prevalence difference. Cross-sectional studies are relatively inexpensive and have data collected on an individual which allows for more complete control for confounding. Additionally, cross-sectional studies allow for multiple outcomes to be assessed simultaneously.

However, this design is unable to provide a cause-effect relationship, as it does not prove the existence of a temporal sequence between exposure and outcome. Cross-sectional studies are not suitable to evaluate rare diseases nor measure the incidence of a clinical condition [12, 14, 16].

- **Case-control studies**

Case-control studies can assess the degree of associations between various risk factors (possible exposures) and results (outcomes) [12]. If the factor is associated with the disease, the proportion of the factor between cases will be greater than the same proportion between controls. This design is commonly applied for rare diseases, which have longer latency periods [12].

In this type of study there are two groups of patients: cases (patients who have a specific disease or health-related condition and are commonly identified in hospitals, clinics or healthcare services) and the control group (patients who do not have the disease or specific condition) [12, 16]. The selection of the control group is a critical step as, ideally, they should have similar characteristics to the case group. The strategy to be adopted for the selection of the control group depends on the objective of the study. One widely used approach is the propensity score, in which it is possible to balance or pair the groups of individuals based on some covariables or characteristics [22, 23].

The main advantages of the case-control design are the low associated cost, the efficiency and the time required to conduct the study, as it is possible to analyze several predictors simultaneously [4, 12]. However, the main limitation of these studies refers to the retrospective design, which requires an adequate recording of information or the patient's memory of the event [12, 24].

The results of the case-control studies are expressed in odds ratio (OR), which is an estimate of the ratio of the morbidity between exposed and not exposed individuals to the risk factor [11], as shown in the following Equation 1:

$$\text{Odds Ratio} = \frac{(A/C)}{(B/D)} \quad (\textit{Equation 1})$$

Where 'A' is the number of cases with the risk factor, 'C' is the number of cases without the risk factor, 'B' refers to the number of controls with the risk factor and 'D' to the number of controls without the risk factor. OR values greater than 1 indicate that the exposure is positively related to the health outcome (disease), which may imply a cause-effect relationship [25].

Still, within case-control studies is possible to have 'cross-case studies', in which an individual in the case group can act as its own control, thus minimizing some potential confounders. In this specific design, cases are assessed for their exposure status immediately before they became a case, and then these are compared to their own exposure at an earlier point where they did not become a case. The selection of the previous point is often at random or depends on an average measure of exposures over time. The main limitation in this design is the memory bias, since the study participants are more likely to remember an exposure before it becomes a case [4, 12, 17].

- **Cohort studies**

Cohort studies are longitudinal studies, appropriate for answering questions about the causes or prognosis of the disease, as well as the incidence of a disease. In this design, the objective is to establish a causal link between the factor to which the group was exposed and the final outcome. Cohort studies can be prospective or retrospective [4].

In this design, patients are usually divided into two groups based on their exposure status. These cohorts are monitored over time to evaluate the natural history of a disease, assess the prognosis compared to the treatment or investigate the disease etiology. Whereas the cohort study is concerned with frequency of disease in exposed and non-exposed individuals, the case-control study is concerned with the frequency and amount of exposure in subjects with a specific disease (cases) and people without the disease (controls). Cohort studies allow to estimate the incidence rates and relative risks. They are usually more costly and require longer driving times when compared to case-control studies, as they can address various outcomes at the time [4, 12, 14, 16, 26].

However, cohort studies are more prone to selection and attrition bias (to (losses) and are not appropriate to evaluate rare diseases [12, 14].

The cohort study is the only observational study that allows calculating the incidence. In addition, because the first part of the study is similar to a cross-sectional study, it is possible to measure the point prevalence, and the prevalence over a period. In a cohort study, the results can be expressed in OR, prevalence ratio, rate ratio, relative risk, risk ratio and hazard ratio [17]. Among these measures, the relative risk (RR) is one of the most used due to its easier interpretation (Equation 2). The RR is defined as the ratio between the incidence

of the outcome in the exposed group and the incidence of the outcome in the non-exposed group [4, 14, 16].

$$\text{Relative risk} = \frac{A/(A+B)}{C/(C+D)} \quad (Equation\ 2)$$

Where 'A' is the number of exposed people with the outcome, 'B' is the number of exposed people without the outcome, 'C' is the number of unexposed people with the outcome and 'D' is the number of unexposed people without the outcome. If the RR is greater than 1, it means that the risk to develop the outcome in the exposed individuals is greater than the risk in the unexposed individuals, which represents a positive association between the exposure of interest and the outcome of interest [25].

Interventional or experimental studies

Interventional studies, also called experimental studies, are those where the researcher intercedes as part of the study design. The most common and strongest interventional study design is a randomized controlled trial, however, there are other interventional study designs, including pre-post study design, non-randomized controlled trials, and quasi-experiments. Studies that only evaluate one single intervention are called "single arm trials", while those that have one or more comparators are called "controlled trials". In controlled clinical trials, the division of patients between study groups can be done by a randomization process [17, 19].

Experimental studies are used to evaluate study questions related to either therapeutic agents or prevention. Therapeutic agents can include prophylactic agents, treatments, surgical approaches, or diagnostic tests. Prevention can

include changes to protective equipment, engineering controls, management, policy, or any element that should be evaluated as to a potential cause of disease or injury [16].

- **Pre–post study (before–after design)**

A pre-post study measures the occurrence of an outcome before and again after a particular intervention is implemented. Pre-post studies may be single arm, one group measured before the intervention and again after the intervention, or multiple arms, where there is a comparison between groups. Often there is an arm where there is no intervention. The no-intervention arm acts as the control group in a multi-arm pre-post study. These studies have the strength of temporality to be able to suggest that the outcome is impacted by the intervention, however, pre-post studies do not have control over other elements that are also changing at the same time as the intervention is implemented. Thus, changes in disease occurrence during the study period cannot be fully attributed to the specific intervention. Outcomes measured for pre-post intervention studies may be binary health outcomes such as incidence or prevalence, or mean values of a continuous outcome such as systolic blood pressure may also be used. The analytic methods of pre-post studies depend on the outcome being measured. If there are multiple treatment arms, it is also likely that the difference from beginning to end within each treatment arm are analyzed [17, 19].

- **Interventional studies without concurrent controls**

When a new intervention is available, it is possible to compare the result obtained after using the intervention with a similar group of people followed in the past without this treatment (i.e. historical controls). This study presents a high risk of bias, as it is susceptible to differences in the severity of the disease or other

factors in both groups and through the improvement over time in the available supportive care [17, 19].

- **Non-randomized clinical trials**

Non-randomized trials are interventional study designs that compare a group where an intervention was performed with a group where there was no intervention. These are convenient study designs that are most often performed prospectively and can suggest possible relationships between the intervention and the outcome. The selection of the group in which the patients will enter is not performed at random, instead, it occurs according to researcher's convenience, or considering the participant's access to the intervention. These study designs are often subject to many types of bias and error and are not considered a strong study design [17, 19].

- **Factorial study design**

The factorial design allows two or more interventions to be performed in the same study without the inclusion of further patients. It also allows the evaluation of whether a combination of the interventions is more effective than the intervention alone. For example, in a simpler factorial design (2x2), participants can be allocated to the following groups: intervention group 1, intervention group 2, intervention group 1 and 2 and group without interventions. Results from both interventions compared to the control can be obtained, which allows to better understand the effect of the interactions between the two treatments [17, 19].

- **Randomized controlled trials**

Randomized controlled trials (RCTs) are the most common type of interventional study, being the gold-standard design to evaluate the cause-effect relationships of interventions and diseases with minimal bias and confounding factors. Confounding factors are those resulting from the non-random distribution of the risk factor both in the population and in the sample that can lead to a misleading estimation of the effect [16, 27].

Yet, several modifications on the standard RCT design can be performed. These trials take a homogenous group of study participants and randomly divide them into two separate groups [12, 16, 28]. Randomization is performed to reduce the systematic differences between the two groups regarding prognostic factors, so that any difference in the results can be reasonably attributed to the effect of the intervention itself. The randomization process can be performed in different ways, including randomization table, coin toss, *software* or randomization center [29]. If the randomization is successful then these two groups should be the same in all respects, both measured confounders and unmeasured factors. The intervention is then implemented in one group and not the other. Both groups are followed prospectively for a specific time interval and afterwards are compared regarding the outcomes of interest. Theoretically, the only difference between the two groups through the entire study is the intervention [12].

Additional methodological elements are utilized among RCTs to further strengthen the causal implication of the intervention's impact. These include allocation concealment, blinding, measuring compliance, controlling for co-interventions, measuring dropout, analyzing results by intention to treat, and assessing each treatment arm at the same time point in the same manner [12]. The blinding process prevents the participants of the study from knowing which

group they belong to, guaranteeing impartiality in the results. According to the blinding process, trials can be classified into: open-label (i.e. all participants and researchers know which intervention is being administered), single-blinded (i.e. when only the patient or the researcher is unaware of the intervention), double blind (both the patient and the observer are do not know about the intervention) and triple-blind (both the observer, the patient, the statistician or the data analyst do are unaware of the groups of the study) [17, 30].

Nonetheless, the RCT also has limitations. It is a very expensive long-term study with limited capacity of data generalization, as the study is restricted to a specific population. Additionally, depending on the clinical condition being assessed, attrition bias may exist (e.g. high rates of losses), requiring large samples of patients [31].

To minimize potential errors, it is paramount that the RCT is well designed, clearly specifying the type of hypothesis tested and the procedures to be used for the analysis of primary outcomes. Among the types of hypotheses evaluated in the study, superiority, non-inferiority, or equivalence between different interventions can be assessed. In a superiority trial, it is assessed whether a treatment intervention is superior (i.e. better than) the control. The aim of a non-inferiority trial is to assess whether a treatment is not worse than the conventional or control treatment. Although this non-inferiority study is not used to prove treatment effectiveness, this method has advantages in situations where a new intervention may be less expensive, less invasive and have fewer adverse events. The equivalence study assesses whether interventions are similar [14, 32, 33].

RCTs can still be classified as with a parallel or cross-over design. After randomization, the subjects can be assigned to receive the interventions

throughout the entire study period (parallel design), or first they are treated with an intervention then switched to the other intervention (cross-over design) [28].

- **Cross-over study design**

A crossover RCT is a type of interventional study design where study participants intentionally 'crossover' to the other treatment arm. This should not be confused with the observational case-crossover design. A crossover RCT begins the same as a traditional RCT, however, after the end of the first treatment phase, each participant is re-allocated to the other treatment arm. There is often a wash-out period in between treatment periods. This design has many strengths, including demonstrating reversibility, compensating for unsuccessful randomization, and improving study efficiency by not using time to recruit subjects [19, 28].

The main advantages of this study are that each participant is his own control, minimizing effects of inter-individual variability. In addition, a smaller number of patients is usually required when compared to a traditional RCT. However, this design can be used only for stable or incurable diseases or health-related, where the intervention provides only temporary effects [19, 28].

- **Cluster clinical trials**

Cluster randomized trials differ from individually RCT in that the unit of randomization is something other than the individual participant or patient (e.g. healthcare centers, hospitals, pharmacies) [34]. This study design is in common use in areas such as education and public health research; they are particularly well suited to testing differences in a method or approach to patient care (as opposed to evaluating the physiological effects of a specific intervention) [35-37].

The disadvantages compared to the randomized controlled trials individually include greater complexity in design and analysis, especially considering the difficulty of recognizing potential sources of confusion, and the need for more participants to obtain the similar statistical power when compared to a traditional RCT [35-37].

Phases of clinical trials

Clinical trials are the main source of evidence on the effects (e.g. efficacy, safety) of a new intervention. After the pre-clinical phase, the phases of clinical research are those in which scientists conduct experiments with humans. The incorporation or approval of a new technology by a national regulatory agency requires the presentation of the results of Phases I, II and III trials; phase IV trials are performed out after marketing [15, 38]:

- **Phase I:** the intervention is evaluated in a small group of volunteers, usually in individuals who do not have the studied disease or health-related condition. The objective is to assess the safety, tolerability, and pharmacokinetics of the product and, when possible, determine its pharmacodynamic profile. At this stage, it is also possible to set the highest dose of the new treatment that can be safely administered without causing serious adverse events. Although pre-clinical research generally provides some general information about the dosage, the effects of a drug on the human body can be unpredictable, that is why this clinical phase I is important to be performed. Around 70% of the products are considered safe after phase II and can continue to phase II trial.
- **Phase II:** they are performed with patients diagnosed with the studied disease or health-related condition. The aim of this phase is to provide data

on both short-term safety, dose-response, and primary efficacy of the product. A small number of patients is included in this phase (around 50-300). Only products presenting promising results in this phase can continue to phase III trials.

- **Phase III:** a larger number of patients diagnosed with the studied disease or health-related condition is included in this phase (300-3,000). The purpose of this trial is to assess the effects of the new drug compared to others available on the market for the same condition. When possible, both short and long-term risk/benefit ratios and the added therapeutic value of the product are evaluated. At this stage, the type and profile of the most frequent adverse drug reactions are investigated. This trial is usually randomized and double-blinded. Around 25-30% of the products tested in a phase III trial are approved.
- **Phase IV:** this phase is usually performed after the product first approval. It includes a higher number of patients that are followed in a real-world setting (observational design) for long periods. This study allows to monitor the long-term effects of the product, including new adverse reactions and effectiveness profile.

Secondary studies

Narrative review

Narrative literature reviews are qualitative studies that have an important role in continuing education as they provide readers with the state of the science of a specific topic or theme from a theoretical and contextual point of view. However, this type of review does not describe the methodological approach that would permit reproduction of data nor answer to specific quantitative research

questions [10, 41]. Because of this selection bias, narrative reviews do not have methodological strength to synthetize evidence and are not recommended approaches to ground health technology assessment processes [10, 41]. Learn more about 'narrative review' on Chapter 03.

Scoping review

A scoping review or scoping study provides an overview or map of a particular subject, especially when it involves the publication of different study designs. This type of review can be conducted to gather the types of evidence available in the literature on the subject of interest, identify the main concepts, factors, or related practices, and point out possible gaps in knowledge [41-43]. See further concepts on 'scoping review' in Chapter 03.

Systematic review

Systematic review is a type of retrospective study that synthesize the available evidence from primary studies to answer a specific research question [10, 46, 47]. It is defined as a review of the evidence on a clearly formulated question that uses systematic and explicit methods to identify, select and critically appraise relevant primary research, and to extract and analyze data from the studies that are included in the review. The main steps of a systematic review include: (1) Formulation of the research question and definition of the eligibility criteria (inclusion and exclusion) of the studies; (2) Search and selection of studies; (3) Data extraction; (4) Assessment of the quality and risk of bias in the included studies; (5) Analysis and presentation of data; (6) Data interpretation and (7) Improvement and updating of the review [46]. These steps are detailed in Chapter 03.

Standards for the conduction and reporting of studies

Several validated guidelines and standards for conduction and reporting epidemiological studies are available worldwide and should be strictly followed by authors, researchers, and other evidence end-users to enhance transparency and data reproducibility [39, 40].

The CONsolidated Standards of Reporting Trials (CONSORT) statement consists of a 25-item checklist and a flowchart that provide guidance to authors on how to report a RCT. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) checklist is comprised of 22-items tool that must be considered in when performing observational studies. Other checklists according to study's design such as The Standards for Reporting of Diagnostic Accuracy (STARD) for studies of diagnostic accuracy and CAse REports (CARE) guideline for case reports are also available [39, 40]. For conducting scoping reviews and systematic reviews recommendations from the Cochrane Collaboration, The Joanna Briggs Institute and the checklists from the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) can be used [43, 44, 46, 47]. To learn more about reporting guidelines, check the Equator Network (Enhancing the QUAlity and Transparency Of health Research) platform (https://www.equator-network.org/). This is an international initiative aimed at promoting transparent and accurate reporting of health research studies to enhance the value and reliability of medical research literature.

Hierarchy of evidence

Scientific evidence is classified into hierarchical levels (represented by a 'pyramid of evidence' according to the credibility (quality) of the information. The base of the pyramid is formed by primary studies with lower methodological

quality, such as expert opinion, case reports and series – which are more prone to bias. Next are cohorts and RCT. All of these primary studies are labeled as 'unfiltered information' as they do not provide a critical evaluation and recommendation about a topic. On the top of the pyramid there are secondary studies, such systematic with or without meta-analyses, critical assessments, decision analyses and economic evaluations. These are considered sources of higher quality information since the evidence is filtered and carefully evaluated. (Figure 2) [3, 46, 47].

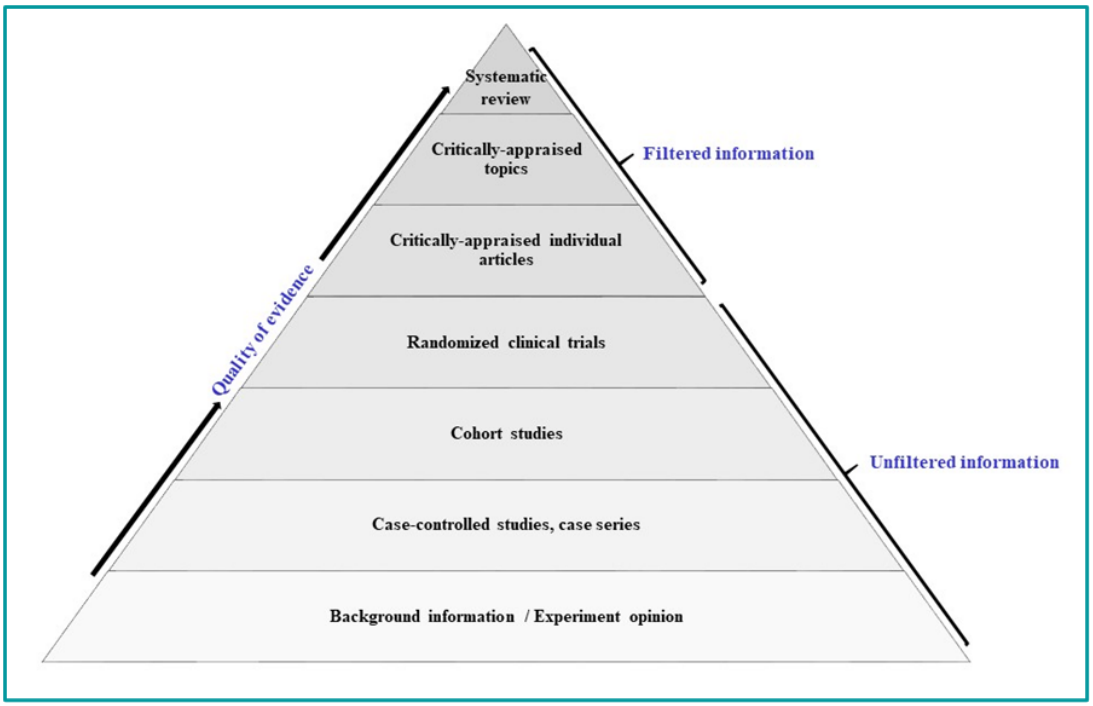


Figure 2. Evidence pyramid
(Adapted from Garattini, 2016)

Recently, new models for the evidence pyramid are being discussed considering that the evidence constantly cross between the different levels (i.e. transitivity). For instance, depending on the clinical context, an observational study design may provide more robust data than a RCT (e.g. long-term safety of

a drug) and so on. In addition, authors consider that the secondary studies ('filtered information') should not be part of the pyramid; instead, they can be considered as a 'magnifying glass' that analyze the evidence from the 'unfiltered information' pyramid [48].

Levels of evidence

The information generated by primary and secondary studies can be classified into degrees of recommendation and levels of evidence. According to the Oxford Center for Evidence-Based Medicine (CEBM) model, the following degrees that should be considered are [49]:

A: consistent level 1 studies

B: consistent level 2 or 3 studies or extrapolations from level 1 studies

C: level 4 studies or extrapolations from level 2 or 3 studies

D: level 5 evidence or troublingly inconsistent or inconclusive studies of any level

Table 1 summaries the degrees of recommendation and levels of evidence according to the CEBM [49]. This classification is an important part of the evidence-based practice, as it helps the reader to interpret the data, grade the evidence, prioritize the information, and used into practice [49, 50].

For inconclusive studies (e.g. RCT with a wide confidence interval and heterogeneous meta-analyses), or in cases where the statistical analyses between the comparative groups is lacking, the evidence should be downgraded. The minus sign (-) should be added after the level of evidence and the degree of recommendation should be rated as 'D' [49].

MARIANA M. FACHI
DANIELLY CHIERRITO

Table 1. Degrees of recommendation and levels of evidence

Degree of Recommendation	Level of Evidence	Study characteristics
A	1 A	Systematic review with meta-analysis of controlled and randomized clinical studies, with homogeneity
	1 B	Controlled and randomized clinical trial with a narrow confidence interval (95% CI)
	1 C	All-or-nothing therapeutic results
B	2 A	Systematic review with meta-analysis of cohort studies, with homogeneity or extrapolations from level 1 studies
	2 B	Cohort study (including low quality randomized controlled trial) or extrapolations from level 1 studies
	2 C	Observation of therapeutic results and ecological studies or extrapolations from level 1 studies
	3 A	Systematic review with meta-analysis of case-control studies, with homogeneity or extrapolations from level 1 studies
	3 B	Case-control study or extrapolations from level 1 studies
C	4	Low quality case reports or analytical observational studies (cohort and case-control), or extrapolations from level 2 or 3 studies
D	5	Expert opinion without explicit critical assessment, based on consensus, physiological studies or animal models, or inconsistent / inconclusive studies of any level

Source: adapted from Center for Evidence-Based Medicine, 2009.

Conclusions

This chapter introduced the role of evidence-based clinical practice as an essential step for guiding decision-making process in healthcare. The appropriate selection, design, and report of the different epidemiological studies – including primary and secondary sources, allow for reliable results and transparent research. Both observational studies (descriptive and analytical studies) and interventional studies (quasi-experimental and randomized clinical trials) have advantages and limitations that should be carefully considered by authors, scientists and healthcare professionals during the conduction of the study and data interpretation. Secondary studies such as scoping reviews and systematic

reviews, when properly performed, are able to summarize the evidence from primary studies aiming at reducing the individual bias. The other chapters of this book further discuss some of points highlighted here and provide practical examples of evidence synthesis in health.

References

[1] Sackett DL, Rosenberg WM, Gray JA, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. BMJ (Clinical research ed). 1996;312(7023):71-2.
[2] Satterfield JM, Spring B, Brownson RC, Mullen EJ, Newhouse RP, Walker BB, et al. Toward a transdisciplinary model of evidence-based practice. The Milbank Q. 2009;87(2):368-90.
[3] Garattini S, Jakobsen JC, Wetterslev J, Bertelé V, Banzi R, Rath A, et al. Evidence-based clinical practice: Overview of threats to the validity of evidence and how to minimise them. Eur J Intern Med. 2016;32:13-21.
[4] Ranganathan P, Aggarwal R. Study designs: Part 1 - An overview and classification. Perspect Clin Res. 2018;9(4):184-6.
[5] Houaiss. Evidência. Dicionário Houaiss da Língua Portuguesa. Objetiva. 2001.
[6] Ioannidis JP. The Mass Production of Redundant, Misleading, and Conflicted Systematic Reviews and Meta-analyses. The Milbank Q. 2016;94(3):485-514.
[7] Akobeng AK. Principles of evidence based medicine. Arch Dis Child. 2005;90(8):837-40.
[8] Djulbegovic B, Guyatt GH. Progress in evidence-based medicine: a quarter century on. Lancet. 2017;390(10092):415-23.
[9] Hatala R, Guyatt G. Evaluating the teaching of evidence-based medicine. Jama. 2002;288(9):1110-2.
[10] Manchikanti L. Evidence-based medicine, systematic reviews, and guidelines in interventional pain management, part I: introduction and general considerations. Pain Physician. 2008;11(2):161-86.

[11] Röhrig B, du Prel J-B, Wachtlin D, Blettner M. Types of study in medical research: part 3 of a series on evaluation of scientific publications. Dtsch Arztebl Int. 2009;106(15):262-8.
[12] Munnangi S, Boktor SW. Epidemiology Of Study Design: StatPearls Publishing, Treasure Island (FL); 2020 2020.
[13] Vetter TR, Mascha EJ. Defining the Primary Outcomes and Justifying Secondary Outcomes of a Study: Usually, the Fewer, the Better. Anesth Analg. 2017;125(2).
[14] Wang B, Wang H, Tu XM, Feng C. Comparisons of Superiority, Non-inferiority, and Equivalence Trials. Shanghai Arch Psychiatry. 2017;29(6):385-8.
[15] Yan F, Thall PF, Lu KH, Gilbert MR, Yuan Y. Phase I;II clinical trial design: a state-of-the-art paradigm for dose finding. Ann Oncol. 2018;29(3):694-9.
[16] Stephenson JM, Babiker A. Overview of study design in clinical epidemiology. Sex Transm Infect. 2000;76(4):244.
[17] Thiese MS. Observational and interventional study design types; an overview. Biochem Med (Zagreb). 2014;24(2):199-210.
[18] Shields KM, DiPietro NA, Kier KL. Principles of drug literature evaluation for observational study designs. Pharmacotherapy. 2011;31(2):115-27.
[19] Aggarwal R, Ranganathan P. Study designs: Part 4 - Interventional studies. Perspect Clin Res. 2019;10(3):137-9.
[20] Noordzij M, Dekker FW, Zoccali C, Jager KJ. Study designs in clinical research. Nephron Clin Pract. 2009;113(3):c218-21.

[21] Sayre JW, Toklu HZ, Ye F, Mazza J, Yale S. Case Reports, Case Series - From Clinical Practice to Evidence-Based Medicine in Graduate Medical Education. Cureus. 2017;9(8):e1546-e.
[22] Månsson R, Joffe MM, Sun W, Hennessy S. On the Estimation and Use of Propensity Scores in Case-Control and Case-Cohort Studies. Am J Epidemiol. 2007;166(3):332-9.
[23] Joffe MM, Rosenbaum PR. Invited commentary: propensity scores. Am J Epidemiol. 1999;150(4):327-33.
[24] DiPietro NA. Methods in epidemiology: observational study designs. Pharmacotherapy. 2010;30(10):973-84.
[25] Vetter TR, Jesser CA. Fundamental Epidemiology Terminology and Measures: It Really Is All in the Name. Anesth Analg. 2017;125(6):2146-51.
[26] Wang JJ, Attia J. Study Designs in Epidemiology and Levels of Evidence. Am J Ophthalmol. 2010;149(3):367-70.
[27] Herrera-Perez D, Haslam A, Crain T, Gill J, Livingston C, Kaestner V, et al. A comprehensive review of randomized clinical trials in three medical journals reveals 396 medical reversals. eLife. 2019;8.
[28] Spieth PM, Kubasch AS, Penzlin AI, Illigens BM-W, Barlinn K, Siepmann T. Randomized controlled trials - a matter of design. Neuropsychiatr Dis Treat. 2016;12:1341-9.
[29] Higgins JPT, Altman DG, Gøtzsche PC, Jüni P, Moher D, Oxman AD, et al. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. BMJ. 2011;343:d5928.
[30] Vasconcelos CEB. O cegamento na pesquisa científica. J Rev Cirur Traumatol Buco-Maxilo-facial. 2016;16(1):5.
[31] Concato J. Study design and "evidence" in patient-oriented research. m J Respir Crit Care Med. 2013;187(11):1167-72.
[32] Lesaffre E. Superiority, equivalence, and non-inferiority trials. Bull NYU Hosp Jt Dis. 2008;66(2):150-4.
[33] Thall PF. A review of phase 2–3 clinical trial designs. Lifetime Data Anal. 2008;14(1):37-53.
[34] Dalma A, Petralias A, Tsiampalis T, Nikolakopoulos S, Veloudaki A, Kastorini CM, et al. Effectiveness of a school food aid programme in improving household food insecurity; a cluster randomized trial. Eur J Public Health. 2020;30(1):171-8.
[35] Puffer S, Torgerson DJ, Watson J. Cluster randomized controlled trials. J Eval Clin Pract. 2005;11(5):479-83.
[36] Bland JM. Cluster randomised trials in the medical literature: two bibliometric surveys. BMC Med Res Methodol. 2004;4:21-.
[37] Murray DM, Varnell SP, Blitstein JL. Design and analysis of group-randomized trials: a review of recent methodological developments. Am J Public Health. 2004;94(3):423-32.
[38] Lipsky MS, Sharp LK. From idea to market: the drug approval process. J Am Board Fam Med. 2001;14(5):362-7.
[39] Bolignano D, Mattace-Raso F, Torino C, D'Arrigo G, Abd ElHafeez S, Provenzano F, et al. The quality of reporting in clinical research: the CONSORT and STROBE initiatives. Aging Clin Exp Res. 2013;25(1):9-15.
[40] Johansen M, Thomsen SF. Guidelines for Reporting Medical Research: A Critical Appraisal. Int Sch Res Notices. 2016;2016:1346026.

[41] Munn Z, Peters MDJ, Stern C, Tufanaru C, McArthur A, Aromataris E. Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. BMC Med Res Methodol. 2018;18(1):143.
[42] Levac D, Colquhoun H, O'Brien KK. Scoping studies: advancing the methodology. Implement Sci. 2010;5:69.
[43] Peterson J, Pearce PF, Ferguson LA, Langford CA. Understanding scoping reviews: Definition, purpose, and process. J Am Acad Nurse. 2017;29(1):12-6.
[44] Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche PC, Ioannidis JPA, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. BMJ. 2009;339:b2700.
[45] Institute TJB. The Joanna Briggs Institute Reviewers' Manual 2015 - Methodology for JBI Scoping Reviews. Adelaide, Australia. The Joanna Briggs Institute. 2015.
[46] Higgins JPT, Green S. Cochrane handbook for systematic reviews of interventions. 2020;Versão 6.1, 2020.
[47] Hartling L, Vandermeer B, Fernandes RM. Systematic reviews, overviews of reviews and comparative effectiveness reviews: a discussion of approaches to knowledge synthesis. Evid Based Child Health. 2014;9(2):486-94.
[48] Murad MH, Asi N, Alsawas M, Alahdab F. New evidence pyramid. Evid Based Med. 2016;21(4):125-7.
[49] Oxford. Oxford Centre for Evidence-Based Medicine: Levels of Evidence (March 2009). 2009.
[50] Burns PB, Rohrich RJ, Chung KC. The levels of evidence and their role in evidence-based medicine. Plast Reconstr Surg. 2011;128(1):305-10.

EVIDENCE-BASED PRACTICE: MAIN CONCEPTS AND DEFINITIONS

MARIANA M. FACHI
DANIELLY CHIERRITO

HISTORY AND STEPS OF HEALTH TECHNOLOGY ASSESSMENT

BEATRIZ BÖGER
MARIANA M. FACHI

Abstract

After the Second World War, the growth in technological innovations resulted in a continuous increase in the stock of healthcare technologies worldwide. A health technology is the application of organized knowledge and skills in the form of devices, medicines, vaccines, procedures, and systems developed to solve a health problem and improve quality of lives. However, the increase in the development and manufacturing of new technologies caused a considerable increment in costs in the health sector. The process called 'Health Technologies Assessment' (HTA) was created grounded on these concerns, aiming at supporting clinical and political decisions regarding the impact and incorporation of health technologies in a given setting in different countries or regions. HTA agencies use standard methods to assess a technology and, thus, ensure consistent evaluations that can guide national and international decisions. Several steps should be followed for the incorporate of a technology: (1) identify of the topic for assessment, (2) clearly specify the problem, (3) gather the available evidence, (4) collect primary data (field evaluation), (5) appraise the evidence, (6) synthesize and consolidate the evidence, (7) conduct an economic evaluation, budget and health systems impact analyses, (8) assess both social, ethical and legal factors, (9) formulate findings and recommendations, (10) disseminate these findings and recommendations and (11) monitor the impact of assessment reports. This chapter describes the history of HTA and highlights the main steps and requirements during the evaluation of a new technology.

Keywords: healthy technology assessment; healthcare evaluation; systematic evaluation

BEATRIZ BÖGER
MARIANA M. FACHI

Introduction

Since the Second World War, an important growth in the development and manufacturing of technological innovation occurred, leading to a continuous increase in the stock of health technologies [1-4]. In addition, advancements in science directly contributed for improving techniques to prevent, diagnose, and treat diseases and other health-related conditions. This resulted in a significant increase in life expectancy worldwide [3-6].

However, the availability of new technologies also caused an important augmentation of health-related costs. The higher health expenditure over the years is justified by the fact that new health technologies are usually more expensive that the older ones. Additionally, one should be aware of the 'cumulative factor of use' of different technologies indicated for the same disease or condition, unlike in other sectors where the inclusion of a new technology tends to replace the previous ones. A 'health technology' is defined by the World Health Organization (WHO) as the application of organized knowledge and skills in the form of devices, medicines, vaccines, procedures, and systems developed to solve a health problem and improve quality of lives [1, 7].

Decisions about the incorporation and use of health technologies may vary according to each country and health system [1, 8]. Nonetheless, several obstacles are commonplace, including the need for balancing the effects and costs of the new technologies, guarantee supply and access to these products, update regulatory instruments, and invest in human and technical resources to perform these activities [4, 5, 9]. The rapid dissemination of information – with hundreds of publications available every day – makes even more difficult to keep updated with the best evidence to ground decision-making process [10, 11].

HISTORY AND STEPS OF HEALTH TECHNOLOGY ASSESSMENT

BEATRIZ BÖGER
MARIANA M. FACHI

The Health Technology Assessment (HTA) is a systematic evaluation of the properties and effects of a health technology, addressing the direct and intended effects of this technology, as well as its indirect and unintended consequences, aiming mainly at informing decision making regarding the benefits or otherwise of the new technology [12-14]. This evaluation is carried out through several studies on the effects of the technologies of interest. Randomized controlled trials (RCT) are usually used to prove the efficacy and safety of the treatment, while observational or real-world studies can help in assessing its effectiveness. Economic evaluations, including cost-effectiveness studies and budget impact analyses are paramount to demonstrate the economic value of the technology, especially considering the limited budget of health systems and the need to effectively allocate resources [12,15].

This chapter aims to present the history of HTA and provide an overview of the main steps for assessing a new health technology.

Definition of 'health technology assessment'

Goodman (1998) conceptualizes HTA as being "(…) a multidisciplinary field of policy analysis, which studies the clinical, social, ethical and economic implications of the development, diffusion and use of technology in health" [12].

The primary purpose of HTA is to contribute to decision-making processes, during the implementation of public policies in health systems, within health services and care practices provided by professionals. For this to occur, HTA needs to provide constant information on the benefits, risks and costs of new technologies and technologies that are already being used [13,14]. In this context, HTA is a continuous process aiming at systematically evaluating the

short and long-term consequences of a given technology or a group of technologies [1–4] to support political or clinical decisions regarding the diffusion and incorporation of these technologies by different players (e.g. healthcare professionals, politicians, managers, and other stakeholders) [16,17].

HTA is a comprehensive and multidisciplinary approach that uses explicit analytical frameworks drawing to assess different stages of the technology life cycle from different perspectives (payers):

- Innovation
- Regulation and diffusion
- Incorporation
- Use/adoption
- Discontinuance/obsolescence

The main attributes of a technology that should be considered during a HTA are its efficacy/effectiveness, safety, and efficiency/appropriateness [18–22]. Figure 1 depicts a hierarchy of the different technologies that can be assessed in healthcare. Technologies that directly interact with patients are called 'biomedical technologies', such as medicines or devices. Medical procedures are part of the training of health professionals and support care delivery. Together with biomedical technologies, these procedures constitute the group of 'medical technologies'. On a broader view, all medical technologies are used within a context that encompasses a structure of technical and administrative support, information systems and healthcare provision workflow [3, 23].

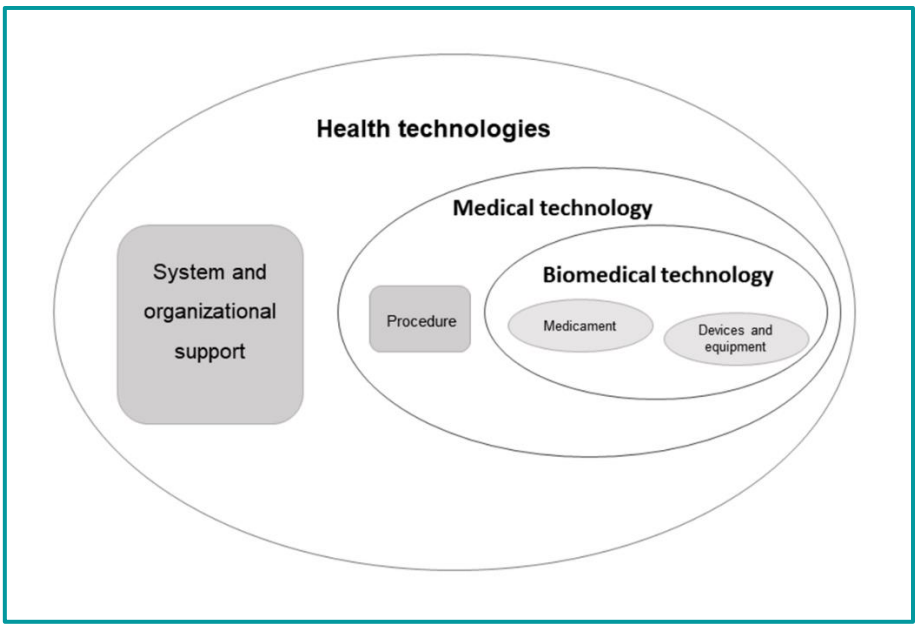


Figure 1. Set of health technologies used at each level of care
(*Adapted from Liaropoulos, 1997*)

Health technologies can also be classified according to their purpose:

- *Prevention, screening, and diagnosis*: to identify the cause and nature or extent of a disease;
- *Treatment*: to improve or maintain patient's health, avoid further deterioration or as a palliative care;
- *Rehabilitation*: to restore, maintain or improve the physical or mental functions of an individual.

Additionally, technologies are classified according to their diffusion stage:

- *Future*: design stage of development;
- *Experimental*: laboratory or animal testing stage;
- *Investigational:* clinical assessments stage;
- *Established:* considered by providers as a standard approach for use;
- *Obsolete/abandoned/outdated:* technologies that are outperformed by other technologies or were rated as ineffective or harmful.

Health technology life cycle

The technology life cycle begins with the development of a new product, process, or practice (*innovation*), and usually ends with its practical use (*adoption*). During this process, economic evaluations and clinical trials are conducted to assess the benefits and risks of the new technology to the target population. However, these evaluations are usually limited to controlled settings which may not truly reflect the impact of the technology in real world scenarios. That is why HTA agencies are often encouraging further observational studies [22, 24, 25].

Several factors can impact on the *innovation* step, such as the clinical features and prevalence of the studied disease or health-related condition, clinical research barriers and limitations (e.g. human and technical resources, sample size, ethical procedures, funding), economic aspects (e.g. technology price) and regulatory legislation of a given country/region. As soon as the new technology reaches the market, the *innovation* phase ends. At this point, other factors impact on the *diffusion* of the technology and its acceptability on the market. Regulatory legislation can slow down the *diffusion* process [24–27].

The recognition of the added value of the new technology by healthcare providers is an important step to further incorporate it on the health system (*incorporation* step). For low-cost technologies, incorporation may go unnoticed. However, for large-scale or high-cost technologies, this stage is critical. Technologies' life cycle is constantly assessed, especially considering the growing introduction of products on the market worldwide – which are usually more effective, but also more expensive. Obsolete technologies or those considered ineffective or harmful should be discontinued from the HTA process *(obsolete/abandoned/outdated* step*)* [22, 24–27].

HTA agencies

HTA is an important process to ground decisions about the incorporation of new technologies, being used in several countries such as Australia, Canada and Western Europe. The HTA can be performed by different bodies, such as government agencies, insurance companies, professional associations, private institutions, and universities [28, 29].

Some international and well-known HTA networks are: the International Network of Agencies for Health Technology Assessment (INAHTA) that currently includes 55 agencies; the European Network for Health Technology Assessment (EUnetHTA) with 51 participating organizations; the Health Technology Assessment Asia Link (HTAsiaLink) with 35 agencies; the Red de Evaluación de Tecnologías en Salud de las Américas (RedETSA) with 38 member institutions; and the Professional Society for Health Economics and Outcome Research (previous known as the International Society for Pharmacoeconomics and Outcomes Research - ISPOR) with more than 20,000 individual members distributed in 110 countries. These networks are often composed by national agencies such as: the National Coordinating Centre for Health Technology Assessment (NCCHTA – United Kingdom), the Canadian Coordinating Office for Health Technology Assessment (CCOHTA – Canada), the Agency for Healthcare Research and Quality (AHRQ – United States of America). For further information see Table 1.

Table 1. HTA agency by country

Country	Agency	Acronym
Argentina	Instituto de Efectividad Clínica y Sanitaria	IECS
Australia	Adelaide Health Technology Assessment	AHTA
	Australian Safety & Efficacy Register of New Interventional Procedures-Surgical	ASERNIP-S
	The Medical Services Advisory Committee	MSAC
	Pharmaceutical Benefits Advisor Committee	PBAC
Austria	Institute of Technology Assessment	ITA
Belgium	The Belgian Health Care Knowledge Centre	KCE
Brazil	National Committee for Health Technology Incorporation	CONITEC
Canada	Canadian Agency for Drugs and Technologies in Health	CADTH
	Institut National d'Excellence en Santé et en Services Sociaux	AETMIS
	Alberta Heritage Foundation for Medical Research	AHFMR
	Canadian Coordinating Office for Health Technology Assessment	CCOHTA
	Institute of Health Economics	IHE
	Medical Advisory Secretariat Ministry of Health and Long-Term Care	MAS
Chile	Evaluación de Tecnologías Sanitarias	ETESA
Cuba	Instituto Nacional de Higiene y Epidemiologia	INHEM
Denmark	Danish Centre for Health Technology Assessment Danish	DACEHTA
	Institute for Health Services Research	DSI
Spain	Health Technology Assessment Agency	AETS
	Andalusian Agency for Health Technology Assessment	AETSA
	Galician Agency for Health Technology Assessment	AVALIA-T
	Catalan Agency for Health Technology Assessment and Research	CAHTA
Finland	Finnish Office for Health Technology Assessment	FinOHTA
France	Agence Nationale d'Accreditation et d'Evaluation en Sante	HAS
	Comité d'Evaluation et de Diffusion des Innovations Technologiques	CEDIT
Germany	German Institute of Medical Documentation and Information	DIMDI
	Institute for Quality and Efficiency in Health Care	IQWiG
Hungary	Health Economics and Health Technology Assessment Research Centre	HunHTA
Israel	Israeli Center for Technology Assessment in Health Care	ICTAHC
Latvia	Health Statistics and Medical Technology Agency	HSMTA
Netherlands	Healthcare Insurance Board	CVZ
	Netherlands Organisation for Health Research and Developmen	ZonMW
Mexico	Mexican Social Security Institute	IMSS
N. Zealand	New Zealand Health Technology Assessment	NZHTA
Norway	Norwegian Knowledge Center for the Health Services	NOKC

Country	Agency	Acronym
Sweden	Centre for Medical Technology Assessment	CMT
	Swedish Council on Health Technology Assessment	SBU
Switzerland	Swiss Network for Health Technology Assessment	SNHTA
USA	Agency for Healthcare Research and Quality	AHRQ
	Centers for Medicare & Medicaid Services	CMS
	Technology Assessment Program	VATAP
UK	Institute of Applied Health Sciences	IAHS
	Health Technology Assessment- Coordinating Centre for Health Technology Assessment	NIHR-NCCHTA
	NICE National Institute for Health and Clinical Excellence	NICE
	NHS Quality Improvement Scotland	NHS QIS

N. Zealand: New Zealand; USA: United States of America; UK: United Kingdom
Source: adapted from the International Network Agencies for Health Technology Assessment, 2011.

History of HTA

Following the Second World War, the creation of most of the health systems, together with the scientific and technological grow in health, led governments implement strategies and policies to effectively incorporate new technologies using a limited budget [1–4]. However, while policymakers were focused only on cost savings, healthcare professionals highlighted the need to further evaluate the value of the technology (e.g. benefits and consequences) in practice, which could directly impact on costs. At that time, several interventions were considered harmful or ineffective and no standards for clinical practice existed [7, 16, 17].

The grow of the HTA field can be historically divided into three moments:

- 1978-1987: strengthening of the scientific field and evidence-based practice
- 1988-2003: enhancement of legitimacy and policy processes in HTA
- 2004-2013: international development of HTA

These milestones offer an empirical illustration of the HTA natural history according to Battista and Hodge that defined it as a linear process involving the stages of emergency-consolidation-expansion [2, 30].

Before 1970, economic evaluations of a given technology were scarce in the literature. One of the firsts discussions on the economic imbalance between health care providers, medical services and patients occurred only in 1963, authored by the Nobel Prize-winning economist Kenneth Arrow with the work 'Uncertainty and the Welfare Economics of Medical Care'. Short after, Martin Feldstein published his research on an economic overview of the United Kingdom' National Health Service (1963-1967), which was later used by several universities to ground economic evaluations in health [1, 30].

Similarly, in depth discussions on 'clinical decision making' raised during the 70s. Alvan Feinstein published a paper on entitled 'Clinical Judgment' in 1967 drawing attention on the role of clinical reasoning and identified biases that can affect it. In 1972, Archie Cochrane published 'Effectiveness and Efficiency', which describes the lack of controlled trials supporting many practices that had previously been assumed to be effective. In the mid-1980s, several clinical epidemiology guidelines, translating epidemiological methods to physician decision making were published. However, the term 'evidence-based medicine' was introduced slightly later, in the context of medical education by Gordon Guyatt in 1992. In 1996, David Sackett clarified the definition of this term as the conscientious, explicit, and judicious use of current best evidence in making decisions about care of individuals patients. In the 2000s, this concept was expanded to 'evidence-based practice' [15, 30].

During the emergency of HTA (1978-1987), the target was to improve and standardize the methodologies for the production and marketing of new

technologies that could immediately be used in health systems. This innovative use of scientific and technical knowledge guaranteed the legitimacy of the assessments in terms of objective and neutral analyses and maximized of the benefits of the technology in terms of standards for efficacy and safety [1, 17, 30]. Shortly after this period, HTA was institutionalized in some countries, such as Australia and Canada. In Europe, Sweden, United Kingdom, France, and the Netherlands were pioneers in implementing complete HTA, with formal agencies created already in the during 1980-1990. Conversely, in the United States of America, although the Office of Technology Assessment (OTA) of the American Congress was created in the late 80s, it lasted until 1995. After several political and organizational changes in the country, today only the veterans' program of War and some private sector entities use HTA [17,30].

During the 1980s, HTA became an autonomous field of medical science, which can be partially due the development of the 'technology assessment' movement, initiated by OTA from United States [31]. This era is marked by the foundation of the International Society of Technology Assessment in Health Care that was responsible for the creation of the International Journal for Technology Assessment in Health Care (IJTAHC) in 1985. After that, a significant increase of scientific publications on this field occurred, with around 10 articles published per year. Additionally, the International Society of Technology Assessment in Health Care created on of the most important global scientific event, the HTAi Annual Meeting, that occurs until today [2, 31].

The Swedish Council on Technology Assessment in Health Care (SBU) was officially the first national HTA agency created in Europe in 1987. The purpose of this agency was only to inform the Swedish government and district councils about the cost of health technologies, yet without a regulatory function.

The agency should provide evidence-based information on health technologies to guide health policies and clinical practices in the country [4, 32].

Similarly, in the 1990s, the National Institute for Health and Clinical Excellence (NICE) was created in the United Kingdom aiming at providing reports to ground healthcare decisions about the benefits of an intervention, especially considering the results of cost-effectiveness analyses [4, 17]. NICE is internationally recognized for its methodological rigor and evaluation guidelines that are used as 'models' by several other agencies. The NICE' HTA process results in 'recommendations' that are considered mandatory in the country. Technologies positively recommended for use should be made available within three months of the publication of the final report [30, 31].

In Australia, the government created the Pharmaceutical Benefits Advisory Committee (PBAC) that officially evaluates the clinical and economic value of new health products (e.g clinical efficacy, safety, and cost-effectiveness). Additionally, drugs that are recommended for use by the PBAC should be introduced in the national list of the Australian Pharmaceutical Benefits Scheme (PBS). PBS is a government program that is responsible for reimbursement of most of the prescription drugs in country [1].

In Canada, in the late 80s, the Canadian Coordinating Office for Health Technology Assessment - CCOHTA was created. Later, in 2006, the government created the Canadian Agency for Drugs and Technologies in Health (CADTH). The agency should provide evidence-based assessments of new technologies at all government levels. Conversely to the NICE' recommendations, the CADTH's recommendations are of advisory nature. That is to say, the Ministries of Health and provincial governments are responsible to decide whether to introduce the new technology in the health system and public drug plans or not [33, 34].

The International Network of Agencies for Health Technology Assessment (INAHTA) was the first network created (1993) aiming at disseminating and standardizing concepts, methods, and quality standards in HTA [1]. The Agency's mission is to become a permanent forum for exchange and collaboration between the different HTA bodies worldwide [35]. During 2000s, other HTA networks were created, reinforcing the role of complete economic and clinical assessment of technologies. Additionally, health economists developed systematic methods for associating technologies' effectiveness and efficiency. This allowed the costs to be included in the decision-making process and promoted more efficient allocation of resources [15, 21, 30].

Unfortunately, HTA agencies worldwide do not have sufficient resources (both human, technical, and economic) to evaluate all the new interventions that enter the market and whether they should be incorporated in the public health systems. The prioritization of the technologies usually takes into account the impact of the disease in the country (e.g. affected population, other available technologies, current costs). For this reason, the influence of HTA on policy formulation may be limited.

Main steps of the HTA process

Standardized methods and processes for technologies evaluation are paramount to ensure a consistent and transparent HTA report [36, 37]. Although these methods may vary according to each country, some main steps are commonplace (e.g. prioritize and identify to be evaluated; address the problem; determine the assessment scenario; retrieve the available evidence; obtain new primary data; interpret and synthetize the evidence; formulate recommendations and disseminate the information; monitor the impact of the HTA report [12].

Step 1: Identify the topic

HTA agencies have criteria for selecting health-related topics to be assessed by HTA reports; however, these criteria are not always explicit. Some examples of criteria for prioritizing areas include: disease mortality or morbidity burdens; number of affected individuals; aggregate costs of the technology; current aggregate costs of the disease or health-related problem; lack of standards procedures in clinical practice; technologies' potential to improve patient health outcomes; technologies' potential to reduce health risks; available scientific evidence to support an official assessment; great interest among health professionals; public or political requirements including regulatory or reimbursement decisions [4, 12, 37].

Several methods and techniques are available for establishing systematic topics' priority in a given HTA agency. Commercial software are often used. Some of the steps in the process include: selection of the criteria to be used for setting the priorities; assignment of weights to the criteria; identification of candidate topics for evaluation; calculation of the degree of priority for each topic; ordering of the topics according to the priority level; and review of the selected topics to ensure that their assessment is consistent with the purpose of the HTA body [38–41].

Step 2: Specify the problem

Prior the formal evaluation of a selected topic, a clear definition of the health problem, the clinical question or the therapeutic indication of the new technology should be provided. This includes: a description of the disease or health-related problem; definition of the target population; description of the

new technology and available comparators; level of care; and benefits and risks (e.g. clinical, humanistic, economic outcomes) that should be assessed [40]. Reports often formulate the question grounded on acronyms, such as PICO - where P corresponds to the patient or population, I is the intervention or indicator, C refers to the comparator or control, and O are the outcomes [12, 42].

Step 3: Gather the evidence

After the definition of the topic to be evaluated, researchers need to find, select, appraise, and synthetize the information. For new technologies, this information can be scarce or difficult to find; for many technologies, the evidence can be conflicting, dispersed and of low methodological quality (prone to bias). Thus, the time and resources required for these activities should be carefully measured when planning the HTA report, especially because they vary according to each topic [12, 40, 42].

One of the very firsts steps for evidence gathering refers to the search of information. The structured question of the HTA report will guide the identification of keywords or descriptors to be used in different information sources. The combination of these descriptors with Boolean operators such as AND, OR or NOT provides more accurate searches to be employed, for instance, in electronic databases as demonstrated in Figure 2 [12, 43]. Although information may be overlapping among sources, multiple sources should be sought to increase the likelihood of retrieving relevant information.

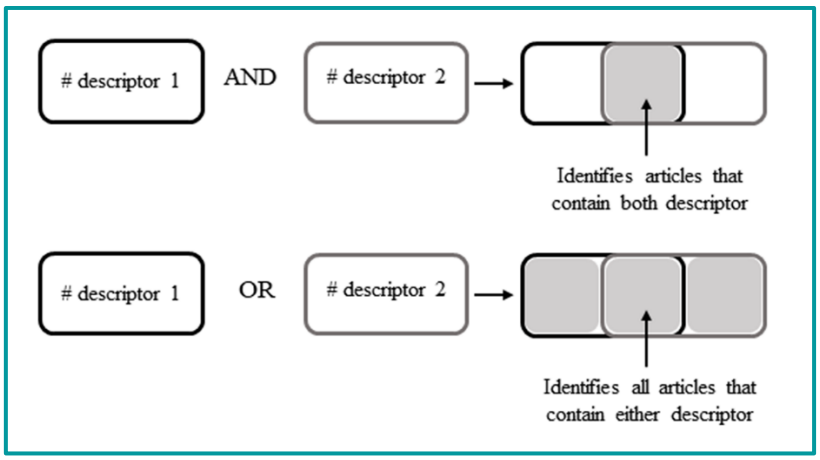


Figure 2. Strategies for using Boolean operators 'AND' and 'OR'

Several information sources, both virtual or physical, can be used in a HTA report including computerized databases of published literature; computerized databases of clinical and administrative data; government reports; systematic reviews and meta-analyses; annals of scientific events and reports; clinical guidelines. These sources are usually classified as primary or secondary according to the format of publication, access and needs of end-users (Table 2).

Table 2. Examples of primary and secondary information sources

Primary sources	Secondary sources
MEDKINE	ACP Journal Club
EMBASE	Evidence Based Medicine
CINAHL	InfoPOEMS
PsycInfo	Evidence Based Medicine Health
CANCERLIT	Evidence Based Nurse
PDQ	Jornal of Evidence Based Health Care
HealthStar	Cochrane Library
LIFE	British Medical Journal
BEHA	Lancet
DISS	Archives of Internal Medicine
WHOLIS	Canadian Medical Association Journal
LILACS	Evidence Based Obstetrics e Gynecology

Source: adapted from Marques, 2004

Step 4: Collect primary data

As described in Chapter 01, systematic reviews, meta-analyses, and other secondary studies are usually used to ground decisions in healthcare as they provide filtered evidence with higher levels. However, sometimes, additional evidence from primary studies (e.g. trials or observational studies) may be required to allow more assertive conclusions on the effects of a given technology, especially when secondary studies are lacking or outdated. This process involves further searches and primary data collection, which may be a limiting factor when developing an HTA dossier as it depends on human resources capacity and time [4, 36, 37].

Step 5: Appraise the evidence

In a next step, all the gathered evidence should be critically evaluated regarding methodological and clinical validity and usefulness. The baseline data and findings from the scientific evidence from different studies' design should be properly collected in standardized forms to allow replicability and transparency. Collecting data from unreliable evidence may led to misleading conclusions and waste of limited resources [36, 37].

Systematic approaches to critically assess the methodological quality and risk of bias of the available studies should be used, which requires, among others, strong epidemiology and statistics knowledges [44-46]. See further information on tools for assessing studies' quality on Chapter 04.

One of the most common approaches to summary studies' quality is to set up an 'evidence table' including: the attributes of the study design (e.g. randomization and blinding process); patient characteristics (e.g. number of

cases, age, gender); results (e.g. mortality, morbidity, adverse events, patients' quality of life); and derived statistical results (e.g. p-value, confidence interval). This information also allows reviewers to systematically compare the characteristics and results of the gathered studies [44-46].

Step 6: Synthetize and consolidate the evidence

The systematic tabulation of the data favors the synthesis and consolidation of the evidence, which can be performed by both qualitative or quantitative analyses [47-48].

The systematic review is a qualitative technique that can increase the accuracy of individual study results by improving the effect estimates of a given intervention. Systematic reviews are commonly followed by a statistical component (meta-analysis), which quantitatively integrates the results of two or more primary studies, increasing the statistical power of the findings [44-46]. Learn more about this topic on Chapter 03.

HTA dossiers may be developed including one or more 'dimensions'. The one-dimension document is grounded only on the systematic synthesis of the evidence. Two-dimension analysis can additionally include an economic evaluation of the data and multiple dimensions dossiers also have expert opinions and mathematical modeling [44-46].

Step 7: Economic evaluation and health systems impact analyses

The studies of costs and consequences related to the use of a technology are one of the main methods of analysis used in HTA [49-51]. Several types and methods of economic evaluation are available and should be selected according

to the purpose of the evaluation and the availability of data and other resources. These studies are usually classified according to how they measure the costs and effects of an intervention (e.g cost-minimization analysis, cost-effectiveness analysis, cost-utility analysis, and budget impact analysis). Both primary (clinical-epidemiological studies) and secondary (systematic reviews and mathematical modeling) level data may be used when performing an economic evaluation [49-51]. See further information on Chapters 07 and 08.

Step 8: Assessment of social, ethical, and legal aspects

Ethical and social values should always be considered by HTA agencies. Yet, only a minority of HTA reports address these issues. This may occur given the scarcity of validated methods for integrating ethics with HTA; professionals' perception on the apparent irrelevance of HTA for policies' formulation; low priority attributed to social and ethical aspects [52-53].

The power of 'choice' suggests that decisions regarding the use and diffusion of health technologies must be guided by the principles of equitable access to technologies and the offer of choice to individuals, facilitated by relationships of trust between patients and providers. Community participation and case reports offer an informed and participatory approach to extracting ethical and social values in HTA. This idea should be further disseminated among HTA agencies, payers, and the society to allow a more complete evaluation of the benefits – or otherwise – of a technology [52].

Step 9: Formulation of findings and recommendations

The conclusions of an assessment and recommendations are the suggestions or advice that accompany the findings. Recommendations can be displayed in several ways, such as a set of options, clinical guidelines, regulations, or hierarchy of therapy/procedures to be used in a given setting. The processes of interpreting and synthesizing the evidence help the evaluating group to determine the strength of the evidence to answer aspects of the proposed question or technology [12].

The Canadian Task Force on Preventive Health Care adopts the following grading scale when recommending a technology: 'conclusive evidence on benefits is recommended', 'highly recommended technology'; 'reasonable evidence to recommend the technology with reservations'; 'the existing evidence is conflicting and does not allow a recommendation to be made against or in favor of the technology'; 'there is reasonable evidence not to recommend the technology'; 'there is good evidence for not recommending the technology'; 'the evidence is insufficient (in quantity or quality) to establish a recommendation' [34].

These degrees can be established according to the level of evidence, the quality of the evaluated studies or the association of the evidence with the outcomes. Agencies can conclude that the existing evidence is insufficient to provide the information needed for policy formulation, and that further studies are needed to generate data for some aspects of the evaluation [12].

Step 10: Dissemination of findings and recommendations

The approaches to disseminating the findings of HTA studies can be grouped into three dimensions: target groups, means of dissemination, and implementation strategies.

The target groups can be patients/consumers, organizations, professional associations; assistance provider organizations; government, health plan operators; quality assurance and accreditation organizations; government policy makers; medical researchers; health sector companies; popular/scientific journalists and educational institutions [12, 40].

The same findings should be presented in different formats and styles, depending on the audience to which they are directed and the means available to disseminate the information (for instance, detailed reports should be delivered to researchers and policy makers, while quick reference guides can be presented for clinicians) [12, 40].

Step 11: Monitoring the impact of assessment reports

Assessing and monitoring the impact of a HTA can be considered a difficult task as different factors can affect this evaluation such as the dissemination technique and the report quality. Concerning the last, important progress in the methodological area of HTA was observed in the past years, which contributes to the publication of complete dossiers with enhanced quality and transparency [28].

Systematic attempts to monitor the impact of the recommendations are still infrequent, despite the growing interest that government agencies, health insurers and international agency consortia have attributed to the subject [48].

Conclusions

The growing development of new health technologies and their implementation in the market highlight the need of standard processes for the evaluation of the impacts of these products in different players, including patients, healthcare professionals, pharmaceutical companies, and the society. This chapter introduced the role of HTA that seeks to show how best to allocate the health budget and to inform health policy makers by using the best scientific evidence on the medical, social, economic, and ethical implications of investments in health care. HTA helps identifying best practices in health care, thereby enhancing safety, improving quality, and saving costs. HTA agencies use standardized methods to evaluate health technology aiming at ensuring consistent evaluations that can guide these decisions. The following steps should be perform to incorporate a technology into a giving setting: identify the topic for assessment, clearly specify the problem, gather the available evidence, collect primary data, assess the evidence, synthesize and consolidate the evidence, conduct an economic assessment, assess social factors, ethical and legal, formulate conclusions and recommendations, disseminate the conclusions and recommendations and monitor the impact of the evaluation reports.

References

[1] O'Donnell JC, Pham S V, Pashos CL, Miller DW, Smith MD. Health technology assessment: Lessons learned from around the world - An overview. Value Health. 2009

[2] Banta H, Kristensen F, Jonsson E. A history of health technology assessment at the European level. Int J Technol Assess Health Care. 2009

[3] Liaropoulos L. Do we need "care" in technology assessment in health care? International journal of technology assessment in health care. 1997

[4] Drummond MF, Schwartz JS, Jönsson B, Luce BR, Neumann PJ, Siebert U, et al. Key principles for the improved conduct of health technology assessments for resource allocation decisions. International Journal of Technology Assessment in Health Care. 2008.

[5] Kelty MF. Health technology assessment and impact. Natl Forum. 1978

[6] Stevens A, Milne R, Burls A. Health technology assessment: History and demand. Journal of Public Health Medicine. 2003.

[7] Manias E, Williams A. Medication adherence in people of culturally and linguistically diverse backgrounds: A meta-analysis. Ann Pharmacother. 2010

[8] Di Bello V, Tonti G, Barletta G, Di Salvo G, Mancino M, La Carrubba S, et al. Introduction to Health Technology Assessment. Journal of Cardiovascular Echography. 2012.

[9] Shekelle PG, Morton SC, Keeler EB. Costs and benefits of health information technology. Evidence report/technology assessment. 2006.

[10] Kristensen FB, Mäkelä M, Neikter SA, Rehnqvist N, Håheim LL, Mørland B, et al. European network for health technology assessment, eunethta: Planning, development, and implementation of a sustainable european network for health technology assessment. Int J Technol Assess Health Care. 2009

[11] Goodman CS, Ahn R. Methodological approaches of health technology assessment. Int J Med Inform. 1999

[12] Cutler DM, McClellan M. Is technological change in medicine worth it? Health Aff. 2001

[13] Newhouse JP. Medical care costs: how much welfare loss? The journal of economic perspectives : a journal of the American Economic Association. 1992.

[14] McDaid D, Cookson R, Maynard A, Sassi F. Evaluating health interventions in the 21st century: Old and new challenges. Health Policy. 2003.

[15] Banta D. The development of health technology assessment. Health Policy (New York). 2003

[16] Banta D, Jonsson E, Childs P. History of the international societies in health technology assessment: International society for technology assessment in health care and health technology assessment international. International Journal of Technology Assessment in Health Care. 2009.

[17] Kassavou A, Sutton S. Automated telecommunication interventions to promote adherence to cardio-metabolic medications: meta-analysis of effectiveness and meta-regression of behaviour change techniques. Health Psychol Rev. 2018

[18] Langebeek N, Nieuwkerk P. Electronic Medication Monitoring-Informed Counseling to Improve Adherence to Combination Anti-Retroviral Therapy and Virologic Treatment Outcomes: A Meta-Analysis. Front Public Heal. 2015

[19] Mahtani KR, Heneghan CJ, Glasziou PP, Perera R. Reminder packaging for improving adherence to self-administered long-term medications. Cochrane Database Syst Rev. 2011
[20] Battista RN, Hodge MJ. The evolving paradigm of health technology assessment: Reflections for the millennium. CMAJ. 1999.
[21] Polisena J, Williams D, Ciani O. Health technology assessment of medical devices. In: Clinical Engineering Handbook, Second Edition. 2019.
[22] Kauppi K, Välimäki M, Hätönen HM, Kuosmanen LM, Warwick-Smith K, Adams CE. Information and communication technology based prompting for treatment compliance for people with serious mental illness. Cochrane Database of Systematic Reviews. 2014.
[23] Jones O a, Lester JN, Voulvoulis N, Kasprzyk-Hordern B, Dinsdale RM, Guwy AJ, et al. Fate and removal of typical pharmaceutical and personal care products in a wastewater treatment plant from Beijing: a mass balance study. Sci Total Environ [Internet]. 2017 Apr 17 [cited 2014 Jun 2];23(4):3–27. Available from: https://doi.org/10.1016/j.jhazmat.2018.04.043
[24] Gutiérrez-Ibarluzea I, Chiumente M, Dauben HP. The life cycle of health technologies. Challenges and ways forward. Frontiers in Pharmacology. 2017.
[25] Mytton OT, Velazquez A, Banken R, Mathew JL, Ikonen TS, Taylor K, et al. Introducing new technology safely. Qual Saf Health Care. 2010
[26] Oliver A, Mossialos E, Robinson R. Health technology assessment and its influence on health-care priority setting. International Journal of Technology Assessment in Health Care. 2004.
[28] Hailey D. Development of the international network of agencies for health technology assessment. International Journal of Technology Assessment in Health Care. 2009.
[29] INAHTA. International Network of Agencies for Health Technology Assessment. http://www.inahta.org/HTA. 2011.
[30] Wagstaff A, Culyer AJ. Four decades of health economics through a bibliometric lens. J Health Econ. 2012;
[31] Benoit C, Gorry P. Health technology assessment: The scientific career of a policy concept. Int J Technol Assess Health Care. 2017
[32] Jonsson E, Reiser SJ. International Journal of Technology Assessment in Health Care: Editorial. International Journal of Technology Assessment in Health Care. 2008.
[33] Menon D, Stafinski T. Health technology assessment in Canada: 20 years strong? Value Heal. 2009
[34] Battista RN, Côté B, Hodge MJ, Husereau D. Health technology assessment in Canada. International Journal of Technology Assessment in Health Care. 2009.
[35] Hailey D. The history of health technology assessment in Australia. International Journal of Technology Assessment in Health Care. 2009.
[36] Assasi N, Tarride JE, O'Reilly D, Schwartz L. Steps toward improving ethical evaluation in health technology assessment: A proposed framework. BMC Med Ethics. 2016;
[37] Booth A. How much searching is enough? Comprehensive versus optimal retrieval for technology assessments. International Journal of Technology Assessment in Health Care. 2010.
[38] Noorani HZ, Husereau DR, Boudreau R, Skidmore B. Priority setting for health technology assessments: A systematic review of current practical approaches. Int J Technol Assess Health Care. 2007

[39] Husereau D, Boucher M, Noorani H. Priority setting for health technology assessment at CADTH. Int J Technol Assess Health Care. 2010
[40] Culyer AJ, Bombard Y. An equity framework for health technology assessments. Med Decis Mak. 2012
[41] De Voogt P, Janex-Habibi ML, Sacher F, Puijker L, Mons M. Development of a common priority list of pharmaceuticals relevant for the water cycle. Water Sci Technol. 2009;59(1):39–46.
[42] Cornet R, De Keizer NF, Abu-Hanna A. A framework for characterizing terminological systems. Methods Inf Med. 2006
[43] Popay J, Roberts H, Sowden A, Petticrew M, Arai L, Rodgers M, et al. Narrative Synthesis in Systematic Reviews: A Product from the ESRC Methods Programme. ESRC Methods Program. 2006
[44] Ryan R. Cochrane Consumers and Communication Review Group: data synthesis and analysis. Cochrane Consum Commun Rev Gr. 2013
[45] Petticrew M, Roberts H. Systematic Reviews in the Social Sciences: A Practical Guide. Systematic Reviews in the Social Sciences: A Practical Guide. 2008.
[46] Dixon-Woods M, Agarwal S, Jones D, Young B, Sutton A. Synthesising qualitative and quantitative evidence: A review of possible methods. Journal of Health Services Research and Policy. 2005.
[47] Murphy E, Dingwall R, Greatbatch D, Parker S, Watson P. Qualitative research methods in health technology assessment: A review of the literature. Health Technology Assessment. 1998.

[48] Kamae I. Value-Based Approaches to Healthcare Systems and Pharmacoeconomics Requirements in Asia. Pharmacoeconomics. 2010
[49] Trippoli S. Incremental cost-effectiveness ratio and net monetary benefit: Current use in pharmacoeconomics and future perspectives. European Journal of Internal Medicine. 2017.
[50] Vanni T, Luz PM, Ribeiro RA, Novaes HMD, Polanczyk CA. Avaliação econômica em saúde: aplicações em doenças infecciosas. Cad Saude Publica. 2009
[51] Hofmann B. Toward a procedure for integrating moral issues in health technology assessment. International Journal of Technology Assessment in Health Care. 2005.
[52] Kiran AH, Oudshoorn N, Verbeek PP. Beyond checklists: toward an ethical-constructive technology assessment. J Responsible Innov. 2015

HISTORY AND STEPS OF HEALTH TECHNOLOGY ASSESSMENT

BEATRIZ BÖGER
MARIANA M. FACHI

METHODS AND TOOLS FOR EVIDENCE GATHERING

ALINE A. GARABELI

Abstract

Review studies are fundamental for the provision of an in-depth knowledge on topics in health sciences. When conducted rigorously, they are able to provide an accurate synthesis of data from primary studies in order to assist researchers and professionals in the interpretation of clinical results. Reviews can be especially useful for clinical practice, helping decision-makers in choosing the most effective health technologies through the best scientific evidence available. However, for a review to be relevant and provide reliable data, researchers should first define what type of review is most appropriate for their research question and comprehensively understand the processes that is required for each type of review. Thus, this chapter aims to describe three types of literature review commonly found in the health sciences literature: narrative, systematic and scoping reviews. The definitions and applications of each type will be presented, as well as a brief description of the main steps as currently established by international organizations.

Keywords: review studies; synthesis of data; evidence-based practice

Introduction

As established by evidence-based practice, decision-making processes in a health technology – such as for a new treatment, a diagnostic method, or any other health intervention – should be based on the best available scientific information [1]. The choice of a health technology for treating a condition first requires a deep analysis of numerous scientific studies and an appropriate synthesis of their results as discussed in Chapter 01. This process, however, can take some time, because it implies having access to primary studies, to critically evaluate the evidence and to synthetize the results of multiple studies that may have contradictory results [2–4].

Reviews are retrospective studies that integrate and synthetize the results of primary studies, making reading and analysis of a large number of health information on a given topic easier [3-6]. There are several types of reviews in literature that may differ according to the methodological process they follow [7]. However, these differences can have implications in the reliability of the study and hinder the quality of the evidence [1]. For this reason, reviews of any nature should be conducted with great scientific rigor and to clearly communicate the findings to facilitate data interpretation and to help health professionals to make more assertive decisions [4, 6].

Understanding the definitions of and the main steps to conduct review studies is essential when choosing the most suitable method for research. This chapter discusses three types of reviews commonly found in the health literature: narrative, systematic and scoping reviews. The definitions, characteristics, applications, and main steps for conducting each of these reviews is presented in the following pages.

Narrative review

When a researcher wants to understand the broad context of a topic at the beginning of the research process, he or she might want to conduct a narrative review. Narrative reviews evaluate the current literature on a theme through a wide survey of the existing information. Generally, these reviews cover several issues or contexts, gathering the information in non-systematical way [3, 7]. A narrative review about arterial hypertension, for example, may describe both the epidemiology of the disease, the pathophysiological mechanisms of increased blood pressure, the diagnostic criteria and treatment options [3]. This type of review is useful when a researcher aims to understand what has already been published on a topic, thus avoiding research duplication, and guiding on the identification of scientific gaps [7-9].

However, narrative reviews do not follow a pre-established protocol nor a systematized methodological process [7, 8]. Searches require no well-defined research questions and several sources of evidence (e.g. from thesis to books or full-text articles) can be used and do not need to be disclosure on the publication. The synthesis of data is usually qualitative. There is no need to systematically integrate the information nor produce maps of evidence and recommendations. Nonetheless, in the past years, there is an increasingly plea to perform more standardized narrative reviews, adding for instance the search strategies and eligibility criteria in order to reduce bias and enhance science transparency [9].

Despite their usefulness in science, narrative reviews are not reliable as sources of evidence for clinical decision-making as they cannot be reproduced by other researchers [3, 10]. In addition, narrative reviews are open to several biases as the authors may select the studies grounded on their preferences or conclusions may be subjected to the researcher's opinion [5, 7, 10].

Main steps for conducting a narrative review

Conducting a narrative review does not require a high degree of scientific rigor. However, to gather relevant information, it is necessary to understand the process and its mains steps as demonstrated in Figure 1:

Figure 1. Summary of the main steps for conducting a narrative review

Defining keywords

For narrative reviews, it is not necessary to define a specific and objective question to be answer by the research, but rather to determine the theme that will be studied. Therefore, the keywords that are best related to the theme should be chosen. The terms should be open enough to include all of the relevant sources of evidence, but also specific enough to retrieve those restricted to the research objectives [5]. In this scenario, 'keywords' and their synonyms are paramount of guide the search. For example, for a review on the treatment for diabetes mellitus, the keywords might be related to 'diabetes mellitus' and 'pharmacotherapy' and their synonyms. As a suggestion, authors can previously read relevant articles related to the theme to find the keywords used. Additionally, authors may easily find related terms or 'descriptors' in controlled vocabularies of electronic databases such as the MeSH (Medical Subjetc Headings) terms in PubMed or the Emtree (Embase Subject Headings) from Embase.

Performing a search

After defining the keywords, the researcher can conduct the search. It is recommended to search in at least two or three reliable and updated sources [9]. Currently, complete online databases including articles on health sciences and that are widely used are MEDLINE (usually searched through PubMed), Scopus, Web of Science and Embase. For searching systematic reviews of randomized clinical trials on efficacy and safety of treatments, there is also the Cochrane Database of Systematic Reviews. In addition, if the researcher wants to search for articles published in non-indexed journals, the Google Scholar tool or any other search engines that allow the searching of grey literature (i.e. reports, conference proceedings, clinical trials registered in clinicaltrials.gov, etc.) can be used [8, 9].

It is important to highlight that each database has its own search engine, so it is necessary to know each tool to use them correctly. As mentioned, some databases present a controlled vocabulary with 'descriptors' of the main issues of the studies. The association of the keywords might be different in each search engine, and their adequate combination is essential for an accurate search. As the search strategies are not restricted and specific for a question, a large number of articles will probably be recovered through the database searches. The researcher might have to evaluate each study manually or import them into a reference manager (e.g. EndNote or Mendeley) to facilitate this analysis [9].

Studies' selection

After the search is complete and duplicates are removed, it is time to read the studies and select those of interest for the research. In narrative reviews, a rigorous process for the inclusion and exclusion of articles is not required, nor is necessary that the process is carried out by two researchers independently [5, 8].

Synthesis

The synthesis of the results might be done in several ways, but authors are not required to appraise all the information they gathered. However, it is recommended that results are presented as clear and objective as possible [5, 10]. The results might be described in the text (e.g. subdivided into sub-themes or topics), or in simple graphs and tables. The studies used for this synthesis must be referenced in the publication [5, 8, 9].

Systematic review

The result of a single study is usually not enough to support clinical decisions. On the other hand, a synthesis of a set of results from several similar individual studies can produce a stronger evidence to ground practical decisions. However, keeping up with information in health care has never been easy: an overload of unfiltered information and lack of open access to information relevant to the well-being of patients currently exist. Even with the assistance of electronic databases such as NLM's MEDLINE, the problem of having to trawl through and sift vast amounts of data has grown. Just between 1865 and 2006, the index in MEDLINE grew from 1,600 references to nearly 10 million. As mountains of synthesized research evidence accumulate, researchers need to keep improving the methods for gathering, filtering, and synthesizing it [2, 3].

Systematic literature reviews and meta-analyses endeavoring to make sense of multiple trials began to appear in a variety of health fields in the 1970s and 1980s. By the mid-1980s, the need to minimize the likelihood of being misled by the effects of biases and the play of chance in reviews of research evidence was being made evident in articles and textbooks. In 1988, regularly updated

electronic publication of systematic reviews and meta-analyses, along with bibliographies of randomized trials, began in the perinatal field. This provided a model for the inauguration of the international Cochrane Collaboration in 1993 to prepare, maintain, and disseminate systematic reviews of the effects of health care interventions [2, 11].

Systematic reviews are defined as retrospective scientific investigations that aim to synthesize the results of primary studies and to produce evidence on the effect of technology [12, 13]. This type of review follows a systematic and explicit method to obtain a critical analysis of the studies, to synthesize the data and to answer clearly and objectively a specific question. This type of review, when well-performed, is reliable, reproducible, and able to be conducted and confirmed by other researchers [3].

Conducting a systematic review includes several steps. A high degree of scientific rigor is mandatory to reduce bias and errors [15]. Systematic reviews are currently considered the highest level of scientific evidence for the clinical decision-making process, as described in Chapter 01. They provide reliable and up-to-date information on a given topic. Through systematic reviews, it is possible to solve controversial issues on the effects of technologies and to decide on the implementation of those with better results [12, 15].

An overview of the mains steps for conducting a SR will be briefly presented in this section. For further detailed processes please consult the online recommendations provided by international organizations that address the methods for performing and reporting high-quality systematic reviews and disseminate information on scientific evidence:

- The Cochrane Collaboration:
  https://training.cochrane.org/handbook/current
- The Joanna Briggs Institute:
  https://wiki.jbi.global/display/MANUAL/JBI+Manual+for+Evidence+Synthesis

Main steps for conducting a systematic review

Figure 2 depicts the main steps that author should follow when performing a systematic review:

Figure 2. Summary of the main steps for conducting a systematic review

Research question

The research question is a key element for obtaining relevant answers in science. A good research question is essential to conducting a systematic review, because it will guide all of the other phases, such as the definition of the eligibility criteria, data extraction and results presentation. The research team needs to evaluate whether the formulated question is answerable, whether it corresponds to reality, what its relevance is and which benefits it will bring to patients and professionals. The question should be clear, specific, and objective, and it should aim to fill the gaps of the scientific knowledge on a specific topic [15-17]. Usually, a research question on healthcare topics should include the population that will

be considered, the technology or intervention and the outcomes of interest. An example of a research question might be:

How effective is celecoxib compared to ibuprofen in relieving pelvic pain in women with deep ovarian endometriosis?

Eligibility criteria

Besides the research question, it is necessary to settle the inclusion and exclusion criteria for the studies. Guidelines [14, 17] and researchers [4, 16, 18] recommend using the PICO mnemonic (Population, Intervention, Comparator and Outcomes of interest) to define the scope of the review and which studies should be eligible for inclusion in the analysis. The following chart presents a hypothetical example of the definition of each PICO item for the research question previously mentioned:

RESEARCH QUESTION
How effective is celecoxib compared to ibuprofen in relieving pelvic pain in women with deep ovarian endometriosis?

PICO
Population: women with deep ovarian endometriosis, with pelvic pain
Intervention: celecoxib
Comparator: ibuprofen
Outcomes: pelvic pain relief, other

In addition, authors can add an item describing the study design they will include (e.g. randomized or non-randomized trials, observational studies). This enhances the specificity of the systematic review and allows further inferences about the results from similar primary studies.

Exclusion criteria are all of the non-desirable characteristics of studies for the systematic review. This should also be defined by the authors prior the search. Exclusion criteria might include publication period, presence of non-desirable comorbidities associated to the clinical condition or any other element that goes beyond the inclusion criteria. Cochrane discourages language limitations as an exclusion criterion, but if there are no resources to include all languages, it might be considered by the authors if properly justified [18, 19].

Search strategy

One of the main characteristics of systematic reviews is an exhaustive literature search. A high sensitivity search strategy can help finding a greater number of relevant studies on a given topic [18, 20].

Some aspects that should be taken into account in the development of this strategy are the type of intervention to be studied, the time or local of the study (e.g. when a search is restricted to a single population), study design and whether searches for non-published data, such as the grey literature, will be conducted. The search strategy is commonly formulated through keywords, which represent the interesting studies. The selection of the keywords should be done using the keywords and synonyms from the literature and, whenever possible, the available descriptors from controlled vocabulary of the databases such as the MeSH terms from PubMed and the ENTREE from Embase. The combination of descriptors is done with the Boolean operators AND, OR. The use of NOT should be avoided in systematic searches because it is a restrictive operator and might exclude possibly relevant articles in the search [18, 20].

Protocol

After the definition of the question, eligibility criteria and search strategy, it is paramount to write a protocol describing the planning of all of the systematic review phases [15, 21]. Registering and publishing the protocol are practices that reduce the possibility of bias, increase the transparency of the process, and help to avoid research duplications [15]. An online portal often used to register of these studies protocols is International Prospective Register of Systematic Review - PROSPERO (**https://www.crd.york.ac.uk/prospero/**), funded by the National Institute of Health Research (NIHR) from the United Kingdom.

The protocol should include a detailed description of all methods and processes of the review such as: a rationale with the context and reasons for the research, aims and methods – that is, the eligibility criteria, databases that will be used for search, research question, search strategies for each search engine, planning for extraction and analysis of data, how the risk of bias and quality of the studies will be measured and methods for data synthesis [15-18].

Additionally, a checklist that includes all of the required items for the elaboration of an systematic review protocol is the Preferred Reporting Items for Systematic Reviews and Meta-analysis - Protocols (PRISMA-P) (**http://prisma-statement.org/Extensions/Protocols**) which can be used to guide the writing of the protocol, from the title to the planning of the data synthesis [21].

Databases

The search should be done in more than one reliable and updated database. The most common databases in health sciences are MEDLINE, Embase, Web of Science, CENTRAL [14, 18, 20]. Other specific databases could be accessed according to the aims and area of the review. In addition, manual search (as in

the reference list of the articles, Google and Google Scholar or other non-indexed sources) and exploration of the grey literature (non-published data, as reports or conference proceedings) should be performed. It is important to highlight that each search engine presents its own mechanisms, and it is fundamental to the reviewers to know those mechanisms to conduct an appropriate search. The number of studies found in each database should be registered. As the cover of the databases can be sometimes similar for a given topic, duplicates registers can be found and should be removed. This can be performed using reference manager software manager (e.g. EndNote or Mendeley). The number of removed duplicates should also be noted during the systematic review.

Screening of titles and abstracts

The screening of the titles and abstracts is the following phase after the search and is intended to remove the clearly irrelevant studies. This phase should be conducted in duplicate, by two researchers independently, and consists of reading and selecting the titles and abstracts that fit the review criteria. After the independent selection of the studies, a consensus meeting between the reviewers should be held to compare the selected studies to obtain a minimum agreement on the articles that should move on to the next phase. When no agreement is reached, a third author can participate in the discussion. The number of the excluded articles should be registered by the authors. The screening should be as inclusive as possible, so doubtful studies should always be evaluated in full in the next step. It is also recommended to calculate the Kappa coefficient to evaluate the concordance level between the reviewers, which should be at least of 0.7 in a scale of 0 (no consensus) to 1.0 (complete consensus) [20, 22].

Full-text reading

After selecting studies according to their title and abstracts, it is time to find the full texts of the included registers for download and full-text reading. The articles should be evaluated based on the concordance to the eligibility criteria. Studies will be excluded if they do not fit one or more criteria. If a study is not available or has denied access, the reviewers should try to get it through library services or by contacting the primary authors directly. As in the screening phase, two reviewers, independently, should conduct the full-text reading. The agreement level between reviewers should be calculate (e.g. Kappa coefficient). A consensus meeting to solve the discrepancies of inclusion or exclusion of studies should be held. If no consensus is reached for some studies, a third reviewer could be called in to solve the divergences. The number of excluded studies should be registered, as well as the reasons for each exclusion.

A flowchart of the systematic review process should be built containing the information about the number of articles found in the databases, the number of excluded and included registers during the screening and eligibility phases. The most common model for flowcharts used and endorsed by scientific journals is the one available on the PRISMA checklist. Please visit the Equator Network website for further information on this and other reporting guidelines: **https://www.equator-network.org/reporting-guidelines/prisma/**

Data extraction

Relevant data from the included studies should be collected in a standardized form. Excel tables or Google forms might be used. Data extraction should also be conducted by two reviewers independently, both of them trained to collect the information. Consensus meetings to reach the final results should

be conducted with the participation of a third reviewer to solve discrepancies when necessary. The information collected from the studies will depend on the aims of the systematic review, but generally, the data refer to the study characteristics (e.g. primary study design, number of participants, period of follow-up, outcome results, funding). The data should be extracted as objectively as possible and presented in tables [14, 18, 24]. All the raw data should be made available in the final publication (manuscript text or supplementary material) or in open-science framework platforms such as the OSF: http://osf.io

Quality assessment

The methodological quality of the included studies can interfere in the results of a systematic review and lead to a overestimation of the effects of an intervention. Thus, authors should critically appraise the included studies.

Several validated tools for each type of study are available in the literature. For randomized controlled trials, the most used is the Cochrane Risk of Bias tool (ROB) [25]. For non-randomized studies, there is the Newcastle-Ottawa Scale or the Cochrane Risk of Bias in Non-Randomized Studies of Interventions (ROBINS-I) [26]. There are also checklists available on the Equator Network website to assess the report of studies, such as the Consolidated Standards of Reporting Trials (CONSORT) [27] for clinical trials; the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) [28] for observational studies; and the Case Report guidelines (CARE) [29] for case reports. In addition, the degree of evidence can be assessed using the Grading of Recommendations, Assessment, Development and Evaluation (GRADE) which generates an evidence grade and the strength of the recommendation for a given intervention and outcome. Learn more on this topic reading Chapter 04.

Evidence synthesis

Data synthesis is a way to gather several studies to produce conclusions on a group of evidence. For a better interpretation of the data, it is mandatory in a systematic review to present a table summarizing the information from the included studies. The data synthesis can begin with a comparison of the PICO items to observe which studies are similar enough to be grouped [30]. The results of the included studies should be qualitatively described when it is not possible to combine the data for statistical analysis, or quantitatively – through a meta-analysis (see Chapter 05) – when studies show the same characteristics and similar data for a given outcome [30-31]. If the data are similar, it is possible to increase the generalization of the results and to conclude that the intervention has an applicable effect to a larger number of patients. However, if the studies present heterogeneity with several variations in the basal characteristics and a meta-analysis is not possible, only a narrative description of the outcomes is recommended [14].

Scoping review

Scoping reviews are exploratory studies addressed to map a topic in a systematic manner, to identify concepts, theories, sources of evidence and gaps in the scientific knowledge [32]. A scoping review is usually conducted when the literature on a theme is heterogeneous and complex or when a mapping the existing evidence is needed [33]. Whereas systematic reviews aim to answer a specific and precise question, scoping reviews are addressed to answer one or more broad questions and to map the key concepts on an issue [34].

Scoping reviews essentially follow the same phases as a systematic review, but their focus and application are substantially different. While systematic

reviews have an analytical and quantitative nature, scoping reviews explore and describe the data without a formal synthesis of the results [35, 36]. The quality assessment of the included studies – a mandatory phase of a systematic review – is not required and often not recommended for scoping reviews, because the objective is to map the available evidence rather than to produce an objective clinical answer for a specific question [35]. For that reason, scoping reviews are not focused on producing recommendations for clinical practice, but on providing an overview of the evidence related to the nature and the array of the available knowledge on a topic.

The choice of a scoping review is appropriate when the research aims to identify the type of available evidence in a field of knowledge, as a precursor phase of a systematic review, to find research gaps, to clarify key concepts in literature, to evaluate how research has been conducted in a specific area and to identify the characteristics related to a concept. A scoping review might be applied also to develop policy maps through the identification of evidence from reports and documents that guide clinical practice in specific fields [33, 37].

A scoping review allows the inclusion of several types study's design and methodologies according to the aims of the review, such as qualitative, narrative, and quantitative studies, or any other policy document or website data [34, 35]. Manual and grey literature searches are extensively conducted in scoping reviews in addition to a search in at least one of the traditional databases. Studies' selection is performed as in systematic reviews by the definition of the inclusion and exclusion criteria can be broader. The results should be presented clearly to summarize and describe the data, generally without complex statistical methods. It is recommended to present the results in maps, tables, figures, or diagrams to facilitate data interpretation. Documentation of all of the processes is

fundamental to the transparency and trustworthiness of the review, from the writing and publication of the research protocol to the final report [35-37].

Unlike systematic reviews, which have been developed and consolidated since the 1970s and 1980s, scoping reviews are relatively new methods. The first methodological guideline on this type of review was published in 2005 [34], and an increase in the publication of scoping reviews by international researchers is being currently observed [35, 38]. However, there is still a lack of standard procedures for conducting and especially reporting these studies. Even if scoping reviews have some methodological flexibility compared to systematic reviews, they should be systematically conducted to reduce bias and to produce reliable results [35]. Here we will briefly present the main steps for conducting scoping reviews according to the Manual from the Joanna Briggs Institute, which is currently the main international organization that conducts, publishes, and guides authors to perform high quality scoping reviews. For further information, please visit: https://synthesismanual.jbi.global

Main steps for conducting a scoping review

See Figure 3 for the main steps of a scoping review. These phases are overall similar to those that should be followed when performing a systematic review.

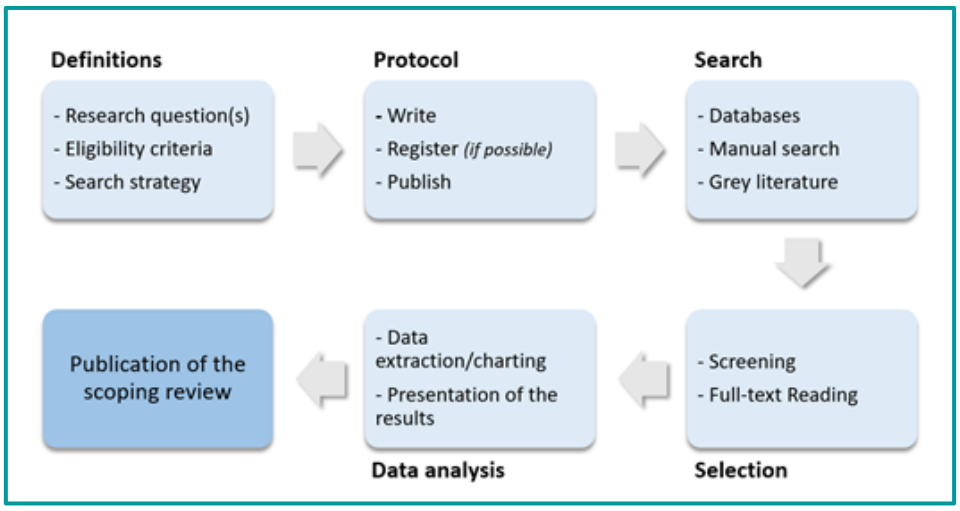


Figure 3. Summary of the main steps for conducting a scoping review

Review question

The definition of the research question should be guide by the review aims. Scoping reviews usually present broader objectives with one or more research questions. The question(s) will guide the formulation of inclusion criteria and should contain the elements corresponding to the study objectives that might be described by the PCC mnemonic (Participants, Concept and Context). It is important to highlight that in scoping reviews it is not necessary to provide a previous description on the outcomes, interventions, or interesting phenomena, although these elements might be useful to some issues in the review. An example of a hypothetical question is presented in the box below:

What strategies have been implemented by health professionals in homecare to the management of depression in patients with type 1 diabetes mellitus?

Eligibility criteria

The definition of the inclusion criteria of a scoping review should consider the sources of evidence that will be included. As scoping reviews might include any source of evidence (e.g. primary studies, narrative reviews, systematic reviews, guidelines, websites). Authors should detail what sources will be used according to their usefulness and adequacy to the review's objectives. The PCC elements should be defined in the inclusion criteria; the characteristics of the population, when applicable, should be specified and might include the age, sex and other attributes according to the objectives. The concept corresponds to the 'interventions', 'interesting phenomena' and 'outcomes' according to their relevance to the scoping review. The context might include geographical (e.g. country), social, cultural factors or other more specific (e.g. hospital, primary health care, community) elements to refine the scope of the review. Considering the above-mentioned question, the PCC might be described as follows:

RESEARCH QUESTION
What strategies have been implemented by health professionals in homecare to the management of depression in patients with type 1 diabetes mellitus?

PCC
Population: patients with type 1 diabetes mellitus
Concept: strategies implemented by health professionals to the management of depression in type 1 diabetes mellitus
Context: homecare

Exclusion criteria should also be defined by authors, and can include the period of publication (according to context) and language limitations (e.g. non-Roman characters). However, the JBI strongly recommends no restriction for language, unless there are clear reasons (e.g. feasibility) for it.

Search strategy

The search strategy for a scoping review should be comprehensive and able to identify all of the published (e.g. primary researches, reviews) and unpublished (grey literature) sources of evidence. The strategy should be formulated with the keywords and indexed terms found in the relevant articles. The controlled vocabulary from the databases can also be also used, as in narrative or systematic reviews. Boolean operators AND, OR should be used to group the terms, besides other resources from each database. The description of the search strategy for at least one of the main databases should be documented and reported by the authors.

Protocol

As in for systematic reviews, a protocol should be developed prior the conduction of the scoping review. Several aspects should be defined and described in the document, such as the objectives, questions, methods, and a plan for the presentation of the results. The protocol should contain a detailed description of the inclusion and exclusion criteria, search strategy, sources of evidence, how the data will be collected and presented and, if some change in the process happens, it should be clarified and explained in the report. It is important that the protocol be available through the register or publication. Authors might use online platforms such as the OSF or Research Gate to make the protocol available [35]. The JBI Evidence Synthesis Journal also publishes both protocols and full scoping reviews that follow the JBI methodology. The PROSPERO platform, used for the register of systematic reviews, is currently not available to register scoping reviews.

Databases

The search should be conducted in at least one traditional database and on other specific databases, journals or websites that are considered appropriate to the research objectives. Manual and grey literature searches are one of the main sources for gathering evidence in scoping reviews. The studies found in the databases can be imported to a reference manager. Duplicated registers should be removed. It is important that the entire process is documented (e.g. register the number of the studies retrieved in each search).

Screening of titles and abstracts

The title and abstracts retrieved in the search should be read by two reviewers independent and the results of the selection should be discussed in consensus meetings. Any disagreement on the inclusion of a study should be solved by a third reviewer. All of the studies that are clearly irrelevant to the aims of the scoping review should be excluded. The process should be documented, including the number of included and excluded sources of evidence. As in systematic reviews, the Kappa coefficient should be calculated to verify the concordance level between the reviewers [22].

Full-text reading

Once again, two reviewers independently should read the full-text and discuss which ones should be selected for data extraction. A third reviewer can participate during the consensus meetings if necessary. The selection is also based on the eligibility criteria, and the reasons for the exclusion of each source of evidence should be registered in a table and presented as an appendix in the

report of the review. The number of included and excluded studies should be documented and the concordance level between the reviewers calculated.

Data extraction and charting

The extraction of the data should be done according to the scoping review aims. A standard table or form should be developed to guide the collection of data, and this may include the characteristics of the studies and the main results. There is an example table for data extraction in the JBI Manual; however, it is recommended that each form be developed individually, in accordance with the objectives the review. To reduce bias and errors, this phase should be done by at least two reviewers independently. A pilot phase with the extraction of data from two or three sources of evidence might be useful to align the data that should be collect [34, 35].

Data analysis

Scoping reviews are not intended to synthesize data from their sources and generally include descriptive analysis of information (e.g. frequency counting) without performing complex statistical analysis. The main purpose of extracting and analyzing data in a scoping review is to identify, characterize and summarize the evidence on a topic, including the identification of research gaps. For a scoping review, data such as concept frequency counts, population characteristics and other descriptive data may be sufficient to answer the research question. Likewise, the qualitative data obtained should also not be treated in depth, but a descriptive analysis of the concepts and definitions on the topic could be performed. In some cases, it may be useful to apply qualitative content analysis methods to encode data in a given category. However, these

analyses should only be descriptive, and the authors should not carry out a thematic analysis or synthesis, as this would go beyond the scope of a scoping review – a systematic review of qualitative evidence may be more appropriate.

Presentation of findings

Presenting the results in a clear format may be useful for identifying gaps in the literature and mapping the available evidence on the topic. The results of the searches and the selection process should be described and presented in a flowchart according to the checklist for scoping reviews (PRISMA-ScR) [39]. Data mapping can be presented as tables, diagrams, graphs or described on the text. The PCC elements may help in choosing the best format to present the results. This presentation can include the type of evaluated studies, characteristics of the population, main outcomes analyzed, period of source publication, country of origin, area of intervention (clinic, policy, educational) and research methods. The results can also be classified into conceptual categories, such as 'type of intervention', 'population', 'objectives', 'main findings' and 'research gaps'.

Conclusion

This chapter provides the definition and the key elements of three types of scientific reviews: narrative, systematic and scoping reviews. The selection of a review will depend on the purpose of the research and the question(s) expected to be answered. Health sciences researchers should know the differences, advantages and limitations of these studies and clearly understand their methods. The conduction and reporting of systematic and scoping reviews should strictly follow international guidelines and checklists aiming at providing reliable evidence for end-users.

References

[1] Sackett DL, Rosenberg WMC, Gray JAM, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. 1996. BMJ (Clinical research ed). 1996;312(7023):71-2.
[2] Mulrow CD. Rationale for systematic reviews. Br Med J. 1994;309:597–9.
[3] Cook DJ, Mulrow CD, Haynes RB. Systematic reviews: Synthesis of best evidence for clinical decisions. Ann Intern Med. 1997;126(5):376–80.
[4] Cordeiro AM, de Oliveira GM, Rentería JM, Guimarães CA. Systematic review: A narrative review. Rev Col Bras Cir. 2007;34(6):428–31.
[5] Green BN, Johnson CD, Adams A. Writing narrative literature reviews for peer-reviewed journals: secrets of the trade. J Chiropr Med. 2006;5(3):101–17.
[6] De Luca-Canto G. Revisões sistemáticas da literatura: guia prático. 1.ed. Brazil Publishing. Curitiba: Brazil Publishing; 2020. 189p.
[7] Grant MJ, Booth A. A typology of reviews: An analysis of 14 review types and associated methodologies. Health Info Libr J. 2009;26(2):91–108.
[8] Demiris G, Oliver DP, Washington KT. Defining and Analyzing the Problem. In: Behavioral Intervention Research in Hospice and Palliative Care: Building an Evidence base. 1st ed. Academic Press; 2019. p. 27–39.
[9] Gasparyan AY, Ayvazyan L, Blackmore H, Kitas GD. Writing a narrative biomedical review: Considerations for authors, peer reviewers, and editors. Rheumatol Int. 2011;31:1409–17.
[10] Rother ET. Revisão sistemática X revisão narrativa. Acta Paul Enferm. 2007;20(2):1–2.
[11] Bastian H, Glasziou P, Chalmers I. Seventy-five trials and eleven systematic reviews a day: How will we ever keep up? PLoS Med. 2010;7(9):e1000326.
[12] Atallah AN, Castro AA. Revisão sistemática da literatura e metanálise. Diagnóstico Trat. 1998;2(2):12–5 (http://www.centrocochranedobrasil.org.br/apl/artigos/artigo_531.pdf)
[13] Clarke M, Horton R. Bringing it all together: Lancet-Cochrane collaborate on systematic reviews. Lancet. 2001;357.
[14] Brasil. Ministério da Saúde. Diretrizes metodológicas: elaboração de revisão sistemática e metanálise de ensaios clínicos randomizados. Ministério da Saúde. Brasília/DF; 2012. 92p.
[15] Lasserson TJ, Thomas J. Chapter 1: Starting a review. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (editors) Cochrane Handbook for Systematic Reviews of Interventions version 61 (updated September 2020). Cochrane; 2020 (www.training.cochrane.org/handbook)
[16] Pollock A, Berge E. How to do a systematic review. Int J Stroke. 2018;13(2):138–56.
[17] Thomas J, Kneale D, McKenzie JE, Brennan SE. Chapter 2: Determining the scope of the review and the questions it will address. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (editors) Cochrane Handbook for Systematic Reviews of Interventions version 61 (updated September 2020). Cochrane; 2020 (www.training.cochrane.org/handbook)
[18] Donato H, Donato M. Stages for undertaking a systematic review. Acta Med Port. 2019;32(3):227–35.
[19] McKenzie JE, Brennan SE, Ryan RE, Thomson HJ, Johnston RV. Chapter 3: Defining the criteria for including studies and how they will be grouped for the synthesis. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (editors) Cochrane Handbook

for Systematic Reviews of Interventions version 61 (updated September 2020). Cochrane; 2020 (www.training.cochrane.org/handbook)

[20] Lefebvre C, Glanville J, Briscoe S, Littlewood A, Marshall C, Metzendorf MI, et al. Chapter 4: Searching for and selecting studies. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (editors) Cochrane Handbook for Systematic Reviews of Interventions version 61 (updated September 2020). Cochrane; 2020 (www.training.cochrane.org/handbook)

[21] Moher D, Shamseer L, Clarke M, Ghersi D, Liberati A, Petticrew M, et al. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. Syst Rev. 2015;4(1):1–9.

[22] Feinstein AR, Cicchetti D V. High agreement but low Kappa: I. the problems of two paradoxes. J Clin Epidemiol. 1990;43(6):543–9.

[23] Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med. 2009 Jul 21;6(7):e1000097.

[24] Li T, Higgins JPT. Chapter 5: Collecting data. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (editors) Cochrane Handbook for Systematic Reviews of Interventions version 61 (updated September 2020). Cochrane; 2020.

[25] Higgins JPT, Savović J, Page MJ, Elbers RG. Chapter 8: Assessing risk of bias in a randomized trial. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (editors) Cochrane Handbook for Systematic Reviews of Interventions version 61 (updated September 2020). Cochrane; 2020 (www.training.cochrane.org/handbook)

[26] Sterne J, Hernán M, Reeves B, Savović J, Berkman N, Viswanathan M, et al. Risk Of Bias In Non-randomized Studies of Interventions (ROBINS-I): detailed guidance. BMJ. 2016;355:i4919.

[27] Schulz KF, Altman DG, Moher D. CONSORT 2010 Statement: Updated guidelines for reporting parallel group randomised trials. BMJ. 2010;340(7748):698–702.

[28] Elm E von, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. Strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. BMJ. 2007;335(7624):806–8.

[29] Gagnier JJ, Kienle G, Altman DG, Moher D, Sox H, Riley D, et al. The CARE guidelines: Consensus-based clinical case reporting guideline development. BMJ Case Rep. 2013.

[30] McKenzie JE, Brennan SE, Ryan RE, Thomson HJ JR. Chapter 9: Summarizing study characteristics and preparing for synthesis. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (editors) Cochrane Handbook for Systematic Reviews of Interventions version 61 (updated September 2020). Cochrane; 2020 (www.training.cochrane.org/handbook)

[31] Deeks JJ, Higgins JPT AD (editors). Chapter 10: Analysing data and undertaking meta-analyses. In: In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (editors) Cochrane Handbook for Systematic Reviews of Interventions version 61 (updated September 2020) . Cochrane; 2020 (www.training.cochrane.org/handbook)

[32] Grimshaw J. A guide to knowledge synthesis: a knowledge synthesis chapter. Canadian Institutes of Health Research. 2020; 56

[33] Anderson S, Allen P, Peckham S, Goodwin N. Asking the right questions: Scoping studies in the commissioning of research on the organisation and delivery of health services. Health Research Policy and Systems. BioMed Central; 2008.

[34] Arksey H, O'Malley L. Scoping studies: Towards a methodological framework. Int J Soc Res Methodol Theory Pract. 2005;8(1):19–32.

[35] Peters MDJ, Marnie C, Tricco AC, Pollock D, Munn Z, Alexander L, et al. Updated methodological guidance for the conduct of scoping reviews. JBI Evid Synth. 2020;18(10):2119–26.

[36] Peterson J, Pearce PF, Ferguson LA, Langford CA. Understanding scoping reviews: Definition, purpose, and process. J Am Assoc Nurse Pract. 2017;29(1):12–6.

[37] Munn Z, Peters MDJ, Stern C, Tufanaru C, McArthur A, Aromataris E. Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. BMC Med Res Methodol. 2018 Nov 19;18(1).

[38] Tricco AC, Lillie E, Zarin W, O'Brien K, Colquhoun H, Kastner M, et al. A scoping review on the conduct and reporting of scoping reviews. Vol. 16, BMC Medical Research Methodology. BioMed Central Ltd.; 2016.

[39] Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. Ann Intern Med. 2018;169(7):467–73.

QUALITY EVALUATION AND EVIDENCE APPRAISAL

ALINE F. BONETTI

Abstract

Systematic reviews and meta-analyses represent major tools for integrating the available information and are therefore placed in the top of the pyramid of scientific evidence hierarchy. However, the validity of systematic reviews and meta-analyses' findings rely on the methodological quality and risk of bias of the included studies, as well as on the strength of the generated evidence. Researchers should be able to critically appraise both primary and secondary studies, especially when formulating conclusions and recommendations to guide decision-making processes in healthcare. This chapter aims to describe the basic concepts of methodological quality evaluation and appraisal of evidence. Validated international tools such as the Revised tool for Risk of Bias in randomized trials (RoB 2.0), the Risk of Bias in non-randomized Studies of Interventions (ROBINS- I tool), and the Grading of Recommendations Assessment, Development and Evaluation (GRADE system) will be presented. Additionally, some recommendations to produce good-quality systematic reviews and meta-analyses, such as following specific checklists and guidelines as the PRISMA statement (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) will be discussed in this chapter.

Keywords: methodological quality; quality of evidence; risk of bias

Introduction

The assessment of methodological quality, risk of bias, and the strength of evidence are essential steps during the conduction, reporting and interpretation of findings of systematic review according the PRISMA statement (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) [1] and the recommendations of the Cochrane Collaboration [2] (Chapters 01 and 03). These assessments can draw attention to potential flaws and bias in the summarized information and contribute to the certainty in the overall evidence [3].

The critical appraisal of the included studies in a systematic review should be performed by trained reviewers using validated tools selected according to study's design. Likewise, the assessment of the quality of evidence should be strictly performed using international recognized instruments such as the GRADE (Grading of Recommendations Assessment, Development and Evaluation) system [2]. This assessment is important for the development of clinical recommendations grounded on the available evidence, which may guide, among others, the improvement of practical guidelines on a given medical field. All of these evaluations should be completely documented, which contributes to science transparency and reliability [4, 5].

Additionally, the authors of systematic reviews and meta-analyses should perform their studies following international guidelines and checklists for studies' conduction and reporting. Raw data should always be available to readers to allow replicability of findings [5].

This chapter addresses the main tools for methodological quality and risk of bias evaluation, the graduation system for the assessment of the quality of evidence, and some recommendations to produce good-quality systematic reviews and meta-analyses.

Tools for methodological quality and risk of bias assessment

Bias refers to systematic error, meaning that multiple replications of the same study would reach the wrong answer on average. For randomized and non-randomized studies, the Cochrane Collaboration defines bias as a tendency for study's results to differ systematically from the results expected from a traditional and appropriate randomized trial, conducted on the same participant group that had no errors in its conduct. This would typically be a large trial that conducts all the required steps appropriately: concealment of randomized allocation; blinding patients, health care professionals, and outcome assessors; guaranteeing the reach of the outcome results in all randomized participants; and reporting intervention effects for all measured outcomes [2]. On the other hand, quality is not well defined and can include study characteristics (e.g. performing a sample size calculation) that are not inherently to bias in the study's results [6].

There are different tools for quality and risk of bias assessment, which are specific to each study design and are available on some international and well-recognized handbooks, such as the Cochrane Handbook [2], and the Joanna Briggs Institute manual [7]. Other instruments are also available, such as the Jadad score [8], and the Newcastle-Ottawa tool [9] for methodological quality assessment of randomized clinical trials and observational studies, respectively. Although validated, these last are less are recommended by experts. We can additionally mention the QUADAS-2 (Quality Assessment of Diagnostic Accuracy Studies – version 2), an instrument with four domains used to evaluate the risk of bias and applicability of primary diagnostic accuracy studies [10]. Table 1 summarizes the main characteristics of the most used tools for quality and risk of bias assessment of primary studies usually included in a systematic review.

Table 1. Main tools for quality and risk of bias assessment

Tool	Study design	Final judgment	Advantages	Limitations
Jadad	RCT	Numerical score (maximum of 5)	Simple and easy to use, freely available	Few addressed questions; may not reveal the real quality or risk of bias
JBI critical appraisal checklist	Checklists available for different study's design	Each tool provides a different final judgment	Overall, these instruments are simple and easy to use. The JBI provides manuals for researchers	Often the tools have many domains of subjective interpretation and are time-consuming
Newcastle-Ottawa Scale (NOS)	Observational studies as cohorts and case-controls	'Star system' (more stars, the higher the quality of the study)	Simple and easy to use, freely available	Few addressed questions; may not reveal the real quality of the study. Items of subjective interpretation
QUADAS-2 tool	Diagnostic accuracy studies	Low, unclear, or high risk of bias	More complete and accurate than its first version, easy to use, freely available	Complex, subjective interpretation, time-consuming
Rob 2.0 tool	RCT	Low risk of bias, some concerns, or high risk of bias	More complete and accurate than its first version, freely available	Complex, subjective interpretation, time-consuming
ROBINS-I tool	NRCT	Low, moderate, serious, or critical risk of bias	More complete and accurate than other tools such as NOS, freely available	Complex, subjective interpretation, time-consuming

JBI: Joanna Briggs Institute; NICE: National Institute for Clinical Excellence; NIH: National Institutes of Health; NRCT: Non-randomized controlled trials; RCT: Randomized controlled trials.
Source: adapted from Ma et al., 2020.

The Cochrane Collaboration recommends the use of two instruments for quality and risk of bias appraisal of primary studies:

- ***RoB 2.0 tool: revised tool for Risk of Bias in randomized trials***
- ***ROBINS- I tool: Risk of Bias in non-randomized Studies of Interventions***

Both tools focus on a study's internal validity. The studies with the low methodological quality or high risk of bias should not be excluded from the systematic review, but the results should be interpreted with caution and they may be considered when synthesizing evidence [2].

Overview of the RoB 2.0 tool

The RoB 2.0 tool (a revised tool for Risk of Bias in randomized trials) is structured into a set of domains of bias, with a series of questions (signaling questions), focusing on trial design, conduction, and reporting. Different from its first version, each assessment using the RoB 2.0 tool should be applied to each outcome of the trial [6].

The full guidance document, the crib sheet summarizing the signaling questions, and the template for completing the assessment of the RoB 2.0 tool are available at **www.riskofbias.info**

The authors of systematic review and meta-analyses must clearly understand each domain by reading the detailed explanations of its concepts in the guidance. It is also recommended that the authors download a free standardized spreadsheet available on website, named 'Excel tool to implement RoB 2.0', which facilitates the handling of the instrument. The tool contains five domains, and for each of them, there are the following components:

- Signaling questions that guide the authors for the evaluation
- A judgement about the risk of bias for the given domain
- Free text boxes to justify the judgment of the signaling questions
- An option to predict and explain the likely direction of bias

The response options for each signaling question are:

- Yes
- Probably yes
- Probably no
- No
- No Information

Responses of 'yes' and 'probably yes', 'no', and 'probably no' have the same implications for risk of bias. The definitive responses ('yes' and 'no') mean that clear evidence is available about the question, whilst the 'probably' versions imply that a judgment has been made by the researchers. When there are insufficient details about a topic, the evaluation should be as 'no information' available. Authors can also answer 'not applicable' for questions that a response is not required [6].

This version of the tool incorporates improvements compared to the original tool (RoB) aiming at increasing the reliability of assessments, considering the developments in understanding how bias arises in randomized trials. Although this tool is completer and more accurate than the first version, researcher's perceptions of its usability in practice are still conflicting.

Ideally, the judgments should be supported by written justifications and be performed by two reviewers independently, with the participation of a third reviewer in case of disagreement. The assessment should be performed for each outcome under study. Therefore, the time to conduct the complete assessment and to reach consensus between reviewers are some of the reported barriers to use the new version of the tool [2, 6].

The domains of the RoB 2.0 tool and the issues addressed by them are summarized in Table 2. The signaling questions should be answered for each outcome and independently (i.e the answer to one question should not affect answers to other questions). After that and according to algorithms that map responses to signaling questions, each domain will be judged as:

- Low risk of bias
- Some concerns
- High risk of bias

Table 2. Bias domains of the RoB 2.0 tool

Domain	Brief explanation	Signaling question	Response
1. Bias arising from the randomization process	This domain considers the whole process of randomization	**1.1** Was the allocation sequence random? **1.2**Was the allocation sequence concealed until participants were enrolled and assigned to interventions? **1.3** Did baseline differences between intervention groups suggest a problem with the randomization process?	• Yes • Probably yes • No • Probably no • No information
2. Bias due to deviations from intended interventions*	This domain relates to biases that arise when there are deviations from the intended interventions, which are those specified in the trial protocol	**2.1** Were participants aware of their assigned intervention during the trial? **2.2** Were care and people delivering the interventions aware of participants' assigned intervention during the trial? **Additional questions (effect of assignment to intervention):** **2.3** If Y/PY/NI to 2.1 or 2.2: Were there deviations from the intended intervention that arose because of the trial context? **2.4** If Y/PY to 2.3: Were these deviations likely to have affected the outcome? **2.5** If Y/PY/NI to 2.4: Were these deviations from intended intervention balanced between groups? **2.6** Was an appropriate analysis used to estimate the effect of assignment to intervention? **2.7** If N/PN/NI to 2.6: Was there potential for a substantial impact (on the result) of the failure to analyze participants in the group to which they were randomized? **Additional questions (effect of adhering to intervention):** **2.3** If Y/PY/NI to 2.1 or 2.2: Were important non-protocol interventions balanced across intervention groups? **2.4** Were there failures in implementing the intervention that could have affected the outcome? **2.5** Was there non-adherence to the intervention regimen that could have affected participants' outcomes? **2.6** If N/PN/NI to 2.3, or Y/PY/NI to 2.4 or 2.5: Was an appropriate analysis used to estimate the effect of adhering to the intervention?	• Yes • Probably yes • No • Probably no • No information Not applicable (for 2.3, 2.4, 2.5, 2.6, 2.7)

Domain	Brief explanation	Signaling question	Response
3. Bias due to missing outcome data	This domain considers if the data for that outcome were available for all, or nearly all, participants randomized, or if there was evidence that the result was not biased by missing outcome data.	**3.1** Were data for this outcome available for all, or nearly all, participants randomized? **3.2** If N/PN/NI to 3.1: Is there evidence that the result was not biased by missing outcome data? **3.3** If N/PN to 3.2: Could missingness in the outcome depend on its true value? **3.4** If Y/PY/NI to 3.3: Is it likely that missingness in the outcome depended on its true value?	• Yes • Probably yes • No • Probably no • No information Not applicable (for 3.2, 3.3, 3.4)
4. Bias in measurement of the outcome	The outcomes in randomized trials should be assessed using appropriate outcome measures, and the same for across included groups	**4.1** Was the method of measuring the outcome inappropriate? **4.2** Could measurement or ascertainment of the outcome have differed between intervention groups? **4.3** If N/PN/NI to 4.1 and 4.2: Were outcome assessors aware of the intervention received by study participants? **4.4** If Y/PY/NI to 4.3: Could assessment of the outcome have been influenced by knowledge of intervention received? **4.5** If Y/PY/NI to 4.4: Is it likely that assessment of the outcome was influenced by knowledge of intervention received?	• Yes • Probably yes • No • Probably no • No information Not applicable (for 4.3, 4.4, 4.5)
5. Bias in selection of the reported result	This domain addresses bias that arises because the reported result is selected (based on statistical significance, for example) from among multiple intervention effect estimates that were calculated by the trial authors.	**5.1** Were the data that produced this result analyzed in accordance with a pre-specified analysis plan that was finalized before unblinded outcome data were available for analysis? **Is the numerical result being assessed likely to have been selected, on the basis of the results, from...** **5.2** ...multiple eligible outcome measurements (scales, definitions, time points) within the outcome domain? **5.3** ...multiple eligible analyses of the data?	• Yes • Probably yes • No • Probably no • No information

For accessing the complete tool and guidance, visit **www.riskofbias.info**

NI: no information; N: No; PN: probably no; PY: probably yes; Y: yes; * for the assessment of this domain, the reviewers may define which effect they will quantify: the effect of assignment to the interventions at baseline, regardless of whether the interventions are received as intended (the 'intention-to-treat effect'), or the effect of adhering to the interventions as specified in the trial protocol (the 'per-protocol effect').

Source: adapted from Sterne et al., 2019

The randomization process, when adequately performed, avoids the influence of confounding factors on the assignment of individual participants to intervention groups. Additionally, allocation sequence concealment aims to prevent bias in intervention assignment by avoiding trial personnel and participants from knowing the allocation sequence before and until the assignment.

Biases that arise when there are deviations from the intended interventions (i.e. those specified in the trial protocol) may also occur when the study is not well-designed. In this domain of the tool, additional questions according to the evaluation of the effect (assignment to intervention or adhering to the intervention) should be answered, including the frequency of deviations (e.g. if they were unbalances between groups), and the impact of the deviations on the overall result. Finally, bias from selection of reported results may be avoided by the publication or availability (e.g protocol) of the trial' planned analyses.

If the authors choose to use the Excel tool to perform the RoB 2.0 analyses (freely provided by Cochrane Collaboration), the judgments are automatically generated based on the tool' algorithm. The overall risk of bias – based on the result of the individual domains – is presented on Table 3.

Table 3. Overall risk of bias judgement (RoB 2.0)

Judgement	Criteria
Low risk of bias	The trial is judged to be at low risk of bias for all domains for this result
Some concerns	The trial is judged to raise some concerns in at least one domain for this result, but not to be at high risk of bias for any domain
High risk of bias	The trial is judged to be at high risk of bias in at least one domain for this result ***OR*** The trial is judged to have some concerns for multiple domains in a way that substantially lowers confidence in the result

For assessing the complete tool and guidance, visit **www.riskofbias.info**

Source: adapted from Sterne et al., 2019.

Overview of the ROBINS-I tool

The ROBINS-I tool (Risk of Bias In Non-randomized Studies of Interventions) is intended to the evaluation of non-randomized studies of interventions (NRSI) that compare the health effect of two or more groups. Controlled trials, which did not use the randomization process to allocate the participants, and observational studies that estimate the effectiveness (harm or benefit) of an intervention, such as cohort studies, case-control studies, controlled before-and-after studies, interrupted-time-series studies, can also be evaluated, in terms of risk of bias, according to this tool. Similar to the RoB 2.0, ROBINS-I considers the evaluation of the domains by each outcome under study [12]. The template and detailed guidance for using this tool is available at **www.riskofbias.info**

Compared to the also well-known New Castle Ottawa instrument (NOS) for observational studies, the ROBBINS-I tool is considered more objective, complete, and accurate. However, the time to perform the assessment and reach consensus between the reviewers may be larger if compared to NOS [2, 9, 13]. Both tools are validated and widely used.

ROBINS-I contains seven different domains, with some particular signaling questions, which may be judged according to the following answers:

- Yes
- Probably yes
- Probably no
- No
- No Information

The rationale for these answers is the same as for randomized controlled trials when using the RoB 2.0 tool. The first two domains address issues before

the start of the interventions and the third domain addresses classification of the interventions themselves. The other four domains address issues after the start of interventions. The judgments should be supported by justifications and require the participation of two reviewers independently; in cases of disagreement, a third reviewer should be contacted [12].

The Table 4 summarizes the ROBINS-I domains. There are four categories of risk of bias for each domain, which depends on the assigned responses for the signaling question: low, moderate, serious, critical or no information. For details of those judgments, it is necessary to read the full guidance carefully [12, 14]. Finally, after judging each domain, the evaluator should provide the overall risk of bias, according to Table 5.

Table 4. Bias domains of the ROBINS-I tool

Domain	Brief explanation	Signaling question	Response
		Pre-intervention	
1. Bias due to confounding	Baseline confounding occurs when one or more prognostic variables also predicts the intervention received at baseline	**1.1** Is there potential for confounding of the effect of intervention in this study? If Y/PY to 1.1: determine whether there is a need to assess time-varying confounding: **1.2** Was the analysis based on splitting participants' follow up time according to intervention received? If N/PN, answer questions relating to baseline confounding (1.4 to 1.6). 1.4 If Y/PY, proceed to question 1.3. **1.3** Were intervention discontinuations or switches likely to be related to factors that are prognostic for the outcome? If N/PN, answer questions relating to baseline confounding (1.4 to 1.6). If Y/PY, answer questions relating to both baseline and time-varying confounding (1.7 and 1.8) **Baseline confounding only:** **1.4** Did the authors use an appropriate analysis method that controlled for all the important confounding domains? **1.5** If Y/PY to 1.4: Were confounding domains that were controlled for measured validly and reliably by the variables available in this study? **1.6** Did the authors control for any post-intervention variables that could have been affected by the intervention? **Baseline, time-varying confounding:** **1.7** Did the authors use an appropriate analysis method that adjusted for all the important confounding domains and for time-varying confounding? **1.8** If Y/PY to 1.7: Were confounding domains that were adjusted for measured validly and reliably by the variables available in this study?	• Yes • Probably yes • No • Probably no • No information Not applicable (1.2, 1.3, 1.4, 1.5, 1.6, 1.7)

Domain	Brief explanation	Signaling question	Response
2. Bias in selection of participants into the study	This domain considers if participants were selected according to specific features after the start of the intervention	**2.1** Was selection of participants into the study (or into the analysis) based on participant characteristics observed after the start of intervention? If N/PN to 2.1: go to 2.4 **2.2** If Y/PY to 2.1: Were the post-intervention variables that influenced selection associated with intervention? **2.3** If Y/PY to 2.2: Were the post-intervention variables that influenced selection likely to be influenced by the outcome or a cause of the outcome? **2.4** Do start of follow-up and intervention coincide for participants? **2.5** If Y/PY to 2.2 and 2.3, or N/PN to 2.4: Were adjustment techniques used that are likely to correct for the presence of selection biases?	• Yes • Probably yes • No • Probably no • No information Not applicable (for 2.2, 2.3, 2.5)
		At intervention	
3. Bias in classification of interventions	The groups included in the NRSI should be well defined, clear, and explicit	**3.1** Were intervention groups clearly defined? **3.2** Was the information used to define intervention groups recorded at the start of the intervention? **3.3** Could classification of intervention status have been affected by knowledge of the outcome or risk of the outcome?	• Yes • Probably yes • No • Probably no • No information
		Post-intervention	
4. Bias due to deviations from intended interventions*	This domain considers the deviations in the intervention groups, and the bias arises when there are systematic differences between them	**If your aim is to assess the effect of assignment to intervention see questions 4.1 and 4.2:** **4.1** Were there deviations from the intended intervention beyond what would be expected in usual practice? **4.2** If Y/PY to 4.1: Were these deviations from intended intervention unbalanced between groups and likely to have affected the outcome? **If your aim is to assess the effect of starting and adhering to intervention, answer questions 4.3 to 4.6** **4.3** Were important co-interventions balanced across intervention groups? **4.4** Was the intervention implemented successfully for most participants? **4.5** Did study participants adhere to the assigned intervention regimen? **4.6** If N/PN to 4.3, 4.4 or 4.5: Was an appropriate analysis used to estimate the effect of starting and adhering to the intervention?	• Yes • Probably yes • No • Probably no • No information Not applicable (for 4.2, 4.6)

Domain	Brief explanation	Signaling question	Response
5. Bias due to missing data	This domain considers the reporting of the results, which, preferably, should be available for all, or nearly all participants	**5.1** Were outcome data available for all, or nearly all, participants? **5.2** Were participants excluded due to missing data on intervention status? **5.3** Were participants excluded due to missing data on other variables needed for the analysis? **5.4** If PN/N to 5.1, or Y/PY to 5.2 or 5.3: Are the proportion of participants and reasons for missing data similar across interventions? **5.5** If PN/N to 5.1, or Y/PY to 5.2 or 5.3: Is there evidence that results were robust to the presence of missing data?	• Yes • Probably yes • No • Probably no • No information Not applicable (for 5.1, 5.4, 5.5)
6. Bias in measurement of outcomes	This domain considers the bias introduced by either differential or non-differential errors in the measurement of outcome data	**6.1** Could the outcome measure have been influenced by knowledge of the intervention received? **6.2** Were outcome assessors aware of the intervention received by study participants? **6.3** Were the methods of outcome assessment comparable across intervention groups? **6.4** Were any systematic errors in measurement of the outcome related to intervention received?	• Yes • Probably yes • No • Probably no • No information
7. Bias in selection of the reported result	This domain considers the selective reporting of results by the authors in a way that depends on the findings	Is the reported effect estimate likely to be selected, on the basis of the results, from... **7.1** ... multiple outcome measurements within the outcome domain? **7.2** ... multiple analyses of the intervention-outcome relationship? **7.3** ... different subgroups?	• Yes • Probably yes • No • Probably no • No information

For accessing the complete tool and guidance, visit **www.riskofbias.info**
NI: no information; N: No; PN: probably no; PY: probably yes; Y: yes; * for the assessment of this domain, the reviewers may define which effect they will quantify: the effect of assignment to the interventions at baseline, regardless of whether the interventions are received as intended (the 'intention-to-treat effect'), or the effect of adhering to the interventions as specified in the trial protocol (the 'per-protocol effect').
Source: adapted from Sterne et al., 2016.

Table 5. Overall risk of bias judgement (ROBINS-I)

Judgement	Criteria	Interpretation
Low risk of bias	The study is judged to be at low risk of bias for all domains	The study is comparable to a well-performed randomized trial
Moderate risk of bias	The study is judged to be at low or moderate risk of bias for all domains	The study provides sound evidence for a non-randomized study but cannot be considered comparable to a well-performed randomized trial
Serious risk of bias	The study is judged to be at serious risk of bias in at least one domain, but not at critical risk of bias in any domain	The study has some important problems
Critical risk of bias	The study is judged to be at critical risk of bias in at least one domain.	The study is too problematic to provide any useful evidence and should not be included in any synthesis
No information	There is no clear indication that the study is at serious or critical risk of bias, and there is a lack of information in one or more key domains of bias	No information on which to base a judgement about risk of bias

For assessing the complete tool and guidance, visit **www.riskofbias.info**

Source: adapted from Sterne et al., 2016.

Quality of evidence appraisal

Systematic reviews provide essential, but sometimes not sufficient, information for guiding the clinical decision-making process. According to the methodological quality of the primary studies, their epidemiological design, and available results, certain technologies may or may not be recommended for use or approved in a given setting. Thus, the evaluation of the levels of evidence and proposal of classes of recommendations for a given intervention may be important to clinical decisions and are paramount for Health Technology Assessment processes [15].

Besides degrees of recommendation and levels of evidence proposed by the Oxford Center for Evidence-Based Medicine (CEBM) model (see Chapter 01), the quality of evidence appraisal can be performed according to the GRADE

system. This system allows an explicit approach to reduce unnecessary confusion on evidence interpretation and represents a useful tool for the development of healthcare guidelines to support decision-making process [15-17]. Access: **www.gradeworkinggroup.org**

The GRADE is a complex system that requires previous training of the researchers and a background knowledge of epidemiology. The evaluation of the quality of the evidence (also referred as 'evidence reliability') is performed by outcome and comparison (e.g. effects of intervention A vs. intervention B in the outcome X) [15-18]. The outcomes can be classified into 'critical', 'important' or 'less important' according to the practical setting and their relevance to the patient. The level of evidence is classified in categories as depicted in Table 6.

The initial classification of the quality of evidence is defined by the study design. Randomized clinical trials represent the more appropriate type of studies for evaluating intervention as demonstrated in Chapter 01. In this case, the level of evidence begins as 'high'. When only observational studies are included for providing evidence, the classification begins in the category 'low'.

From this initial classification, the judgment of some aspects of the studies allows reviewers to reduce or raise the level of evidence. Factors for ***downgrading*** the level of evidence include [15-24]:

- Study limitations (risk of bias)
- Inconsistency
- Evidence indirectness
- Imprecision
- Publication bias

On the other hand, if the level of evidence has not been reduced by these factors, the evidence from observational studies can be ***upgraded*** considering:

- Magnitude of the effect
- Dose-response gradient
- Plausible confounders

Similar to the other above mentioned methodological tools, it is recommended that two reviewers judge the quality of evidence independently, and in case of disagreements, a third reviewer should be contacted. The results of the quality of evidence appraisal should be presented in tables with the summary of findings for each issue [25].

Table 6. Level of evidence: GRADE approach

Level	Definition	Implications	Source of information
High	There is a great confidence that the true effect lies close to that of the estimate of the effect	Further research is very unlikely to change the confidence in the estimate of effect	Well-conducted randomized clinical trials, some well-conducted observational studies with consistent findings
Moderate	There is a moderate confidence in the effect estimate. The true effect is likely to be close to the estimate of the effect, but there is a possibility that it is substantially different	Further research is likely to have an important impact on the confidence in the estimate of effect and may change the estimate	Randomized clinical trials with small limitations, some observational studies with consistent findings
Low	The confidence in the effect estimate is limited. The true effect may be substantially different from the estimate of the effect	Further research is likely to have an important impact on the confidence in the estimate of effect and is likely to change the estimate	Randomized clinical trials with moderate limitations, observational studies with limitations

Level	Definition	Implications	Source of information
Very low	There is a truly little confidence in the effect estimate. The true effect is likely to be substantially different from the estimate of effect	Any estimate of effect is very uncertain	Randomized clinical trials with serious limitations, observational studies with many limitations. Expert opinion.

For assessing the complete tool and guidance, visit **www.gradeworkinggroup.org**
Source: adapted from Balshem et al, 2011.

These levels of evidence can guide the decision on the **strength** of the evidence, that can be classified into 'weak' or 'strong' and 'in favor' or 'against' an intervention. Grounded on these findings, clinical recommendations, practical guidelines, and further discussions (e.g. Delphi panel) can be built in a given healthcare area. For further uses in practice of the GRADE approach, read Chapter 08.

Quality of systematic reviews and meta-analyses

The clinical decisions should be based on high-quality studies, such as systematic reviews and meta-analyses, once they provide a synthesis of the current best information about a specific topic. However, to produce consistent and reliable systematic reviews and meta-analyses, the process of conduction and reporting of these studies should be strictly performed by researchers according to available guidelines and checklists. Suboptimal systematic reviews and meta-analyses are still common in the literature – accounting for over 80% of the literature. Major reasons for this science waste include lack of appropriate protocols, research duplication or plagiarism, methodological flaws beyond repair, selective reporting, or lack of transparency [5].

The PRISMA statement and its extensions (for network meta-analyses and scoping reviews, for example) provide checklists for authors to report their studies. The original PRISMA contains 27-items and was created in July 2009 to

designate the minimum standards for conducting and reporting a systematic review with meta-analyses [1]. This statement is also available on the Equator website (Enhancing the Quality and Transparency of Health Research), a platform that gathers more than 400 guidelines to guide the researchers to report their publication in the health area: **https://www.equator-network.org/**

On the other hand, the methodological quality and risk of bias of systematic reviews and meta-analyses can be evaluated by some other instruments, such as the Assessment of Multiple Systematic Reviews – version 2 (AMSTAR-2), the Revised Assessment of Multiple Systematic Reviews (R-AMSTAR) and the Cochrane Collaboration Risk of Bias in Systematic Reviews (ROBIS) [26-28].

Recently, a preliminary tool for assessing the risk of non-reporting biases in evidence syntheses was published, named ROB-ME (Risk of Bias due to Missing Evidence). This tool is directed for authors or users of systematic reviews to assess the risk of bias due to missing evidence in pairwise meta-analyses of the effects of interventions. Similar to RoB 2.0 and ROBINS-I, this instrument contains eight signaling questions, which should be judged by two reviewers independently. The tool also includes an algorithm that maps responses to the questions onto a proposed risk-of-bias judgment: low risk of bias; high risk of bias; some concerns [29].

Conclusions

The reliability of systematic reviews and meta-analyses depends, among others, on their quality. International guidelines and checklists such as the PRISMA statement and Cochrane Handbook, should be rigorously followed during the conduction and reporting of these studies. The quality and risk of bias

assessment of the primary studies included in a systematic review should also be strictly evaluated by using validated instruments, such as RoB 2.0 and ROBINS-I for randomized and non-randomized clinical trials, respectively. The quality of evidence can be assessed using the GRADE approach.

References

[1] Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. BMJ. 2009;339:b2535.

[2] Higgins JPT TJ, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (editors). Cochrane Handbook for Systematic Reviews of Interventions version 6.1 (updated September 2020): Cochrane; 2020 (www.training.cochrane.org/handbook)

[3] Guyatt G, Oxman AD, Akl EA, Kunz R, Vist G, Brozek J, et al. GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. J Clin Epidemiol. 2011;64(4):383-94.

[4] Bonetti AF, Della Rocca AM, Lucchetta RC, Tonin FS, Fernandez-Llimos F, Pontarolo R. Mapping the characteristics of meta-analyses of pharmacy services: a systematic review. Int J Clin Pharm. 2020;42(5):1252-60.

[5] Bonetti AF, Tonin FS, Della Rocca AM, Lucchetta RC, Fernandez-Llimos F, Pontarolo R. Methodological quality and risk of bias of meta-analyses of pharmacy services: A systematic review. Res Social Adm Pharm. 2020;S1551-7411(20)31236-5.

[6] Sterne JAC, Savović J, Page MJ, Elbers RG, Blencowe NS, Boutron I, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ. 2019;366:l4898.

[7] Joanna Briggs Institute. Joanna Briggs Institute Reviewers' Manual 2020. Available from: https://wiki.joannabriggs.org/display/MANUAL/JBI+Reviewer%27s+Manual+-+.pdf+updated+March+2020.

[8] Jadad AR, Moore RA, Carroll D, Jenkinson C, Reynolds DJ, Gavaghan DJ, et al. Assessing the quality of reports of randomized clinical trials: is blinding necessary? Control Clin Trials. 1996;17(1):1-12.

[9] Wells G, Shea B, O'Connell D, Robertson J, Peterson J, Welch V, et al. The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses: Available from: http://www.ohri.ca/programs/clinical_epidemiology/oxford.asp.

[10] Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med. 2011;155(8):529-36.

[11] Ma LL, Wang YY, Yang ZH, Huang D, Weng H, Zeng XT. Methodological quality (risk of bias) assessment tools for primary and secondary medical studies: what are they and which is better? Mil Med Res. 2020;7(1):7.

[12] Sterne JA, Hernán MA, Reeves BC, Savović J, Berkman ND, Viswanathan M, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ. 2016;355:i4919.

[13] Jeyaraman MM, Rabbani R, Al-Yousif N, Robson RC, Copstein L, Xia J, et al. Inter-rater reliability and concurrent validity of ROBINS-I: protocol for a cross-sectional study. Syst Rev. 2020;9(1):12.

[14] Sterne JAC, Higgins JPT, Elbers RG, BC R, ROBINS-I. atdgf. Risk Of Bias In Non-randomized Studies of Interventions (ROBINS-I): detailed guidance. 2016. Available from: http://www.riskofbias.info

[15] Petrisor BA, Keating J, Schemitsch E. Grading the evidence: levels of evidence and grades of recommendation. Injury. 2006;37(4):321-7.

[16] Andrews J, Guyatt G, Oxman AD, Alderson P, Dahm P, Falck-Ytter Y, et al. GRADE guidelines: 14. Going from evidence to recommendations: the significance and presentation of recommendations. J Clin Epidemiol. 2013;66(7):719-25.

[17] Andrews JC, Schünemann HJ, Oxman AD, Pottie K, Meerpohl JJ, Coello PA, et al. GRADE guidelines: 15. Going from evidence to recommendation-determinants of a recommendation's direction and strength. J Clin Epidemiol. 2013;66(7):726-35.

[18] Balshem H, Helfand M, Schünemann HJ, Oxman AD, Kunz R, Brozek J, et al. GRADE guidelines: 3. Rating the quality of evidence. J Clin Epidemiol. 2011;64(4):401-6.

[19] Guyatt GH, Oxman AD, Vist G, Kunz R, Brozek J, Alonso-Coello P, et al. GRADE guidelines: 4. Rating the quality of evidence--study limitations (risk of bias). J Clin Epidemiol. 2011;64(4):407-15.

[20] Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, et al. GRADE guidelines: 7. Rating the quality of evidence--inconsistency. J Clin Epidemiol. 2011;64(12):1294-302.

[21] Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, et al. GRADE guidelines: 8. Rating the quality of evidence--indirectness. J Clin Epidemiol. 2011;64(12):1303-10.

[22] Guyatt GH, Oxman AD, Kunz R, Brozek J, Alonso-Coello P, Rind D, et al. GRADE guidelines 6. Rating the quality of evidence--imprecision. J Clin Epidemiol. 2011;64(12):1283-93.

[23] Guyatt GH, Oxman AD, Montori V, Vist G, Kunz R, Brozek J, et al. GRADE guidelines: 5. Rating the quality of evidence--publication bias. J Clin Epidemiol. 2011;64(12):1277-82.

[24] Guyatt GH, Oxman AD, Sultan S, Glasziou P, Akl EA, Alonso-Coello P, et al. GRADE guidelines: 9. Rating up the quality of evidence. J Clin Epidemiol. 2011;64(12):1311-6.

[25] Guyatt GH, Oxman AD, Santesso N, Helfand M, Vist G, Kunz R, et al. GRADE guidelines: 12. Preparing summary of findings tables-binary outcomes. J Clin Epidemiol. 2013;66(2):158-72.

[26] Kung J, Chiappelli F, Cajulis OO, Avezova R, Kossan G, Chew L, et al. From Systematic Reviews to Clinical Recommendations for Evidence-Based Health Care: Validation of Revised Assessment of Multiple Systematic Reviews (R-AMSTAR) for Grading of Clinical Relevance. Open Dent J. 2010;4:84-91.

[27] Shea BJ, Reeves BC, Wells G, Thuku M, Hamel C, Moran J, et al. AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. BMJ. 2017;358:j4008.

[28] Whiting P, Savović J, Higgins JP, Caldwell DM, Reeves BC, Shea B, et al. ROBIS: A new tool to assess risk of bias in systematic reviews was developed. J Clin Epidemiol. 2016;69:225.

[29] Page M, JAC S, Boutron I, Hróbjartsson A, Kirkham J, Li T, et al. A tool for assessing Risk Of Bias due to Missing Evidence in a synthesis 2020. Available from: https://www.riskofbias.info/welcome/rob-me-tool.

BASIC METHODS AND TOOLS FOR STATISTICAL SYNTHESIS

ALEXANDRE F. COBRE
LETICIA P. LEONART
ROBERTO PONTAROLO

Abstract

The purpose of this chapter is to introduce the reader to basic concepts of statistics analyses frequently used in data synthesis on health technology assessment. First, the most used effect size measures in clinical studies are presented. Following, the basic concepts to acquaint the reader with survival analyses are discussed. Finally, an introduction to pairwise meta-analysis and its most important aspects are exposed. As an introductory text, formulas and equations are avoided, and simple practical examples are provided. After reading this chapter, we hope that the reader will be able to build foundation in basic statistics methods commonly used in health technology assessment and will be prepared to dive into more advanced texts on the subject.

Keywords: risk ratio; mean difference; survival analysis; meta-analysis

BASIC METHODS AND TOOLS FOR STATISTICAL SYNTHESIS

ALEXANDRE F. COBRE
LETICIA P. LEONART
ROBERTO PONTAROLO

Introduction

Statistics is a mathematical science that analyzes research data with the aim of describing it, synthetizing it and testing hypotheses. For example, a group of researchers can conduct a clinical trial to evaluate the efficacy of a new intervention to treat diabetes mellitus versus an old drug. One of the outcomes of interest can be the levels of glycated hemoglobin (HbA1c) after six months of treatment. First, to present the baseline data of both groups, the researchers will need descriptive statistics, such as patients' mean age along with the standard deviation for each group, as well as the proportion of males and females. To evaluate the efficacy of the new drug, the researchers will calculate the of HbA1c and the respective dispersion measure before and after the intervention. Then, to test the hypothesis that the new drug is better than the old one, the difference in means of each group will be compared.

Beyond that, researchers may be interested in conducting a long-term study to evaluate if the new drug reduces mortality and assess the overall survival in each group of patients. In this case, a survival analysis can be performed. Authors can also identify if other clinical trials have already been conducted comparing these interventions, and therefore decide to perform a meta-analysis to improve the statistical power and enhance confidence in the estimates. Without statistics, none of the above-mentioned analyses are possible. Statistical analyses can transform data into valuable information.

Thus, in this chapter, some basic aspects of statistics applied to health technology assessment will be discussed including: (i) effect size measures: calculation and interpretation; (ii) survival analyses; (iii) pairwise meta-analysis conduction.

ALEXANDRE F. COBRE
LETICIA P. LEONART
ROBERTO PONTAROLO

Effect-size measures

Effect size is a statistical concept that measures the strength of the relationship between two variables on a numeric scale. The first step in the statistical analysis of clinical studies is to understand the type of data. Among the most common, which will be discussed in this chapter, there are: (i) binary (or dichotomous) data: the variable must be classified in one of two categories (e.g. alive or dead); (ii) continuous data: the variable is a continuous number, i.e., it can assume any value in a given interval (e.g. height: 1.65, 1.76, 1.80); (iii) count data: corresponds to the number of events experienced by each individual in the study (e.g. number of fractures); and (iv) time-to-event data: also called survival data, it is the time until the occurrence of a binary event (e.g. the time until death) [1–3]. Each one of these cases will be discussed in the following sections.

One important concept when interpreting effect measures is to know which is the null value (i.e. the value that points to no difference between groups). For differences between two measures, the null value is the number zero (0). In the case of ratios (one number divided by another) the number one (1.00) corresponds to the null value. It is paramount to mention that, depending on how the groups are ordered during the comparison (for example, intervention B versus intervention A or intervention A versus intervention B), the interpretation of the effect measure can change (damage or benefit) [4,5].

ALEXANDRE F. COBRE
LETICIA P. LEONART
ROBERTO PONTAROLO

Effect measures for dichotomous outcomes

Dichotomous outcomes are those which can assume one of two values: for example, recovery from a disease or non-recovery, dead or alive, occurrence of an adverse event or non-occurrence. Some of the usually employed effect measures for this type of data are: a) Risk Ratio (RR), b) Odds Ratio (OR); c) Number needed to treat (NNT); and d) Risk difference (RD) [4].

Risk ratio

Risk is the probability that an event will occur. It can be presented in the form of a decimal number ranging from zero to one (0 – 1), or converted into a percentage (0% - 100%) [1]. For example, if among 100 patients 40 experienced an outcome, we can say that the risk of that outcome is 0,4 or 40%.

Following the same reasoning, the risk ratio (RR), also called the relative risk, compares the risk of an event between two groups. This comparison is made by dividing the risk in group A by the risk in group B [2,3]. A contingency table can be elaborated to facilitate the calculation and understanding of the RR, as shown in the Table 1, which helps organizing the data from each group into those who achieved or not the outcome.

Table 1. Contingency table

Outcome	Group 1	Group 2
-	a	b
+	c	d
Total	a + c	b + d

ALEXANDRE F. COBRE
LETICIA P. LEONART
ROBERTO PONTAROLO

Below is presented the RR formula and how it can be written considering the contingency table (Equation 1):

$$\text{RR} = \frac{\text{Risk in Group 1}}{\text{Risk in Group 2}} = \frac{\frac{a}{a+c}}{\frac{b}{b+d}} \quad (Equation\ 1)$$

Consider the following example: a randomized controlled trial compared valsartan 160 mg/day (n=35) with placebo (n=37) in patients with arterial hypertension. The treatment effect was measured by monitoring blood pressure at the outpatient level. At baseline, no patient had controlled blood pressure. After the end of follow-up, 11 patients in the valsartan group were found to have reduced the blood pressure, compared with 3 patients in the placebo group [4]. To calculate the RR between valsartan and placebo, first a contingency table can be built, see example in Table 2. Then, we can calculate the risk of not controlling the blood pressure with valsartan versus placebo, according to Equation 2.

Table 2. The contingency table for the example of valsartan versus placebo for arterial hypertension

Outcome	Valsartan	Placebo
Did not reduce blood pressure	24	34
Reduce blood pressure	11	3
Total	35	37

ALEXANDRE F. COBRE
LETICIA P. LEONART
ROBERTO PONTAROLO

$$RR = \frac{\frac{24}{35}}{\frac{34}{37}} = \frac{0.686}{0.919} = 0.746 \textit{ (Equation 2)}$$

This result means that the risk of not controlling the blood pressure with valsartan is 0.746 times the odds with placebo, i.e., valsartan reduced the risk of uncontrolled blood pressure in 25.4%.

Odds ratio

The odds of an event is the probability of occurrence of this event divided by the probability of its non-occurrence. The odds ratio (OR) is the odds of the event in one group divided by the odds of the same event in another group [5,6]. The understanding and calculation of the OR can be also elaborated with the use of a contingency table (Table 1) and its Equation 3 is shown below.

$$OR = \frac{\text{Odds in Group 1}}{\text{Odds in Group 2}} = \frac{\frac{a}{c}}{\frac{b}{d}} \textit{ (Equation 3)}$$

Still considering the example of valsartan versus placebo (see Table 2), the OR for uncontrolled blood pressure would be, according to Equation 4:

$$OR = \frac{\frac{24}{11}}{\frac{34}{3}} = \frac{2.18}{11.33} = 0.19 \textit{ (Equation 4)}$$

ALEXANDRE F. COBRE
LETICIA P. LEONART
ROBERTO PONTAROLO

This result means that the odds of not controlling the blood pressure with valsartan is 0.19 times the odds with placebo.

Risk difference

The risk difference (RD), which can be also termed the attributable risk (AR) or excessive risk, is defined as the difference between the risks of two groups [6]. The contingency table can be used to build the Equation 5:

$$\text{RD} = \text{Risk in Group 1} - \text{Risk in Group 2} = \frac{a}{(a+c)} - \frac{b}{(b+d)} \quad \textit{(Equation 5)}$$

Considering the example of valsartan versus placebo for hypertension, we can calculate the RD as follows (Equation 6):

$$\text{RD} = \frac{24}{(24+11)} - \frac{34}{(34+3)} = 0.686 - 0.919 = -0.233 \quad \textit{(Equation 6)}$$

In this example, the RD is a negative value, which indicates that the intervention has a protective effect. Therefore, the absolute difference in risk between valsartan and placebo for uncontrolled blood pressure is 23.3%.

Important note: risk ratio is not the same as odds ratio

As seen in the valsartan example, risk ratio and odds ratio are not the same (RR=0.746, OR=0.19). However, the non-equivalence of both measures does not mean that one of them is wrong: both OR and RR are fully valid to describe the magnitude of an effect. Problems can arise if the OR is

misinterpreted as the RR. For example, for clinical interventions that increase the occurrence of the event, the OR will be greater than the RR. In this case, the error in interpretation will tend to overestimate the effect of the intervention, especially when it comes to common events (i.e., risk of events greater than 20%). Unfortunately, errors of this type are frequently found in several studies published in the literature [7]. The exception is when the event is rare: in this case, the OR and the RR will provide similar values and an extrapolation is acceptable.

Number needed to treat

The number needed to treat (NNT) estimates the number of patients who need to be treated to prevent an additional bad outcome. NNT is an intuitive measure, because it is known that not all patients benefit from the intervention: some can improve, some can remain stable and some can deteriorate [4,8]. The NNT is the inverse of the RD, as shown in Equations 7 and 8:

$$\text{NNT} = \frac{1}{\frac{a}{(a+c)} - \frac{b}{(b+d)}} \textit{ (Equation 7)}$$

For the valsartan versus placebo example, the NNT is:

$$\text{NNT} = \frac{1}{0.233} = 4.29 \textit{ (Equation 8)}$$

As the number of patients is a variable than can only assume integer values, we can affirm that the number needed to treat with valsartan to avoid an additional case of uncontrolled blood pressure is 5.

Effect measures for continuous outcomes

Variables such as body mass index, height, time, temperature, which can assume infinite values in a given interval, are called continuous variables [9]. In clinical studies, it is common to calculate the mean value for a continuous outcome for each group. To compare this outcome between two groups, the mean difference (MD) can be employed, which is simply the mean of one group minus the mean of the other group [10]. In a single study, it is expected that both groups will be measured in the same way. For example, the same questionnaire of severity of symptoms will be applied to patients of the intervention and control groups, so the means are directly comparable. However, in the case of meta-analyses, where different clinical studies are pooled together, it is possible that the different clinical studies have assessed the same outcome using different scales. In this case, in order to make the studies comparable, the standardized mean difference (SMD) has to be used [11] [12]. The SMD is calculated using the Equation 9 below:

$$\text{SMD} = \frac{\text{difference in mean outcome between groups}}{\text{Standard deviation of outcome among patients}} \quad \textit{(Equation 9)}$$

ALEXANDRE F. COBRE
LETICIA P. LEONART
ROBERTO PONTAROLO

Effect measures for counts

Some medical events are recurrent, i.e., can happen to a person more than once during the study, such as hospitalization, adverse event, or cancer recurrence. In these situations, besides knowing if the event occurred or not, it is also important to know how many times it occurred in a specified period. The number of events divided by the person-time is called the incidence rate [13]. The person-year (or person-month, person-week) is the number of persons which were followed in the study by a whole year. For example, in a study with 100 patients who were followed for 12 months there are 100 person-years. However, if half of the patients (n=50) stayed on the study for only 6 months, the number of person-years would be 75 (50 persons who stayed for the whole year contributed as 50 person-years and the other 50 persons who stayed for only 6 months contributed as 0.5 person-year each, resulting in 25 person-years). Therefore, the incidence rate (sometimes referred simply as rate), can be calculated as follows (Equation 10):

$$\text{Rate} = \frac{\text{number of events in a specified period}}{\text{population at risk (person}-\text{time)}} \quad \textit{(Equation 10)}$$

A cohort study assessed the mortality of HIV/AIDS in a military hospital of Mozambique in patients under tenofovir + lamivudine + efavirenz treatment. During 624 person-years of follow-up, 39 patients died [14]. The mortality rate can be calculated according to Equation 11:

ALEXANDRE F. COBRE
LETICIA P. LEONART
ROBERTO PONTAROLO

$$\text{Rate} = \frac{39}{624} = 0.062 \textit{ (Equation 11)}$$

The result shows that the mortality rate is 0.062 per person-year, which can be also expressed as 6.2 per 100 person-years. To compare the rate between two groups (rate ratio), one should simply divide the rate in the first group by the rate in the second group (see Equation 12).

$$\text{Rate ratio} = \frac{\text{Events per person-time in Group 1}}{\text{Events per person-time in Group 2}} \textit{ (Equation 12)}$$

This result means that the risk of not controlling the blood pressure with valsartan is 0.746 times the risk with placebo, i.e., valsartan reduced the risk of uncontrolled blood pressure in 25.4%.

Survival analysis

In survival analysis, the outcome has a binary component (the occurrence or non-occurrence of the event) and a continuous component (the time to the occurrence of the event). With survival analyses, it is possible to compare groups in relation to the time until the emergence of several different outcomes, such as death and disease progression. As the patients are followed for long periods, this analysis is usually applied to chronic diseases and to assess outcomes that take time to occur, such as mortality [15]. It is important that, despite the name 'survival', survival analysis is not used only to assess the

time until death. Any binary event of interest can be explored: time to disease recurrence, time to surgery, time to failure and so on.

Censoring is a fundamental concept in survival analysis. In studies with long follow-up periods, some patients may not be followed until the occurrence of the event of interest, that is, they will have an incomplete follow-up time. Censorships can occur for several reasons, such as: the conclusion of the study before the patient has experienced the event, the patient is withdrawn from the study (for example, the patient must discontinue the study before the event of interest because of an adverse event) or loss to follow-up (it is not possible to contact the patient anymore for any reason, for example, change of city). If a person does not experience the event during the follow-up time, they will inevitably be censored [16].

In the following topics, two functions that are essential for survival analysis are described: the survival function and the hazard function. Later, some of the most common survival analysis methods are described.

The survival function

The survival function s(t) is a chronological, non-negative function, which starts at a certain moment in a well-defined time where all patients under study are alive (100%). As time goes by, patients die and consequently the survival of the population decreases. This function gives the probability of a patient surviving beyond a specified time [17,18]. The survival function s(t) is mathematically expressed by the Equation 13:

ALEXANDRE F. COBRE
LETICIA P. LEONART
ROBERTO PONTAROLO

$$s(t) = P(T > t) \text{ (Equation 13)}$$

Where 'T' is a random variable corresponding to a person time to the event of interest, 't' is any value that time can assume in the analysis and 'P' refers to probability. Therefore, the survival function allows to calculate the probability that an event has not happened in a chosen time, which can be done comparing the probability of 'T' being greater than 't'.

The hazard function

The hazard function h(t) shows the instantaneous probability of an individual experiencing the event of interest given the time that has passed without the event. Therefore, the function h(t) describes how the event rate varies depending on the time of patient follow-up [17,18]. Contrary to the survival function, the hazard function focus on the occurrence of the event. The h(t) is expressed mathematically by Equation 14:

$$h(t) = \lim_{\Delta t \to 0} = \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t} \text{ (Equation 14)}$$

Where 'T' is a random variable corresponding to a person time to the event of interest, 't' is any value that time can assume in the analysis and 'P' refers to probability. Figure 1 illustrates the hazard function distribution of a cohort of women with vulvar squamous cell carcinoma followed for 20 years

(January 1, 2000 - December 31, 2019), in which one group underwent surgery and the other did not. It is possible to see that at all times of follow-up, patients who did not receive the surgery had a higher hazard of experiencing the event of interest (death from cancer) than the patients who underwent the surgery.

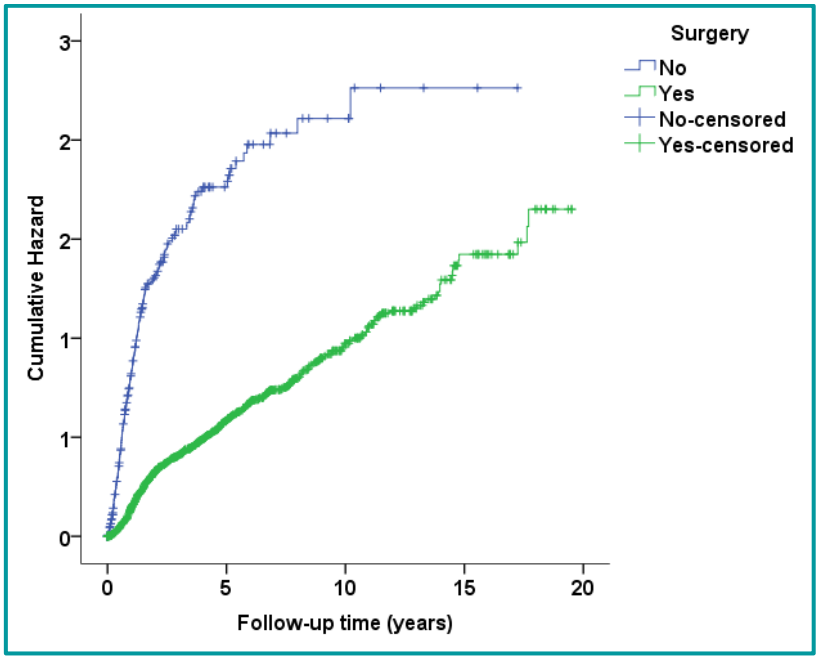


Figure 1. Hazard function distribution for death from cancer according to surgery status in patients with vulvar squamous cell carcinoma, São Paulo, Brazil (2000-2019)

Life table

At population level, the life table is used to estimate the life expectancy. It shows the probability of a person dying before their next birthday considering their current age. Beyond that, the life table method can also be applied to specific research questions. In this method, the accumulated

probability of survival is calculated at fixed intervals previously stipulated by the researcher. The number of patients at risk corresponds to the number of patients alive at the beginning of each interval (Table 3). The cumulative probability of survival for a given period depends on the probability of survival for the previous period. For example, in Figure 1, the 10-year survival probability of a patient who has undergone surgery is equal to the 10-year survival ratio of this patient (86% probability) multiplied by the cumulative survival ratio of the previous 5-year period (13% probability). In other words, the probability of survival at 10 years is equal: 0.86x0.13 = 0.11. Thus, the probability of a patient's survival in a certain period of time has a conditional character. It is worth mentioning that in this method censorship is not used in the calculation of accumulated probability of survival, for this reason, the number of patients at risk is different from the number of patients who started the interval, because the number [19].

Table 3. Life table for patients with vulvar squamous cell carcinoma survival according to surgery status, São Paulo, Brazil (2000-2019)

N. events	Interval start time (years)	N. enter interval	N. exposed to risk	N. deaths	Proportion of deaths	Proportion surviving	Cumulat. proportion survival
N. surgeries	0	453	410.5	333	0.81	0.19	0.19
	5	35	26.5	8	0.30	0.70	0.13
	10	10	7.0	1	0.14	0.86	0.11
	15	3	2.0	1	0.50	0.50	0.06
Surgeries	0	1457	1241.5	550	0.44	0.56	0.56
	5	476	372.0	112	0.30	0.70	0.39
	10	156	112.0	34	0.30	0.70	0.27
	15	34	18.5	3	0.16	0.84	0.23

ALEXANDRE F. COBRE
LETICIA P. LEONART
ROBERTO PONTAROLO

Kaplan-Meier

The Kaplan-Meier (KM) method, also known as product limit method, is non-parametric; therefore, it does not require any assumption about normality. Different from the life table, it considers censoring. In addition, KM allows to compare survival curves between different groups [20]. Below, an example to illustrate the KM method is presented.

In Table 4, a hypothetical cohort of 100 patients is presented. At study begin (0 year), no patient has experienced the event, so the survival is 1 (100%). Therefore, 100 patients continue in the analysis (i.e., are at risk of experiencing the event). At year 1, 5 patients have experienced the event and 1 patient was censored. Thus, the survival probability is 0.95 (95%). For the next year, we must discount from the 'at risk' column the number of patients that experienced the event (n=5) and the number of censored patients (n=1). Therefore, for the analysis at year 2, the denominator will be 94, and as 10 patients experienced the event, the survival probability is 85%. The same logic applies for all time intervals until the end of analysis.

Table 4. The Kaplan-Meier method

Interval beginning (years)	Patients at risk	Patients with event	Censored patients	Survival probability
0	100	0	0	1
1	100	5	1	$1 \times \frac{95}{100} = 0.95$
2	94	10	2	$0.95 \times \frac{84}{94} = 0.85$
…	…	…	…	

ALEXANDRE F. COBRE
LETICIA P. LEONART
ROBERTO PONTAROLO

Methods for comparing Kaplan-Meier survival curves

The most common used statistical test to compare KM curves is the chi-square test log-rank. However, the log-rank can lose statistical power as the time passes and the population at risk decreases. To solve this problem, some variations of the log-rank test can be used, such as the Wilcoxon (Breslow) and Tarone-Ware tests. The log-rank test has greater statistical power for comparison of curves at the beginning to the end of follow-up; Wilcoxon test is more powerful for statistical comparison of curves at the beginning of follow-up whereas the Tarone-Ware test is more powerful for comparison of curves at the middle of follow-up. If there are significant differences in all of these tests, it is reliable to say that there are significant differences in survival between the groups throughout the follow-up period [21–23].

Figure 2 shows the survival Kaplan-Meier curves of women who underwent and who did not undergo surgery for vulvar squamous cell carcinoma. From the graph it is possible to see that surgery provided gains in survival at the beginning, in the middle and at the end of the follow-up (log rank, $p < 0.001$; Breslow, $p < 0.001$; Tarone-Ware, $p < 0.001$).

ALEXANDRE F. COBRE
LETICIA P. LEONART
ROBERTO PONTAROLO

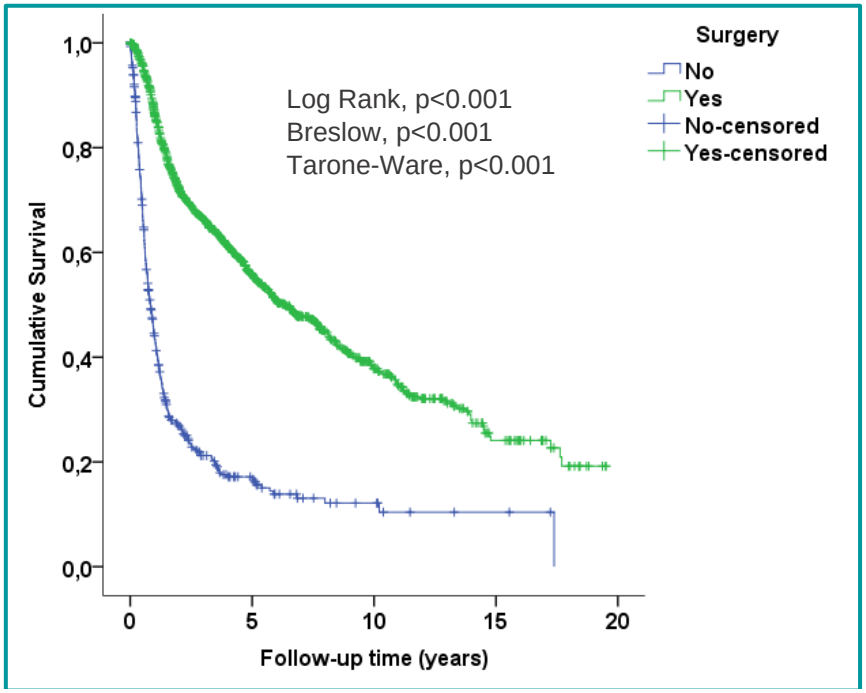


Figure 2. Kaplan-Meier curves for survival according to surgery status of patients with vulvar squamous cell carcinoma, São Paulo, Brazil (2000-2019)

Cox proportional hazards model

The Cox proportional hazards model has the advantage over the Kaplan-Meier method of adding covariates to the analysis. Therefore, it allows to adjust the model for confounding, which is not uncommon in observational studies. Besides that, it is possible to study how and if the covariates interact with each other and decide among the models performed which one gives the results with greater precision. For example, the Cox proportional hazards model is useful in clinical oncology when the objective is to investigate prognostic factors, since it evaluates the joint effect of the covariates on the survival of cancer patients. As this model is more mathematically advanced than the KM method, it cannot be

easily manually performed. Fortunately, common statistic software such as STATA and R are all prepared to perform this analysis. The Cox proportional hazards model is represented by a hazard function, described as follows (Equation 15):

$$h(t) = h_0(t) \times exp(\beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n) \quad (Equation\ 15)$$

Where 't' is the survival time, 'h0' is the baseline hazard, 'X' is the covariate and 'β' is the coefficient that measures the impact of the covariate. The hazard ratio (HR) corresponds to 'expβ'. If the HR is greater than 1.00, this means that as the value of the covariate increases, the survival decreases, i.e., the covariate is positively associated with the probability of the event of interest to occur.

For the use of the Cox risk-proportional regression model, certain prerequisites must be met: the groups must be independent, and the hazards must be proportional throughout the follow-up period. As it is possible to see in the formula, the HR is independent of time (HR = expβ). The simplest method for diagnosing the hazards proportionality is by visual inspection of curves in the graph: they should be proportional and do not cross; otherwise, the proportionality of hazards is violated and other model should be used, such as the variation of the Cox proportional hazards model that allows the inclusion of time-dependent covariates [35].

ALEXANDRE F. COBRE
LETICIA P. LEONART
ROBERTO PONTAROLO

Meta-analysis

A systematic literature review aims to collect all study data from clinical studies related to a particular scientific question. The data extracted from the clinical studies will then be synthesized. If possible (i.e. there is reasonable homogeneity between the studies), a statistical synthesis can be performed, which is called meta-analysis. In order to obtain reliable results, the meta-analysis is usually conducted with randomized controlled clinical trials, as this type of study usually presents a high level of evidence. However, other types of studies, such as observational cohorts, can also be pooled together in meta-analysis. The main objective of a meta-analysis is to improve the analytic statistical power, as the number of patients at risk and number of events increases when the studies are pooled together. Therefore, it is possible to say that meta-analysis aims to mimic a mega trial [17,27].

Meta-analysis must follow pre-defined mathematical criteria. Pooling the studies without these criteria can return misleading results. In a pairwise meta-analysis, each study has its own intervention and control group. Therefore, first it is necessary to estimate the effect measure of each individual study, and only then calculate the global estimate. This way, the randomization process inherent to each clinical trial is preserved [24].

An example of a meta-analysis is shown in Figure 3. This analysis compared fluoxetine 60 mg per day versus placebo in overweight or obese patients, and the outcome was weight loss at the end of trial [25]. The outcome is continuous (weight), so the chosen effect measure was mean difference (MD).

ALEXANDRE F. COBRE
LETICIA P. LEONART
ROBERTO PONTAROLO

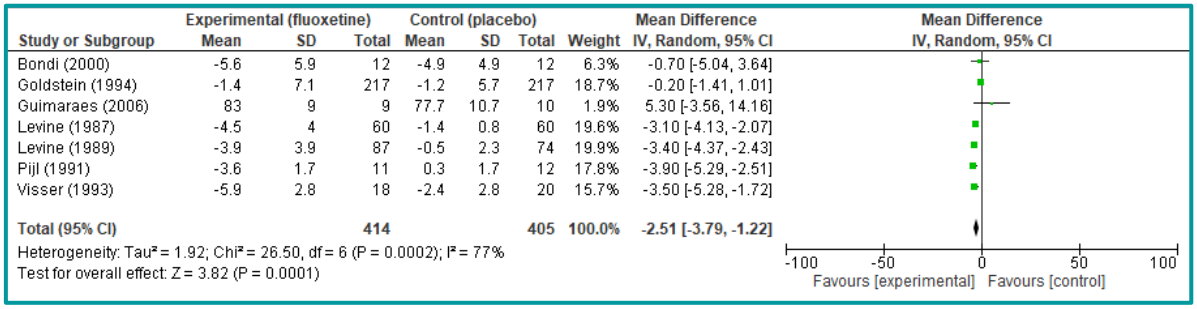


Figure 3. Meta-analysis of fluoxetine 60 mg per day versus placebo in overweight or obese patients (outcome weight loss at the end of trial)

In the first column, the identification of the studies included in the meta-analysis is shown (first author name and year of publication). In the following columns to the right, the number of patients included in each trial for the groups fluoxetine and placebo are presented (N) as well as the mean value of weight loss with the correspondent standard deviation for each group. In the last column, the mean difference and the associated 95% confidence interval (CI) of each individual study is presented, and in the last row of this column, the global effect of all studies is shown. Each study receives a different weight, as can be seen in the penultimate column. The allocation of weight depends on the method applied, which will be discussed further up. The graph illustration of the meta-analysis is shown as a forest plot.

The effect measure and 95% CI of each study is represented by a small green square and a black line, respectively. The global effect of all studies pooled together is presented as a black diamond. The extremities of the diamond represent the 95% CI. In this example, it possible to see that the diamond does not touch the vertical line, which represents the null value,

meaning that there is significant statistical difference between the two groups, in this case, favoring fluoxetine. This can also be confirmed by checking the numerical result: MD -2.51 (95% CI -3.79, -1.22). The 95% CI does not cross the null value (0).

Fixed and random effects models

One of the main points that differentiates the meta-analysis from a simple pooled analysis is the weight attributed to each of the studies. Two statistical models can be used for weight assignment: the fixed effect model and the random effects model.

In the fixed effect model, it is assumed that there is only a single effect that is common to all studies, and any variation observed between individual studies is due to sampling error. In practice, this is a rare scenario, which should restrict the application of this type of model. In the random effects model, the true effect can vary among the studies, which would be a more realistic scenario. Thus, there is more than one true effect, and the meta-analysis provides an average of those effects. In this model, in addition to the sampling error (intra-study variance), there is also the variance between the different studies (Tau2), also known as heterogeneity, which can be estimated using the DerSimonian and Laird method [24,26,27].

As mentioned, the weight attributed to each study depends on if the model is considering fixed effect or random effects. In the fist case, only the intra-study variance is considered. In the second case, the Tau2 is also inserted into the analysis. Thus, in the fixed effect model, studies with a larger sample

size tend to receive a greater weight, since the global effect is unique and more accurate in larger studies. On the other hand, in the random effects model it is assumed that there is more than one true effect. Thus, the distribution of weights tends to be more balanced. Consequently, since studies with a smaller sample size (more inaccurate) are given greater weight in the random effects model, the confidence intervals will be wider. The only exception is when Tau2 is 0. In this case, both models will provide the same result [24,28].

Heterogeneity

To assess whether there is statistically significant heterogeneity between the studies (Tau2), a X2 (chi-square) test considering the null hypothesis that all studies have the same effect estimate should be conducted. If the value obtained is less than the defined α (e.g. p = 0.1), the null hypothesis is rejected and it can be concluded that there is heterogeneity between the studies. It is important to note that a $p>\alpha$ does not necessarily mean that the studies are homogeneous, since the lack of significance may be due to the lack of statistical power [24].

A common way the assess the magnitude of the heterogeneity is through the I2 statistics, which can vary from 0 to 100%. The I2 is related to the overlapping of the confidence intervals of the studies. Thus, this estimate can be seen as a measure of inconsistency between studies. It is important to mention that studies with low precision can result in a low I2 value due to the overlap of excessively wide confidence intervals. Therefore, in these cases, the value of I2 should be interpreted with caution [24,29].

ALEXANDRE F. COBRE
LETICIA P. LEONART
ROBERTO PONTAROLO

Sensitivity analysis

Although meta-analysis plays an important role in the synthesis of evidence for scientific decision-making, some of these decisions are unclear because the included studies themselves did not present the necessary information, or due to poorly defined eligibility criteria, or even the lack of consensus on the best statistical method to be used for a specific problem. All these factors hinder the reproducibility of the meta-analysis, making it necessary to carry out a sensitivity analysis. In the sensitivity analysis, the way of analyzing the data is varied, in order to understand the impact of this change on the results. One way is to include only studies with specific characteristics, (e.g. group of patients, article' publication date, country). If a change in the results is detected, and if more homogeneous results are obtained with the new form of analysis, the authors should present the results in a subgroup analysis [30,31].

Meta-regression

The regression method is commonly used in primary studies to assess the association between one or many independent variables with the outcome variable. In meta-analysis a similar method, called meta-regression, can be employed. It allows evaluating the effect of multiple factors on heterogeneity, respecting a minimum limit of ten studies for conducting the analysis [32].

ALEXANDRE F. COBRE
LETICIA P. LEONART
ROBERTO PONTAROLO

Meta-analytic methods

The most used methods are the inverse of variance, Mantel-Haenszel and Peto. The inverse variance method can be used for both continuous and dichotomic data, and its name describes how it works: the weight given to each study is the inverse of the variance of the effect estimate. It can be performed using both fixed or random effects measure, and the difference, as previously mentioned, relies in the addition of the heterogeneity into the analysis when using the random effects model. When there are few events or a small sample size, the inverse of variance may not provide accurate estimates, and therefore, alternative methods using a different weighting scheme can be used, such as the Mantel-Haenzel. When events are very rare, the Peto's method is recommended, as corrections for zero cells counts are not necessary [33].

Publication bias

The data used to conduct the meta-analysis may be affected by publication bias. Publication bias is when published results are different from reality. One study showed that of all clinical trials registered on the ClinicalTrial.gov registration platform, less than 70% of those are published. The non-publication of researches may be due to lack of interest on the part of the scientific editors, not being interested in publishing negative results (without statistical significance), or due to authors not submitting the study due to unfavorable results. The techniques used to detect these types of bias are funnel graphic pole and statistical tests. Generally, these methods are applied to meta-analyses with ten or more studies, and are based on questions of precision

and estimates. Publication bias can also be estimated in the study registration database, looking for those studies that were not published [34,35].

Recommendations for improving meta-analyses reproducibility

- **Recruit experience:** for guidance on how to do a literature search and calculation of reproducible effect magnitude, before starting any literature search consult a librarian and before starting coding effect magnitudes consult a statistician
- **Facilitate cumulative scientific knowledge through future-proof meta-analyses**: Report all data used in the meta-analysis. Report all meta-analysis results data for each data point. To avoid confusion, report all relevant information from studies included in the meta-analysis. In case of analysis by subgroups, cite the original article justifying the chosen classification, and specifying any subjective decision
- **Facilitate quality control**: Specify which author performed the extraction and encoding of the effect size data, knowing that it is advisable that extraction be performed independently with at least two researchers. Clearly inform the calculations adopted to estimate the magnitude of effect, and what assumptions were made for each effect size, obtained from the articles included
- **Conduct the study following the guidelines of an official guide**, for example PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses). The reporting guidelines ask authors to provide a completed checklist, such as supplementary material, before and after publication of the study

ALEXANDRE F. COBRE
LETICIA P. LEONART
ROBERTO PONTAROLO

- **Pre-registration:** It is recommended that authors pre-register their systematic review protocol and meta-analysis, in order to distinguish between exploratory and confirmatory analysis
- **Facilitate reproducibility:** Provide a spreadsheet containing the meta-analysis data, to be easily analyzed in any statistical program. If possible, inform the software used to conduct the meta-analysis[35].

Conclusion

In this chapter, basic statistical concepts for those interested in health technology assessment were presented. First, the most common effect measures found in clinical studies were presented for continuous, dichotomous and count data. Following, the main concepts on survival analyses were described, including the survival and hazard functions, and the most frequently used methods for survival analyses, which include life table, the Kaplan-Meier method, and the Cox proportional hazards model. Finally, the basic aspects of pairwise meta-analyses were presented, including how to interpret a forest plot, the difference between fixed and random effects model and which are the most employed analytic methods.

ALEXANDRE F. COBRE
LETICIA P. LEONART
ROBERTO PONTAROLO

References

[1] Panter-Brick C. Health, risk, and resilience: Interdisciplinary concepts and applications. Annu Rev Anthropol 2014;43:431–48.
[2] Andrade C. Understanding relative risk, odds ratio, and related terms: As simple as it can get. J Clin Psychiatry 2015;76:e857–61.
[3] Furcada JM, Patino CM, Ferreira JC. Estimating risk in clinical studies: Odds ratio and risk ratio. J Bras Pneumol 2020;46:1.
[4] Gogtay NJ, Thatte UM. Number needed to treat. J Assoc Physicians India 2017;65:90–4.
[5] Stare J, Maucort-Boulch D. Odds ratio, hazard ratio and relative risk. Metod Zv 2016;13:59.
[6] Kim H-Y. Statistical notes for clinical researchers: Risk difference, risk ratio, and odds ratio. Restor Dent Endod 2017;42:72.
[7] Higgins JPT, Li T DJ (editors). Choosing effect measures and computing estimates of effect. Cochrane Handb. Syst. Rev. Interv. version 6.1 (updated Sept. 2020), 2020.
[8] Mendes D, Alves C, Batel-Marques F. Number needed to treat (NNT) in clinical literature: An appraisal. BMC Med 2017;15:1–14.
[9] Fletcher J. Continuous variables. BMJ 2008;337.
[10] Guyatt GH, Thorlund K, Oxman AD, Walter SD, Patrick D, Furukawa TA, et al. GRADE guidelines: 13. Preparing Summary of Findings tables and evidence profiles - Continuous outcomes. J Clin Epidemiol 2013;66:173–83.
[11] Higgins JPT, Whitehead A, Turner RM, Omar RZ, Thompson SG. Meta-analysis of continuous outcome data from individual patients. Stat Med 2001;20:2219–41.
[12] Faraone S V. Interpreting estimates of treatment effects: Implications for managed care. P T 2008;33.
[13] Shmueli G, Minka TP, Kadane JB, Borle S, Boatwright P. A useful distribution for fitting discrete data: Revival of the Conway-Maxwell-Poisson distribution. J R Stat Soc Ser C Appl Stat 2005;54:127–42.
[14] Cobre AF, Pedro C de AA, Fachi MM, Vilhena RO, Marson BM, Nicobue V, et al. Five-year survival analysis and predictors of death in HIV-positive serology patients attending the Military Hospital of Nampula, Mozambique. AIDS Care - Psychol Socio-Medical Asp AIDS/HIV 2020;32:1379–87.
[15] Clark TG, Bradburn MJ, Love SB, Altman DG. Survival Analysis Part I: Basic concepts and first analyses. Br J Cancer 2003;89:232–8.
[16] Kleinbam DG, Klein M. Survival Analysis - A Self-Learning Text. 3rd ed. Springer; 2012.
[17] Hermeto RT, Gomes DG. Dissertação de mestrado: análise de Sobrevivência na Modelagem do tempo de Vida de redes de sensores sem Fio. 2014.
[18] Alencar AP, Alessandra C Goulart GTS. Análise de sobrevivência 2019.
[19] Bustamante-Teixeira MT, Faerstein E, Latorre M do R. Técnicas de análise de sobrevida. Cad Saude Publica 2002;18:579–94.
[20] Kaplan EL, Meier P. Nonparametric Estimation from Incomplete Observations. J Am Stat Assoc 1958;53:457–81.
[21] Ross DJ, Mohsenifar Z, Koerner SK. Survival characteristics after neodymium: YAG laser photoresection in advanced stage lung cancer. Chest 1990;98:581–5.
[22] MIOT HA. Análise de sobrevivência em estudos clínicos e experimentais. J Vasc Bras 2017.

[23] Cox DR. Analysis of survival data.Chapman and Hall/CRC 2018. https://doi.org/https://doi.org/10.1201/9781315137438.
[24] Borenstein M, Hedges LV, Higgins JP RH. Introduction to Meta-Analysis. Chichester, UK: John Wiley & Sons, Ltd; 2009.
[25] Serralde-Zúñiga AE, Gonzalez Garay AG, Rodríguez-Carmona Y, Melendez G. Fluoxetine for adults who are overweight or obese. Cochrane Database Syst Rev 2019.
[26] Gurevicth J, Koricheva J, Nakagawa S, Stewart G. Meta-analysis and the science of research synthesis. Nature 2018;555:175–82.
[27] DerSimonian R, Laird N. Meta-analysis in clinical trials revisited. Contemp Clin Trials 2015;45:139–45.
[28] Brockwell SE, Gordon IR. A comparison of statistical methods for meta-analysis. Stat Med 2001;20:825–40.
[29] Crowther M, Lim W, Crowther MA. Systematic review and meta-analysis methodology. Blood 2010;116:3140–6.
[30] Crowther M, Lim W, Crowther MA. Systematic review and meta-analysis methodology. Blood 2010;116:3140–6.
[31] Copas JB, Shi JQ. A sensitivity analysis for publication bias in systematic reviews. Stat Methods Med Res 2001;10:251–65.
[32] Ahn E, Kang H. Introduction to Systematic Review and Meta-Analysis: A Health Care Perspective. Korean J Anesthesiol 2018;71:1–38.
[33] Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al. Cochrane Handbook for Systematic Reviews of Interventions version 6.1 (updated September 2020) 2020.
[34] Lin L, Chu H. Quantifying publication bias in meta-analysis. Biometrics 2018;74:785–94.
[35] Lakens D, Hilgard J, Staaks J. On the reproducibility of meta-analyses: Six practical recommendations. BMC Psychol 2016;4:1–10.

BASIC METHODS AND TOOLS FOR STATISTICAL SYNTHESIS

ALEXANDRE F. COBRE
LETICIA P. LEONART
ROBERTO PONTAROLO

PHARMACOECONOMIC ANALYSES: MAIN CONCEPTS

IGNACIO AZNAR-LOU
VASCO M. PONTINHA

Abstract

In the last years, the growing interest of healthcare professionals, policy-makers, and other stakeholders in enlarging the role of economic evaluations was driven by several factors such as evidence-based healthcare culture, patient-centered actions, and quality-linked incentives, associated with an important increase of financial constraints and pressures on healthcare budgets. Pharmacoeconomics, as a branch of health economics, focuses on balancing the costs and benefits (i.e. consequences) of an intervention or technology towards the use of limited resources, aiming at maximizing value to patients, healthcare payers and society. These concepts are part of the Health Technology Assessment (HTA) process, that informs governmental players about medical, social, and economic implications of development, diffusion, and use of health technologies. This chapter aims to provide an overview of the important concepts in pharmacoeconomic analysis methods, including studies classification (e.g. budget-impact analysis, cost-minimization analysis, cost-effectiveness analysis, cost-utility analysis), types of costs and outcomes, and modelling approaches (e.g. decision trees or simulation models as the Markov model) and additionally discuss some recommendations for future studies.

Keywords: economic evaluation; costs; outcomes; pharmacoeconomics

PHARMACOECONOMIC ANALYSES: MAIN CONCEPTS

IGNACIO AZNAR-LOU
VASCO M. PONTINHA

Introduction

Health Technology Assessment (HTA) intend to provide a bridge between scientific research and decision-making processes, including setting priorities in healthcare or guiding the selection or incorporation of new treatments. In the past years, the growing awareness of the importance of HTA worldwide also highlighted the needed of using well-designed and standardized studies and tools to support these reports, especially considering the economic component. This is because studies of costs and related economic implications comprise a major group of methods used in HTA [1-3].

Pharmacoeconomics is a branch of health economics that usually focuses on balancing the costs and benefits of an intervention towards the use of limited resources, aiming at maximizing value to patients, healthcare payers and society. This is important as cost containments are currently common for the management of healthcare systems worldwide, yet the development of innovative and cheaper interventions is scarce. However, although most of the newer technologies are costlier than the existing ones, they also usually provide added benefits over previous interventions. In this scenario, decisions-makers (e.g. healthcare professionals, politicians, and other stakeholders) have to consider whether or not the new intervention is affordable and an efficient use of resources. Additionally, full pharmacoeconomic evaluations, defined as analyses that identify, measure and compare the outcomes (i.e. costs and consequences) among available interventions, are key studies to inform pricing and reimbursement decisions in several countries [4-7].

In recent years, there is a general trend on the increase in number of the pharmacoeconomic studies on different levels of complexity worldwide. These studies can involve attributes of either or both of primary data collection and

integrative methods. That is, cost data can be collected, for example, as part of randomized controlled trials and other clinical studies as well as administrative databases used in health care payment. Cost data from one or more sources often are combined with data from primary clinical studies, epidemiological studies, and other sources to conduct one (or more) economic evaluations (e.g. cost-minimization analysis, cost-benefit analysis, cost-effectiveness analysis, cost-utility analysis) that involve weighing health and economic impacts of health technology [8,9].

The suitability of any of this variety of approaches to economic analysis depends on the purpose of an assessment and the availability of data and other resources. Thus, this chapter aims to provide an overview of the main pharmacoeconomic analysis methods that should be used for the assessment of healthcare technologies, and additionally discuss some recommendations for future studies.

Guidelines for economic evaluations

Economic evaluations of healthcare interventions pose a particular challenge for conduction and reporting because substantial information must be conveyed to allow scrutiny of study findings. Globally, a number of countries have developed guidelines that describe the design and conduct of economic evaluations as part of HTA or pharmacoeconomic analysis for decision making [10-13].

A recent scoping review summarized the recommendations made on methods of economic evaluations by the national healthcare economic evaluation (HEE) guidelines. A total of 31 national HEE guidelines, published between 1997 and August 2020 were evaluated. Almost half of them (45%) targeted the

evaluation of pharmaceuticals. The nature of the guidelines was either mandatory (31%), recommendatory (42%), or voluntary (16%). There was a substantial consensus among the guidelines on several key principles, including the primarily or preferable type of economic evaluation (cost-utility analysis), time horizon of the analysis (long enough), health outcome measure (quality-adjusted life-years - QALY) and use of sensitivity analyses. The recommendations on study perspective, comparator, discount rate and type of costs to be included varied according to the country given the differences in the health systems and financing mechanisms, capacity of local researchers, and data availability [4].

The Professional Society for Health Economics and Outcomes Research (ISPOR), for instance, is responsible for creating guidelines in the pharmacoeconomic field to be used worldwide. The CHEERS (Consolidated Health Economic Evaluation Reporting Standards) statement is one of the most well-known checklists that can help during the process of performing and reporting an economic study. It provides examples and explanations for each of the 24 items and accompanying recommendations, with some specific recommendations for single study-based and model-based economic evaluations. The final recommendations are subdivided into six main categories: title and abstract, introduction, methods, results, discussion, and can be accessed through the website: **http://www.ispor.org/TaskForces/EconomicPubGuidelines.asp**

Overall, a well-designed pharmacoeconomic analysis involves 10 steps:

- Defining the problem
- Determining the study's perspective
- Determining the alternatives and outcomes

- Selecting the appropriate pharmacoeconomic method
- Placing monetary values on the outcomes
- Identifying study resources
- Establishing the probabilities of the outcomes
- Applying decision analysis
- Discounting costs or performing a sensitivity or incremental cost analysis
- Presenting the results, along with any limitations of the study

Key attributes in economic evaluations

An economic evaluation is defined as a comparative analysis of at least two health interventions used to assess both the costs and consequences of the different technologies in a given population, providing a decision framework. The two main components of this analysis are 'costs' (i.e. the monetary component of the economic analysis. It can be divided into direct, indirect, and intangible costs and 'outcomes') and 'outcomes' (i.e. also called 'benefits' or 'consequences', the outcomes are the expected healthcare or humanistic results from an intervention) [14-16].

The main inputs and definitions of pharmacoeconomic evaluations as a brief summary of the CHEERS statement are described below. These parameters should be considered in the same way as those from clinical trials (e.g. population, intervention, comparator, outcome and timing – PICOT). The population comprises the modelled population, sources of input data and assumptions for which must be clearly articulated so that its generalizability and applicability can be ascertained. The intervention is the technology of interest; all assumptions made about its use should be clearly described. The identification

and measurement of outcomes and costs will depend on the intervention characteristics and perspective adopted. The appropriate expression of the time horizon is important because interventions costs and benefits vary with time

- **Target population:** the group or subgroup of patients who will benefit from the health intervention
- **Target intervention:** main interventions being assessed in the economic evaluation such as drugs, vaccines, medical procedures, or services
- **Comparator:** other intervention or standard of care (current best practice), minimum practice, or no intervention that is being compared to the target intervention
- **Setting:** the context in which the intervention will occur
- **Perspective:** the different viewpoints from which outcomes and costs are being assessed (e.g., patient, provider, payer, society)
- **Time horizon:** the duration over which costs and outcomes are calculated in an economic analysis
- **Discounting:** cost analyses should account for the effect of the passage of time on the value of costs and outcomes. This method is used to account for individuals time preference (i.e. most individuals have a positive rate of time preference whereby benefits are preferred sooner rather than later, and costs incurred later rather than sooner)
- **Modelling:** decision analyses from economic evaluations can be operationalized through modeling processes such as decision trees or simulation models

- **Sensitivity analyses:** estimation of data stability; a means of representing uncertainty in the results of economic evaluations (e.g. one-way simple sensitivity analysis, multiway sensitivity analysis, threshold sensitivity analysis, probabilistic sensitivity analysis)

Additionally, the concept of 'opportunity cost' refers to the loss of potential benefits from other options when one option is chosen. This concept is based upon the idea that the scarcity of resources leads players to expend capital on one healthcare activity by sacrificing services elsewhere. Thus, understanding the potential missed opportunities foregone by choosing one technology over another allows for better decision-making [17].

Types of costs and costing methods

One of the firsts steps in any cost analysis is the identification of the various costs (monetary outcomes), that are usually classified into [8, 9]:

Direct costs that represent the value of all goods, services, and other resources consumed in providing health care or dealing with side effects or other current and future consequences of healthcare. They are paid directly to healthcare service (i.e. associated with patients' treatment). These costs can be additionally classified into fixed or variable (according to the changes in the volume of services provided) and medical or non-medical direct costs, depending on their nature. Medical costs include, for instance, staffing (e.g. physicians, nurses), consumables (e.g. drugs, treatments), consultations, exams, procedures, hospital and intensive care admissions, equipments and installations, ambulance services. Non-medical costs include extra expenses from treatments, travels costs and temporary residence.

Indirect costs, sometimes referred to 'productivity losses' represent the costs experienced by patients', family, or society, as the loss of earnings or productivity resulting from patients' illness. These include the costs of lost work due to absenteeism or early retirement, impaired productivity at work (sometimes known as 'presenteeism'), and lost or impaired leisure activity. Indirect costs also include the costs of premature mortality.

Intangible costs are attributed to the amount of suffering that occurs due to illness or healthcare intervention. This cost is usually difficult to measure in monetary terms yet is being increasingly included in utility assessments. They include costs related to pain, suffering, and grief.

The cost can be measured as cost/unit, cost/treatment, cost/person, cost/person/year, cost/case prevented, cost/life saved, cost/DALY (disability-adjusted life year), cost/QALY (quality-adjusted life year), cost/LYG (life years gained) [5, 6].

The validity of a cost-related study depends on the sources of the data for costs and outcomes. The costing methods generally fall on a spectrum between a bottom up, micro-costing and a top down, gross costing approach, each with trade-offs between accuracy, precision, and the burden research. The choice of the method determines the cost estimates. Commonly, hybrid approach is found to be appropriate under given feasibility restraints.

Additionally, increased attention is being given to collection of cost data in rigorous, prospective clinical studies. The closer integration of economic and clinical studies can promote more informed resource allocation for new technologies and generate reliable cost and outcomes data during the early part of a technology's lifecycle [18, 19]. There is also a growing interest in using

observational data to assess the safety, effectiveness, and cost-effectiveness of medical technologies. However, operational, technical, and methodological challenges may limit its widespread use. Common data models and federated data networks offer a potential solution to many of these problems. For instance, the open-source Observational and Medical Outcomes Partnerships (OMOP) common data model standardizes the structure, format, and terminologies of otherwise disparate datasets, enabling the execution of common analytical code across a federated data network in which only code and aggregate results are shared. The use of open-source and standardized analytics improves transparency and reduces coding errors, thereby increasing confidence in the results [20].

Assessments should also make clear whether average costs or marginal costs are being used in the analysis. Whereas average cost analysis considers the total (or absolute) costs and outcomes of an intervention, marginal cost analysis considers how outcomes change with changes in costs (e.g. relative to the standard of care or another comparator), which may provide more information about how to use resources efficiently. Marginal cost analysis may reveal that, beyond a certain level of spending, the additional benefits are no longer worth the additional costs [9, 21].

Outcome measures

The second component of any economic analysis is the outcome or consequence to be measured, that is defined as the expected benefits from an intervention. 'Benefit' measurement aims to be equally comprehensive by incorporating all of the impacts upon the patients' life that arise as a consequence of the use of a healthcare technology. The benefits derived from an intervention

might be measured in: (i) Natural units (e.g. years of life gained, events prevented, avoided medical procedures); (ii) Utility units that attempt to evaluate the quality of a state of health (and not just its quantity) or the satisfaction derived from moving from one state of health to another as a consequence of the application of an intervention. Such utility estimates are frequently informed by some measurement of 'quality of life' in different disease states [6, 22].

Healthcare related quality of life (HRQoL) measures attempt to incorporate into the analysis the physical, social, and emotional aspects of the patient's well-being, which are not directly measurable in clinical terms. One of the most common summaries of quality and quantity of life is the Quality Adjusted Life Year (QALY) measure. In order to generate QALYs, health utilities (or HRQoL weights) are needed. Utilities are preference weights, where preference can be equated with value or desirability. Utilities are measured on a cardinal scale of 0 to 1, where 0 indicates death and 1 indicates full health. States worse than death can also be accounted for, with such states taking a negative value [23]. The QALY is able to combine the effects of health interventions on mortality and morbidity into a single index thereby providing a 'common currency' to enable comparisons across different disease areas. QALYs are calculated simply by multiplying the duration of time spent in a given health state by the HRQoL weight (i.e. utility score) associated with that health state. For instance, if an individual is in a health state for 10 years, where the health state has an associated utility of 0.7, this would generate 7 undiscounted QALYs.

Many different methods have been proposed to value HRQoL based upon widely different techniques and value systems, that are broadly divided into direct or indirect methods (also called generic preference-based measures). Authors may also use imputed data from literature or expert opinion. The direct

methods that tend to be used most regularly for eliciting preferences include the visual analogue scale (VAS), the time trade-off (TTO) and the standard gamble (SG). The VAS (a form of rating scale) involves the use of a scale shown on a single line. The top of the scale indicates the 'best imaginable health', whereas the bottom of the scale indicates the 'worst imaginable health'. The TTO method presents individuals with two alternative scenarios and asks which they would prefer: the choice is between living for the rest of their life in an impaired health state or living in full health for a shorter period of time. Then, participants are asked how much time they would be willing to sacrifice to avoid an impaired health state. Finally, the SG involves an element of risk in the decisions faced by participants. The choice is between the certainty of remaining in a particular health state, or taking a gamble of either being in full health or risking death. Indirect methods involve the use of pre-scored generic preference-based measures (also called 'off-the-shelf' questionnaires or generic multi-attribute systems), which are routinely used in healthcare trials. In this context, health states are described using standardized generic utility questionnaires, which cover general aspects of health. The most commonly generic questionnaires are the EuroQol (EQ)-5D, the Short Form 6D (SF-6D) and the Health Utilities Index (HUI). The measures of these tools differ in terms of aspects such as the dimensions of health (attributes) that are included, the number and description of levels defined for each dimension, and the population on which the preferences are based. The instruments also differ in terms of the valuation method: the TTO was used to value the EQ-5D, whereas the SF-6D and HUI were grounded on SG. Once completed, the questionnaires generate a score using an algorithm based on values that have been obtained from a sample of the general public [23, 24].

However, many controversies in using QALY approach exist, especially considering its limitations in terms of capturing health benefits its blindness towards equity concerns, the underlying theoretical assumptions, and the most appropriate generic preference-based measure of utility. Additionally, there is growing debate relating to whether a QALY is the same regardless of who accrues it, and also the issue as to who should value health states. In this context, other approaches such as DALY (disability-adjusted life year) can be used. Another approach that can be used to assess benefits in economic evaluations is the willingness to pay (WTP) analysis, which main feature is to value health outcomes in monetary terms. In this analysis, individuals are asked the maximum they are willing to pay ('sacrifice') to achieve a given benefit of a intervention. For this, players have to consider all the important attributes of the technology/service under evaluation. Using WTP to estimate the benefits of healthcare allows individuals to value both health outcomes, non-health outcomes and process attributes. WTP can be estimated using techniques such as open-ended, bidding, payment card and closed-ended [2, 26].

Perspective, time horizon and modelling

The economic evaluations have another important component, called 'perspective', that represents the point of view adopted when deciding which types of costs and health benefits are to be included in the analysis. Typical viewpoints are those of the patient, health insurance companies and employer (e.g. payers), hospital/clinic or healthcare professionals (e.g. providers), healthcare systems or society [5, 27]. The most comprehensive perspective is societal as it includes the perspectives of all stakeholders in healthcare, aiming at reflecting a full range of social opportunity costs associated with different

interventions. In particular, this includes productivity losses arising from patients' inability to work, and changes in these losses associated with a new technology. The UK NICE (The National Institute for Health and Care Excellence) recommends that any pharmacoeconomic analyses submitted to the regulators should include a societal perspective – called the 'reference case'. Other perspectives may be also evaluated [6].

Interpretation of cost analyses must consider that the time horizon (or time-frame) of a study is likely to affect the findings regarding the relative magnitudes of costs and outcomes, as they usually do not accrue in steady streams over time. The choice of time horizon is an important decision for economic modelling and depends on the nature of the disease and intervention under consideration and the objectives of the analysis. They should be long enough to capture streams of health and economic outcomes (including significant intended and unintended ones). These could encompass a disease episode, patient life, or even multiple generations of life. For instance, longer time horizons are recommended for chronic conditions associated with on-going medical management, rather than a cure. A shorter time horizon may be appropriate for some acute conditions, for which long-term consequences are less important. HTA agencies usually recommend a lifetime horizon, although it may be useful in sensitivity analysis to test out intermediate time-horizons (e.g. 5 to 10 years), for which there may be more robust data available. Additionally, it is important to consider that the use of long-term time horizon is likely to involve extrapolating the cohort (group of patients) experience into the future and making assumptions about the continued efficacy of interventions and costs of care, as well as discounting of future inputs [28, 29].

The discounting is a method that accounts for individuals time preference, considering that costs and outcomes can occur at different times when using a technology. Most individuals have a positive rate of time preference whereby benefits are preferred sooner, and costs incurred later. That is to say, costs and outcomes that occur in the future usually have less present value than costs and outcomes realized today. In economic evaluations, the discount rates of costs and outcomes is performed if the costs and effectiveness outcomes are considered beyond 12-month time periods. The present value of money, as well as better health, is higher than future costs and outcomes. Currently, the NICE recommends a discount rate of 3.5%, but overall rates of 3% or 5% per year can be used. Cost analyses should also correct for the effects of inflation (which is different from the time preference accounted for by discounting), such as when cost or cost-effectiveness for one year is compared to another year [30].

Modeling is defined as the reproduction of events and possible consequences due to alternative policy options at the cohort or individual levels using mathematical and statistical frameworks. Models of costs and benefits are paramount in economic evaluations that are part of decision-making processes for incorporation and financing of technologies of healthcare systems. These decision analyses can be operationalized through decision trees or simulation models [21, 31, 32]. To address uncertainty involved in estimations of costs, outcomes, and other variables used in a decision-model analysis, sensitivity analysis should be performed. This type of analysis may find that including variables such as indirect costs in the model or using a reasonable higher discount rate, changes the cost-effectiveness of one intervention compared to another. The four main types of sensitivity analyses are: one-way simple sensitivity analysis,

multiway sensitivity analysis, threshold sensitivity analysis, probabilistic sensitivity analysis [21, 33].

Type of economic analyses

The main types of economic analyses are depicted in Table 1:

Table 1. Pharmacoeconomic analysis according to costs x outcomes

Type of analysis	Valuation of costs*	Valuation of health outcomes	Calculation
Cost of illness analysis (COI)	$	None	At disease level
Budget-impact analysis (BIA)	$	None or various**	Compare interventions
Cost-minimization analysis (CMA)	$	Assume same	Compare interventions
Cost-consequence analysis (CCA)	$	Natural units	Compare interventions
Cost-effectiveness analysis (CEA)	$	Natural units	Cost-benefit ratio
Cost-utility analysis (CUA)	$	Utility units	Cost-benefit ratio
Cost-benefit analysis (CBA)	$	$	Ratio or net costs and benefits

*Any currency
**It can determine the impact of a technology on a designated nonfixed budget or it can maximize some health outcome within a designated fixed budget
Adapted from the US National Information Center on Health Services Research and Health Care Technology (NICHSR) **www.nlm.nhi.gov/nichsr/hta101/ta10107.html**

Cost-of-illness analysis (COI)

Determination of the economic impact (burden) of a disease or condition on a given population or region/country including the associated treatments costs [8]. This analysis can be useful to prioritize between diseases. However, it is not sufficient to ground efficient healthcare allocation for coverage and reimbursement decisions of a particular intervention (e.g. a high-cost burden

does not mean that treatments are available to reduce this burden) [34, 35]. In this case, budget-impact analysis (BIA) is preferable, as affordability is also important for short-run economic purposes.

Cost-minimization analysis (CMA)

Aims to determine the least costly among alternative technologies that are assumed to produce equivalent healthcare outcomes (~same efficacy/safety profiles). The evidence on the equivalence must be referenced by the author conducting the study and should have been done prior to the cost-minimization analysis [6, 8].

Cost-effectiveness analysis (CEA)

One of the most used economic evaluation worldwide, it is defined by ISPOR as a comparison of interventions regarding costs in monetary units and outcomes expressed in quantitative non-monetary health units (e.g. reduced mortality or morbidity, symptom-free days gained, cases prevented, life years gained) [22, 36]. The CEA usually considers a long-term or lifetime time horizon, apply discounting rates and the inputs consider the population average. If there are just two alternative technologies being assessed by the CEA, their difference in cost (incremental cost) is compared to their difference in outcomes (incremental effect) by dividing the former by the latter. This ratio is known as the incremental cost-effectiveness ratio (ICER) as showed below (Equation 1). If there are more than two alternatives, they are compared on a systematic pairwise basis using their ICERs [22, 30].

IGNACIO AZNAR-LOU
VASCO M. PONTINHA

$$Incremental\ cost\ effectiveness\ ratio\ (ICER) = \frac{Cost_A - Cost_B}{Effect_A - Effect_B} \quad (Equation\ 1)$$

Based on the ICER results, a cost-effectiveness plane diagram can be built to illustrate the different situations during a decision analysis and also used to demonstrate the meaning and use of the ceiling ratio, where it is often referred to as demonstrating cost-effectiveness acceptability (Figure 1).

In this diagram, ICERs are presented graphically as a combination of the costs and the effects of a health intervention, compared to an alternative. Costs are conventionally placed on the top-bottom (north-south) axis and effects on the right-left (east-west) axis. In both cases, these effects can be negative, zero or positive. If the intervention lies in the top left quadrant (NW), as demonstrated by point A (Figure 1), the costs of the intervention are higher than its alternative, and its benefits are lower. As this is unambiguously worse than its alternatives, the intervention is considered 'dominated' and should be rejected (unacceptable). Similarly, in the bottom right quadrant (SE), point B refers to a technology with lower costs and higher benefits than its alternatives, so the new treatment 'dominates' the previous one and should always be accepted. For both the top right (NE) and bottom left (SW) quadrants – represented by points C and D, respectively, neither alternative dominates. In these cases, ICERs should be calculated (measured as the slope of the line from the origin to the point). For point C, the ICER represents the cost per unit of effect gained, while for point D it refers to a cost saving per unit of effect lost.

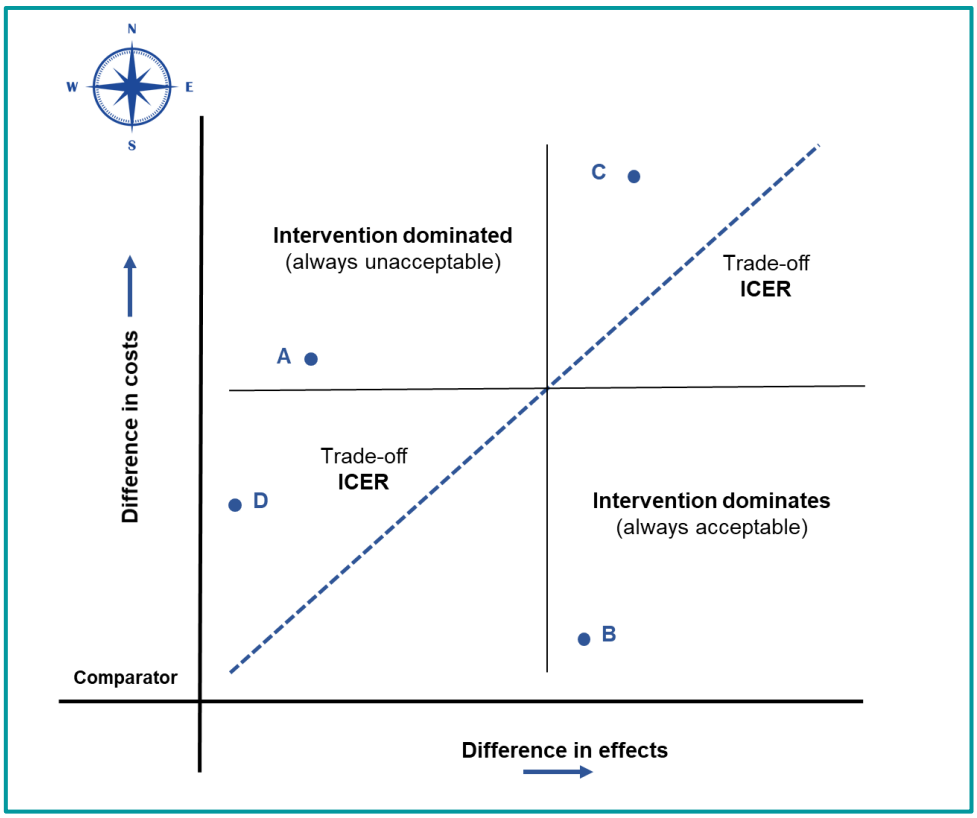


Figure 1. Summary of a cost-effectiveness plan diagram

Cost-utility analysis (CUA)

A form of CEA that compares costs in monetary units with health outcomes regarding their utility and mortality, which is expressed in QALYs. This is the preferred type of economic evaluation as it allows the use of the same health outcome for all interventions and diseases, and thus to help decision-makers to allocate resources efficiently [22, 24, 36]. Similar to CEA, the ICER in the CUA is calculated considering a ratio of costs over benefits, in this case, over QALYs as follow (Equation 2):

$$Incremental\ cost\ utility\ ratio = \frac{Cost_A - Cost_B}{Utility_A - Utility_B}\ (Equation\ 2)$$

Cost-consequence analysis (CCA)

A form of CEA that presents costs and health outcomes in discrete categories, without aggregating or placing weights on the costs and health outcomes [5, 6].

Cost-benefit analysis (CBA)

Compares costs and health benefits (and risks), all of which are quantified in common monetary units as follow (Equations 3 and 4) [5, 6].

$$Cost - benefit\ ratio\ approach = \frac{Cost_A - Cost_B}{Benefit_A - Benefit_B}\ (Equation\ 3)$$

$$Cost - benefit\ net\ approach = (Cost_A - Cost_B) - (Benefit_A - Benefit_B)(Equation\ 4)$$

Budget-impact analysis (BIA)

This type of analysis estimates the impact of implementing or adopting a new technology or technology-related policy or service on a designated healthcare budget. This method is able to assesses the affordability of a healthcare intervention if the intervention is used within an environment as compared to not used within that same environment [1, 37] (see Figure 2). The BIA is usually performed from the payer perspective (model inputs), considers the size of the population, and has a short-term and steady time horizon (for instance 3 to 5 years). Here, the only model output is the cost.

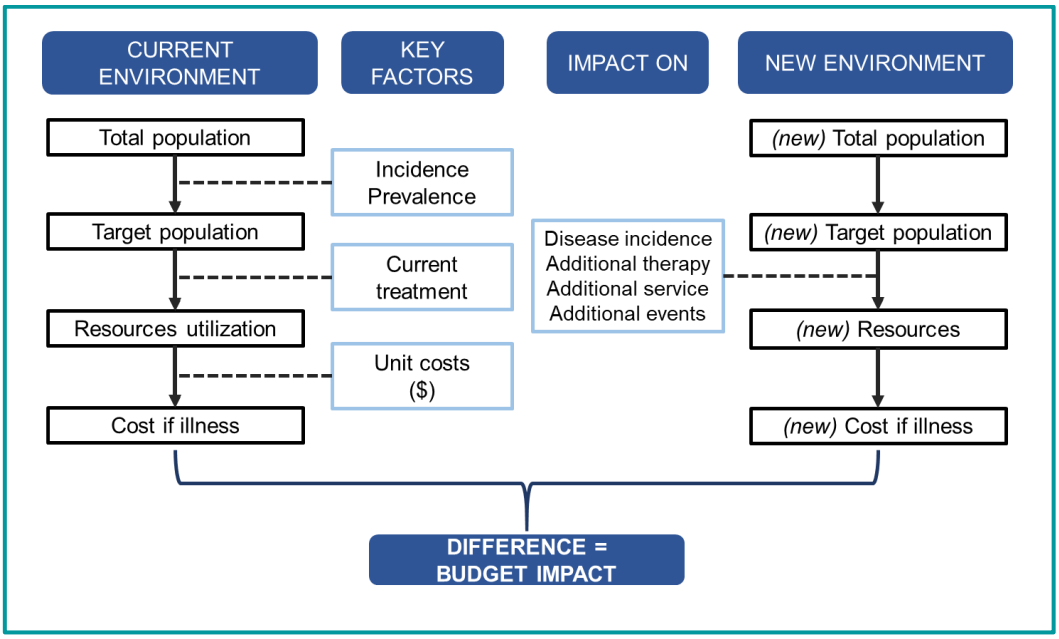


Figure 2. Basic framework of a budget impact analysis

Decision analyses

In economic evaluations, the decision analyses can bey operationalized through different decision analytic models that should be selected considering the context and available data. Decision trees and simulation models are the most commonly known type of analysis. Regardless of their structural form, several similarities across healthcare decision analyses exist:

- They require the clinical and policy relevant features of the problem, the time horizon of the analysis and the description of the target population
- They require information on the probability of experiencing a health state or a health event
- They require data on the value associated with a health state or health event (e.g. cost, health effect or both)

- Almost all healthcare decision analyses use inputs from multiple studies or sources given to limitations on data availability

A decision analysis tree outlines and quantifies the consequences of the two or more options. It can be represented by means of a tree diagram as showed in Figure 3 that is constituted by one decision node at the root; branches representing all the strategies that are to be compared; a series of chance nodes off every strategy branch from which emanate the possible consequences of each choice (e.g. transition states); and outcomes depicted at the end of each pathway. The underlying likelihoods of the occurrence of the transition states are called 'transition probabilities'. The sum of all transition probabilities emanating from a chance node is always one. The terminal nodes, where the health impact of each consequence, called payoff (e.g. utilities), is quantified in the analysis [14, 16].

Considering computer-based decision-analytic modeling, the state-transition modeling is frequently used as is considered an intuitive, flexible, and transparent approach. It can be used to model a cohort of patients (called Markov cohort model) or individuals (called microsimulation or first-order Monte Carlo model) [21, 27].

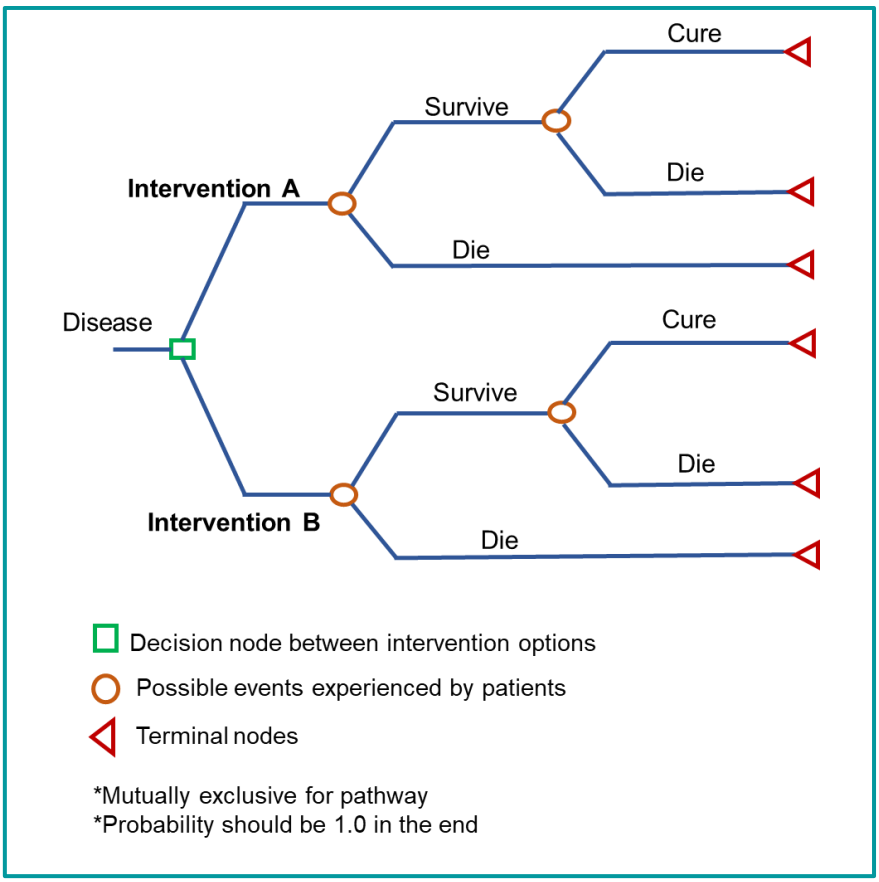


Figure 3. Simple hypothetical example of a decision analysis tree

Markov models are analytical frameworks that represent disease processes evolving over time and are suited to model progression of chronic disease as this type of model can handle disease recurrence and estimate long-term costs and life years gained/QALYs. This type of model use 'disease states' to represent all possible consequences of an intervention of interest. These are mutually exclusive and exhaustive and so each individual represented in the model can be in one and only one of these disease states at a specified or fixed period of time. Individuals move ('transition') between disease states as their condition changes over time. Time itself is considered as discrete time periods called 'cycles' (e.g. number of weeks or months), and movements from one

disease state to another (in the subsequent time period) are presented as 'transition probabilities'. At the end of each cycle, the individual can either stay in the same health state or move to another health state. Time spent in each disease state for a single model cycle (and transitions between states) is associated with a cost and a health outcome (see Figures 4) [16, 27].

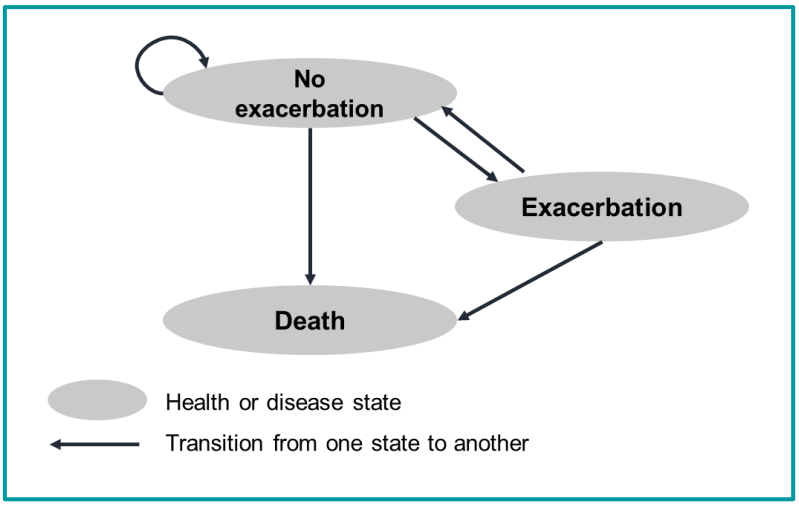


Figure 4. Simple hypothetical example of the Markov model

Another model with flexible frameworks is the discrete event simulation that is useful in emergency care or in the transmission of infections such as in constrained resources environments. In these cases, the events must be mutually exclusive and individuals progress through the model only if they experience a new event. However, because this type of model is used to represent complex systems, it is usually more difficult to develop, implement, and analyze [16, 27, 38].

Sensitivity analyses

Sensitivity analysis is used to illustrate and assess the level of confidence that may be associated with the conclusion of an economic evaluation. It is performed by varying key assumptions made in the evaluation (individually or severally) and recording the impact on the result (i.e. output) of the evaluation.

In model-based economic evaluations this includes varying the values of key input parameters, as well as structural assumptions concerning how the parameters are combined in the model. These analyses are usually classified into

- **One-way or simple sensitivity analysis**: input parameters are varied one by one as demonstrated in Figure 5 in a so called 'tornado chart' or 'tornado diagram'. In this graph, the X-axis is the net present value (NPV) (i.e. total present value over a period). Longer bars indicate more sensitive variables.

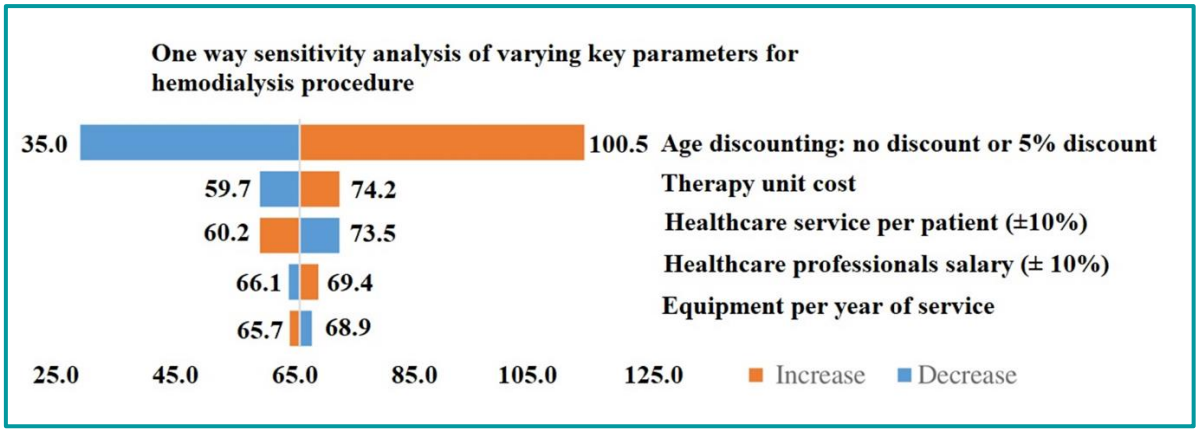


Figure 5. Hypothetical example of a one-way sensitivity analysis

- **Multi-way sensitivity analysis:** more than one parameter is varied at the same time. It should be noted that multi-way

sensitivity analysis becomes more difficult to interpret as progressively more variables are varied in the analysis

- **Threshold analysis**: the model is used to assess the tipping point for an input parameter (at what value of this parameter would the decision based on the output of the evaluation be altered?)
- **Probabilistic analysis:** a stochastic approach is taken to produce a distribution of outputs based 'n' distributions of input parameters. In this scenario, faced with the choice of whether or not to reimburse a new technology, the decision maker will likely be interested in the probability that the new technology is cost-effective compared to the existing alternative. This probability can be identified from ICER plane with reference to the decision-maker's defined maximum acceptable ceiling ratio (λ). This probability is simply the proportion of the scatter plot points that fall to the south and east of a ray with slope of λ drawn through the origin (i.e., proportion of incremental cost-effect pairs with a value below λ). Since the maximum acceptable ceiling ratio will generally not be stated explicitly, a sensitivity analysis should be undertaken with the probability determined for a range of λ s. The cost-effectiveness acceptability curves (CEAC) provides a plot of these probabilities (y-axis) against λ (x-axis) (see Figure 6).

Cost-effectiveness acceptability curves were introduced as an alternative to calculating confidence intervals for ICERs with statistical methods. The CEAC indicates the probability that an intervention is cost-effective compared with the alternative, given the observed data, for a range of λ values. Given a specified

value of this 'acceptable' cost-effectiveness ratio (i.e., λ on the x-axis), the CEAC shows the probability (read off on the y-axis) that the data are consistent with a true cost-effectiveness ratio falling below that value.

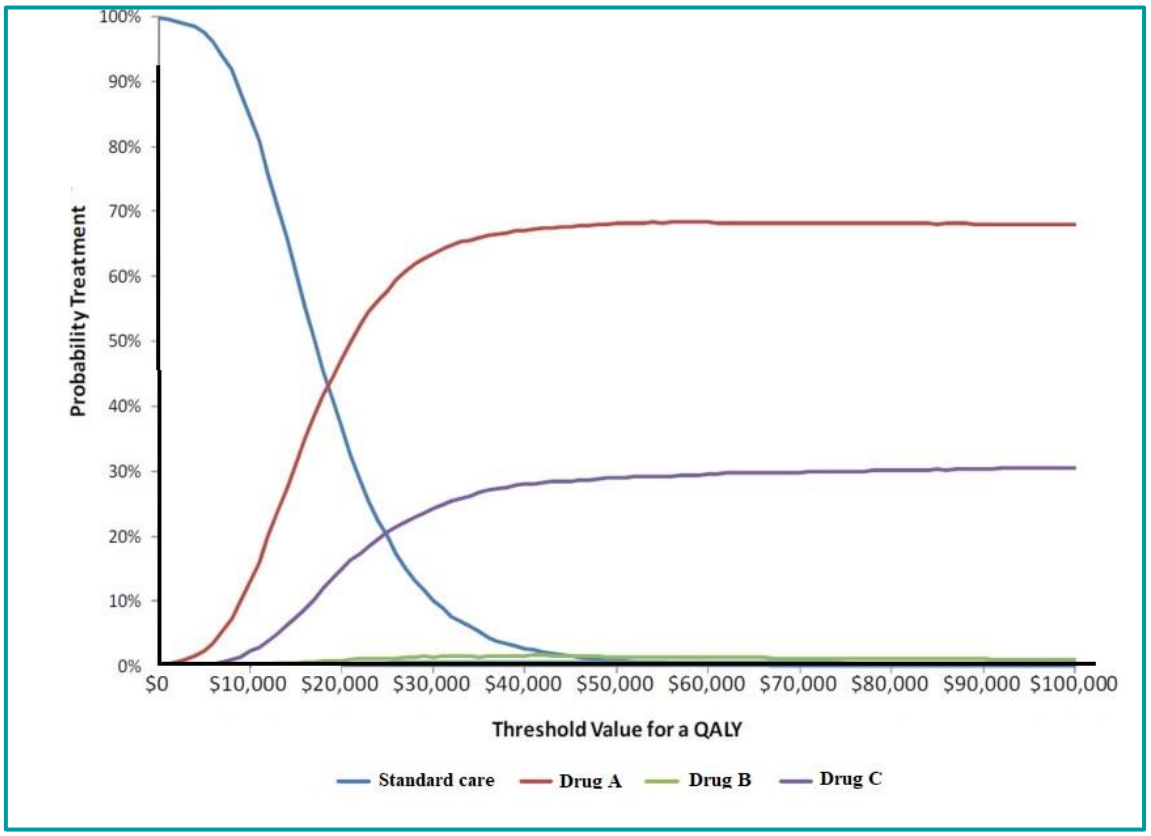


Figure 6. Hypothetical example of a CEAC

Sensitivity analysis is an important part of the evaluation process and gives valuable information to decision-makers about the robustness of their decision based on the findings of an economic evaluation, as well as the potential value of collecting more information before making a decision.

IGNACIO AZNAR-LOU
VASCO M. PONTINHA

Conclusions

The content presented in this chapter is paramount for the conduction of HTA economics studies. Some important concepts and methods were presented, such as the cost and outcomes that can be used in health economics analyses, the type of economics analyses (cost-minimization, cost-effectiveness, budget impact analysis) and the available models to perform a simulation cost analysis. Beyond the concepts presented here, other may be of interest to readers, such as how to deal with uncertainty in the analysis, which is inherent to any health economics analysis and may encompasses variability, heterogeneity, parameter, and structural uncertainty. Beyond that, to achieve more robust results and decrease uncertainty, individual patient data and real-world evidence are increasingly being focus of researches and incorporated into the economic analysis. Authors should strictly follow international guidelines and checklists (e.g. CHEERS) to perform and report an economic evaluation.

References

[1] Rabarison KM, Bish CL, Massoudi MS, Giles WH. Economic Evaluation Enhances Public Health Decision Making. Front Public Health. 2015;3:164.

[2] Clement FM, Harris A, Li JJ, Yong K, Lee KM, Manns BJ. Using effectiveness and cost-effectiveness to make drug coverage decisions: a comparison of Britain, Australia, and Canada. JAMA. 2009;302(13):1437-43.

[3] Allen N, Liberti L, Walker SR, Salek S. A Comparison of Reimbursement Recommendations by European HTA Agencies: Is There Opportunity for Further Alignment? Front Pharmacol. 2017;8:384.

[4] Sharma D, Aggarwal AK, Downey LE, Prinja S. National Healthcare Economic Evaluation Guidelines: A Cross-Country Comparison. Pharmacoecon Open. 2021.

[5] Walley T, Haycox A. Pharmacoeconomics: basic concepts and terminology. Br J Clin Pharmacol. 1997;43(4):343-8.

[6] McIntosh E, Luengo-Fernandez R. Economic evaluation. Part 1: Introduction to the concepts of economic evaluation in health care. J Fam Plann Reprod Health Care. 2006;32(2):107-12.

[7] Murphy EM, Rodis JL, Mann HJ. Three ways to advocate for the economic value of the pharmacist in health care. J Am Pharm Assoc (2003). 2020;60(6):e116-e24.

[8] Goodman C. HTA 101: Introduction to Health Technology Assessment. Bethesda, MD: National Library of Medicine (US). 2014.

[9] Drummond M, Sculpher M, Torrance G, O'Brien B, Stoddart G. Methods for the Economic Evaluation of Health Care Programmes. 3rd Edition Oxford: Oxford University Press. 2005.

[10] Eccles M, Mason J. How to develop cost-conscious guidelines. Health Technol Assess. 2001;5(16):1-69.

[11] Nixon J, Stoykova B, Glanville J, Christie J, Drummond M, Kleijnen J. The U.K. NHS economic evaluation database. Economic issues in evaluations of health technology. Int J Technol Assess Health Care. 2000;16(3):731-42.

[12] McGhan WF, Al M, Doshi JA, Kamae I, Marx SE, Rindress D. The ISPOR Good Practices for Quality Improvement of Cost-Effectiveness Research Task Force Report. Value Health. 2009;12(8):1086-99.

[13] Fenwick E, Steuten L, Knies S, Ghabri S, Basu A, Murray JF, et al. Value of Information Analysis for Research Decisions-An Introduction: Report 1 of the ISPOR Value of Information Analysis Emerging Good Practices Task Force. Value Health. 2020;23(2):139-50.

[14] Husereau D, Drummond M, Petrou S, Carswell C, Moher D, Greenberg D, et al. Consolidated Health Economic Evaluation Reporting Standards (CHEERS) statement. Value Health. 2013;16(2):e1-5.

[15] Jolicoeur LM, Jones-Grizzle AJ, Boyer JG. Guidelines for performing a pharmacoeconomic analysis. Am J Hosp Pharm. 1992;49(7):1741-7.

[16] Ademi Z, Kim H, Zomer E, Reid CM, Hollingsworth B, Liew D. Overview of pharmacoeconomic modelling methods. Br J Clin Pharmacol. 2013;75(4):944-50.

[17] Danzon PM, Drummond MF, Towse A, Pauly MV. Objectives, Budgets, Thresholds, and Opportunity Costs-A Health Economics Approach: An ISPOR Special Task Force Report [4]. Value Health. 2018;21(2):140-5.

[18] Graves N, Walker D, Raine R, Hutchings A, Roberts JA. Cost data for individual patients included in clinical studies: no amount of statistical analysis can compensate for inadequate costing methods. Health Econ. 2002;11(8):735-9.
[19] Briggs A, Clark T, Wolstenholme J, Clarke P. Missing... presumed at random: cost-analysis of incomplete data. Health Econ. 2003;12(5):377-92.
[20] Kent S, Burn E, Dawoud D, Jonsson P, Ostby JT, Hughes N, et al. Common Problems, Common Data Model Solutions: Evidence Generation for Health Technology Assessment. Pharmacoeconomics. 2020.
[21] Epstein D, Sutton A. Modelling correlated clinical outcomes in health technology appraisal. Value Health. 2011;14(6):793-9.
[22] Brazier J, Ara R, Azzabi I, Busschbach J, Chevrou-Severac H, Crawford B, et al. Identification, Review, and Use of Health State Utilities in Cost-Effectiveness Models: An ISPOR Good Practices for Outcomes Research Task Force Report. Value Health. 2019;22(3):267-75.
[23] Whitehead SJ, Ali S. Health outcomes in economic evaluation: the QALY and utilities. Br Med Bull. 2010;96:5-21.
[24] Torrance GW, Feeny D. Utilities and quality-adjusted life years. Int J Technol Assess Health Care. 1989;5(4):559-75.
[25] Schwarzer R, Rochau U, Saverno K, Jahn B, Bornschein B, Muehlberger N, et al. Systematic overview of cost-effectiveness thresholds in ten countries across four continents. J Comp Eff Res. 2015;4(5):485-504.
[26] Berger M. Willingness to pay versus willingness to buy: what defines value in healthcare? Value Health. 1998;1(4):201-3.
[27] Stahl JE. Modelling methods for pharmacoeconomics and health technology assessment: an overview and guide. Pharmacoeconomics. 2008;26(2):131-48.
[28] Basu A, Maciejewski ML. Choosing a Time Horizon in Cost and Cost-effectiveness Analyses. JAMA. 2019;321(11):1096-7.
[29] Kim DD, Wilkinson CL, Pope EF, Chambers JD, Cohen JT, Neumann PJ. The influence of time horizon on results of cost-effectiveness analyses. Expert Rev Pharmacoecon Outcomes Res. 2017;17(6):615-23.
[30] Westra TA, Parouty MB, Wilschut JC, Boersma C, Postma MJ. Practical implications of differential discounting of costs and health effects in cost-effectiveness analysis. Value Health. 2011;14(8):1173-4; author reply 4-5.
[31] Hay JW. Economic modeling and sensitivity analysis. Value Health. 1998;1(3):187-93.
[32] Annemans L, Geneste B, Jolain B. Early modelling for assessing health and economic outcomes of drug therapy. Value Health. 2000;3(6):427-34.
[33] Boshuizen HC, van Baal PH. Probabilistic sensitivity analysis: be a Bayesian. Value Health. 2009;12(8):1210-4.
[34] Greenberg D, Mohamed Ibrahim MIB, Boncz I. What Are the Challenges in Conducting Cost-of-Illness Studies? Value Health Reg Issues. 2014;4:115-6.
[35] Onukwugha E, McRae J, Kravetz A, Varga S, Khairnar R, Mullins CD. Cost-of-Illness Studies: An Updated Review of Current Methods. Pharmacoeconomics. 2016;34(1):43-58.
[36] Bertram MY, Lauer JA, De Joncheere K, Edejer T, Hutubessy R, Kieny MP, et al. Cost-effectiveness thresholds: pros and cons. Bull World Health Organ. 2016;94(12):925-30.

[37] Sullivan SD, Mauskopf JA, Augustovski F, Jaime Caro J, Lee KM, Minchin M, et al. Budget impact analysis-principles of good practice: report of the ISPOR 2012 Budget Impact Analysis Good Practice II Task Force. Value Health. 2014;17(1):5-14.
[38] Zhang X. Application of discrete event simulation in health care: a systematic review. BMC Health Serv Res. 2018;18(1):687.

ADVANCED STATISTICAL CONCEPTS: NETWORK META-ANALYSES AND VALUE OF INFORMATION

LETICIA P. LEONART

Abstract

In this chapter, four themes relevant to those interested in health technology assessment are presented: network meta-analysis, living systematic review, model based meta-analysis and value of information analyses. Using a simple language and avoiding the excess of mathematical equations, a straightforward introduction to these themes is given, which should be an easy starting point for those who want to delve into more advanced subjects on health technology assessment. Network meta-analyses is becoming an essential knowledge to those who work with systematic reviews and meta-analysis, as it enables the comparison of multiple interventions in one sole analysis. Living systematic review is also being increasingly used due to the rapid pace of clinical studies publication, helping to guarantee that the evidence synthesis is always up to date. Model based meta-analysis is becoming more and more used as it enables the introduction of data that goes beyond that reported on clinical trials into the analysis, in an effort to use the synthesized information during drug development and not only after the drug is on the market. Value of information analyses bring a meaning to the results of cost-effectiveness probabilistic sensitivity analyses, aiding in the identification of what parameters contribute the most to the uncertainty of the results and where to focus further research.

Keywords: network meta-analysis; living systematic review; model based meta-analysis; value of information

Introduction

With the global status that science has acquired in the last decades, in which information is available in a matter of seconds, and with the high increase in processing power of machines and software, a vast range of new techniques and possibilities to improve clinical and economic data became available. In this chapter, an introduction to relevant and increasingly used techniques in health technology assessment is presented, encompassing network meta-analysis, living systematic review, model-base meta-analysis, and value of information.

Knowledge of network meta-analysis is now critical when performing studies for potential incorporation of health technologies into health systems. NMAs are easily encountered in analyses performed for and by health agencies, such as the National Institute for Health and Care Excellence (NICE) from the United Kingdom and the Canadian Agency for Drugs and Technologies in Health (CADTH). Knowing how to perform an NMA is, therefore, essential for those who wish to work in the field of health technology assessment. Or, at least, know how to interpret an NMA.

Living systematic review is also a technique that has been increasingly used in evidence synthesis due to the fast publication of new clinical studies, guaranteeing that the evidence generated by systematic reviews does not become obsolete and that the decision-making is based on up-to-date high-quality evidence. Model-based meta-analysis is an expansion to the limits of traditional meta-analysis and, in this chapter, we illustrate its importance in drug development and as a tool to exploit other types of evidence beyond phase III clinical trials.

We also present the value of information analyses and its potential to convert the sometimes hard to digest results of probabilistic sensitivity analysis from cost-effectiveness studies into an enlightening and useful information that can help to identify the most critical and uncertain parameters and, thus, clarify where further research should be focused.

Network meta-analysis

Through traditional pairwise meta-analysis, it is not possible to compare more than two interventions simultaneously. Sometimes, the results obtained from pairwise meta-analyses are not enough to answer the scientific question being asked and to solve the decision-making problem, as it is common that more than two technologies (e.g. drugs, diagnostic tests, etc.) are available in the market for the same condition. Furthermore, comparing all the relevant technologies for the same health condition simultaneously can provide an overview of the state of the art about its treatments, such as existing approaches, number of trials, which technologies have been directly compared, what are the relative risks and benefits, etc. and guide the direction for new studies and minimum/expected performance of new technologies. Considering that, network meta-analysis can be an alternative to traditional meta-analysis [1-6].

Using network meta-analysis (NMA), effect measures of interventions that have not been directly compared in clinical studies can be indirectly calculated through a common comparator. For example: for a given disease there are three drugs, A, B and C. Drug A was compared in randomized controlled trials with drugs B and C, but B and C were never directly compared.

Using the common comparator, A, it is possible to estimate the relative effect of B vs C. This process is exemplified in Figure 1.

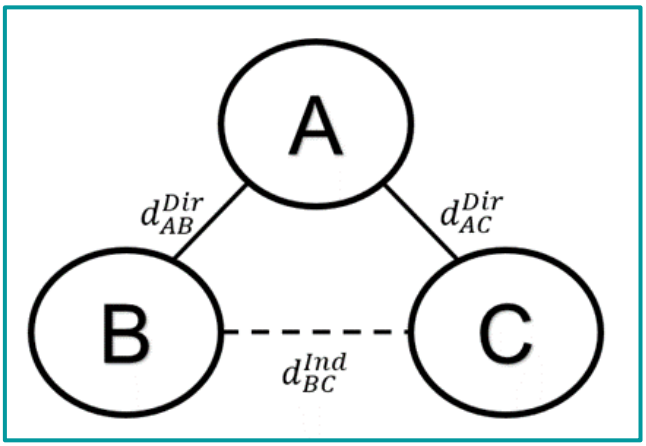


Figure 1. Scheme exemplifying the indirect effect estimate process

Thus, the indirect estimate between B and C can be obtained according to Equation 1. The variance of the indirect estimate is equivalent to the sum of the variance of the direct estimates (see Equation 2):

$$d_{BC}^{Ind} = d_{AC}^{Dir} - d_{AB}^{Dir} \ (Equation\ 1)$$

$$V_{BD}^{Ind} = V_{AC}^{Dir} + V_{AB}^{Dir} \ (Equation\ 2)$$

In addition, when both direct and indirect evidence are available for the same comparison, it is possible to calculate a mixed effect measure [2,7].

A brief history of indirect comparisons

In 1997, Bucher and colleagues [8] proposed a method for indirectness comparison, preserving the randomization of the originally assigned patient groups. This method is known as 'adjusted indirect treatment comparison' (ITC) or 'anchored ITC' and was developed for the comparison of three interventions, estimating the odds ratio (OR) as effect measure. Despite the merit of being an important advance on the field, the proposed method has some limitations: the restricted number of interventions that can be compared (n=3), it can only be applied to two-arm trials and it assumes that the relative effect is the same through all the studies [9,10]. Despite not being the number one choice of indirect comparison anymore, it has an important educational value as it is easily understandable, as shown in the previously depicted Figure 1, which represents the approach proposed by Bucher.

The widely accepted term 'network meta-analysis' was presented by Lumley (2002) [11]. Before that, other terms were employed, which are sometimes still used: mixed treatment comparisons or multiple-treatments meta-analysis [4]. Lumley presented a frequentist approach for indirect comparison. With this approach, it is possible to compare simultaneously more than three interventions: A versus B through more than one common comparator, for example, C and D, in a scenario where there is no trial comparing A versus B but there are studies comparing A versus C, A versus D, B versus C and B versus D. Besides that, it is possible to estimate the level of agreement of the effect measures obtained through the different common comparators: if the indirect comparisons between A and B through C or between A and B through D produce similar results, the confidence that the overall effect measure represents the true value is high (more than one

comparison points out to the same result). On the other hand, if the effect measures using the common comparators (C and D) are disparate, there is inconsistency in the network and the result may not be trustworthy.

In 2004, Lu and Ades [12] refined the NMA technique and allowed for the comparison of multiple strategies using direct and indirect evidence. They proposed a Bayesian framework and created a classification system to rank the interventions being compared considering the probability of yielding the best result, second best and so on for each outcome [12,13]. Since then, the use of NMA has exponentially grown and it is largely applied and accepted as an important tool in the decision-making process.

Understanding the network structure

Depending on the number of clinical studies available and the number of comparators, the network can have different geometries. A graphical representation of the network diagram can help to better understand and visualize the analysis being performed. Some examples of possible network connections are shown in Figure 2. With the increase of direct and indirect comparisons, the complexity of the network also enhances as can be visualized from a) to d).

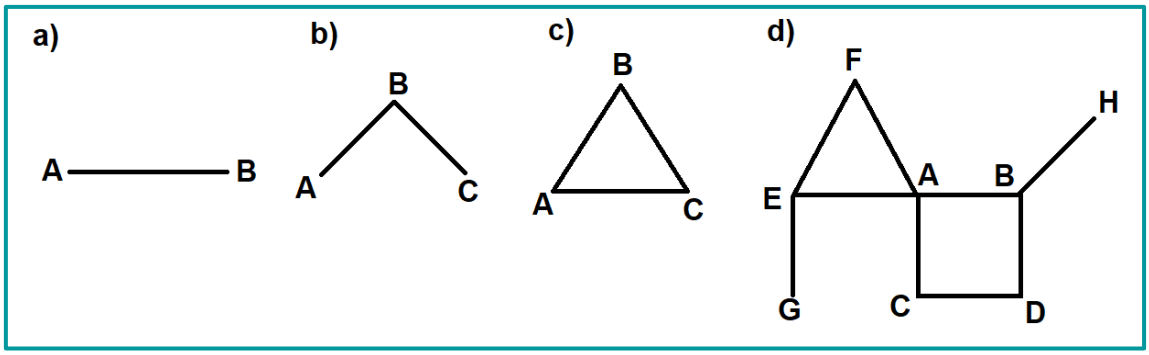


Figure 2. Examples of different possible connections in a network
a) a simple pairwise comparison; b) an indirect comparison between A and C via B; c) a triangle network with a closed loop; d) a complex network with open and closed loops.

In Figure 2, the examples shown in letters b and c have the same interventions being compared: A, B and C. The difference between these two networks is that in Fig. 2-b the estimated effect of A vs C comes only from indirect comparison via B. On the other hand, in Fig 2-c, the estimated effect of A vs C is available via B, but there is also direct comparison, from a trial comparing A vs C, forming what is called a 'closed loop'. In this case, we have available both direct and indirect estimates, which can be pooled together to originate a mixed effect and allow for the evaluation of network consistency by assessing the agreement between the results derived from these two types of comparisons [14].

An example of a traditional network diagram is shown in Figure 3. The circles, called 'nodes', represent the interventions being compared. The lines connecting the nodes represent which interventions were direct compared by clinical studies. The width of the line can be used to demonstrate the number of studies that assessed each comparison. For example, in Figure 3, it is possible to see that A-C and C-D are connected by a thicker line, meaning that there is

more direct evidence for these two comparisons. Other way of representing this is simply putting the number of trials beside the connecting line. A closed loop can be seen in the C-D-E triangle, and an open loop is shown in the A-B-C-D square. In open loops, the evidence can be considered weaker because for some technologies (e.g. B vs D) it relies only on indirect evidence [14-17].

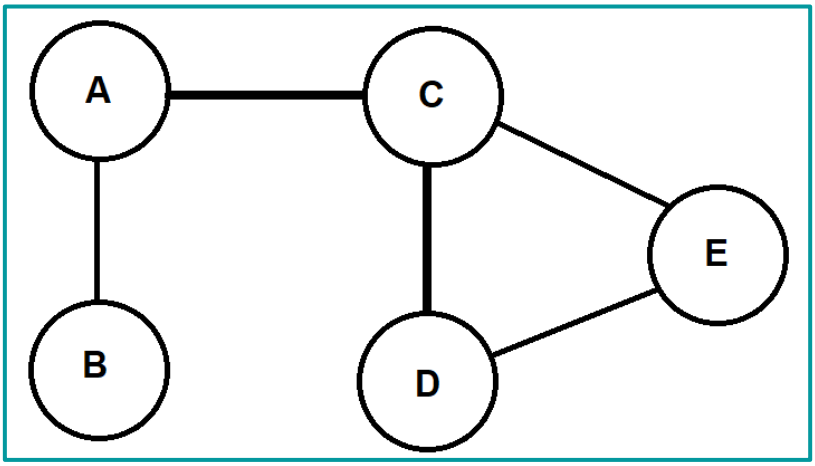


Figure 3. An example of a network diagram.

Assumptions for conducting network meta-analysis

For conducting a reliable NMA, some assumptions should be met. First of all, the same principles that apply to pairwise meta-analysis also apply to NMA, after all, NMA incorporates direct evidence from pairwise meta-analysis in its results. The most important principle, which should be assessed before conducting any meta-analysis, is the guarantee that the included trials are comparable, i.e., 'not mixing apples and oranges', or misleading conclusions can be achieved [18].

The main assumptions that should be considered when conducting NMA are transitivity and consistency, which are discussed in the following paragraphs.

The transitivity ensures that the studies compared by the NMA are similar. The characteristics of patients and treatments must be comparable, that is, the effect modifiers must have a similar distribution between studies. This assumption cannot be assessed statistically, but its validity can be examined in a conceptual and epidemiological way. Ideally, participants in the different studies should be similar enough that they could be part of a mega-trial [6,7,12,19].

An example of a violation of transitivity is shown in Figure 4. In this example, four treatments are compared for weight loss: A, B, C and D. The interventions A, B and C have been evaluated in patients with normal body mass index (BMI), while the interventions B and D have been tested in patients with morbid obesity. There is no evidence of trials assessing A and C for morbidly obese patients and of D for normal BMI people. Therefore, it is inappropriate to indirectly estimate the effect of A vs D and C vs D for weight reduction, as the studies populations are not comparable and will likely show different results. To avoid this kind of mistake, it is essential that the research question is very well described. In this case, the question should clarify which BMI the patients included in the analysis should present.

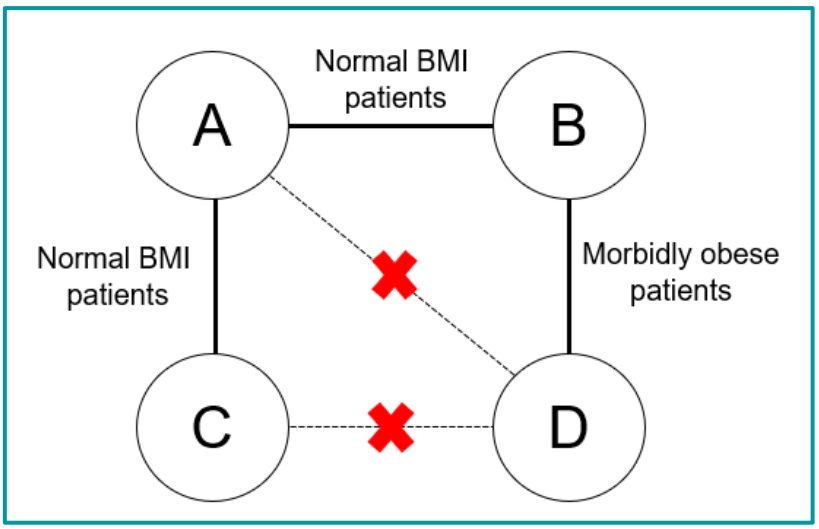


Figure 4. An example of a network that violates the principle of transitivity.
The solid line represents the direct comparisons, and the dashed line represents the indirect comparison.

Consistency can be considered the statistical manifestation of transitivity. This measure indicates whether the directly and indirectly estimated values agree. When there is no statistical agreement, it is said that there is 'inconsistency' in the model. Inconsistency can be seen as a special type of heterogeneity (a concept from pairwise meta-analysis): the latter is the result of varying effect modifiers between different studies, while the former is the result of varying effect modifiers between different comparisons. However, it is only possible to assess the inconsistency when there is direct and indirect evidence for the same comparison. That is, the formation of closed loops in the network is essential for the inconsistency to be estimated. Several approaches have been described for the evaluation of inconsistency, which can be categorized as local and global approaches. Local approaches evaluate separated regions of the

network, whilst global approaches evaluate the consistency in the entire network [6,7,18-22].

A widely used local approach recommended by Cochrane for evaluating inconsistency is called SIDE (Separating Indirect from Direct Evidence), introduced by Dias et al. in 2010 as 'node-splitting'. In this method, the direct evidence for a specific comparison is excluded from the network and compared with the remaining indirect evidence. The two estimates are then compared with a Z-test. This same process is made for each pairwise comparison in the network. However, it is important to mention that tests for inconsistency have low power, which means that occasionally they may suggest that there is no inconsistency when in fact there is. It is advisable to inspect the confidence interval of the inconsistency test to judge if there are really no clinically important discrepancies between direct and indirect evidence [6,7,18-22].

Bayesian network meta-analysis

With the increase in the number of interventions and the complexity of the networks, more advanced calculations are necessary to estimate the measures of effect of the NMA. In this context, the Bayesian meta-analysis using Monte Carlo simulation via Markov Chains (MCMC) becomes an appropriate approach, as it allows the use of simulation to estimate values that cannot be calculated analytically. This approach has been gaining space in recent years due to the advancement of computational capacity and the development of specific software, since these analyses may require a high processing power [4,21,23].

Bayesian statistics was named after the Bayes' Theorem. In this approach, an *a priori* probability distribution is combined with a likelihood to result in an *a posteriori* distribution. That is, Bayes' Theorem is used to review the previous probability distribution in light of new acquired data, resulting in an updated probability distribution [4,24,25]. This concept is described in the Equation 3.

$$p(\theta|Y) = \frac{p(Y|\theta)}{p(Y)} \times p(\theta) \text{ (Equation 3)}$$

Where:

p = probability

θ = parameter of interest

Y = new acquired data

p(Y) corresponds to a normalizing factor that can be omitted. Thus, the Equation 3 can be interpreted as follows:

$$Posteriori \propto Likelihood \times Priori \text{ (Equation 4)}$$

In the Bayesian statistics, an *a priori* probability distribution must be defined, which reflects an uncertainty (belief) about the effect before conducting the analysis. When there is no prior information, the prior distribution may be non-informative or vague (e.g. $d \sim Normal(0, 100^2)$), strategy generally adopted in NMA [4,23-26].

Contrary to what is applied to the frequentist pairwise meta-analysis, the choice of the most appropriate Bayesian model (fixed effect or random effects)

depends on the assessment of the deviance information criteria (DIC). The DIC is equal to the mean posterior deviation plus the number of parameters. Models with smaller DIC are preferable as they are less complex [4,20].

Instead of estimating confidence intervals, as in frequentist statistics, Bayesian statistics estimates credibility intervals (CrI). A 95% CrI indicates a 95% probability that the true effect is in the shown range. A 95% confidence interval indicates that 95% of the calculated confidence intervals will include the true value of the estimated effect. In practice, the Bayesian interval is more easily understandable, and it is not rare that frequentist results are mistakenly interpreted according to the Bayesian definition [23,24,27].

The use of MCMC in statistic Bayesian analysis has become frequent as it allows the execution of complex networks through simulation. First, an initial parameter value is chosen at random. At each iteration, this parameter is updated following a stochastic process. After many iterations, an accurate estimate of the model is obtained. As each chain starts with a different random value, it is suggested that at least three chains be tried. In addition, the first iterations must be discarded (burn-in). In order to check if model convergence has been achieved, some strategies can be used. For example, the Brooks-Gelman-Rubin method runs several chains with different starting values and compares the results. Numerically, convergence can be assessed by calculating the Potential Scale Reduction Factor (PSRF), which considers the variance of the chains. If the PSRF value is close to 1, it means that convergence was achieved [4,25,28].

The results of a NMA can be presented in a consistency table, which usually shows the mixed effect measure (the combination of direct and indirect

effects). Figure 5 is an example of how the results of an NMA can be presented. In the example, five interventions were compared: A, B, C, D and E. It was possible to estimate effect measures for all the comparisons, which are shown in pairs: A vs B, A vs C, B vs C and so on. In this example, the results are presented as *odds ratio* (OR) along the 95% CrI. If the result is greater than 1, it favors the intervention that is shown in the row of the cell where the effect measure is. If the result is lower than 1, it favors the intervention that is shown in the corresponding column. Statistically significant results are shown on bold and underlined. For A vs B, the estimated effect is OR 2.30 (CrI 95% 1.50 – 3.00). This result is placed in the row A and column B. Therefore, as 2.30 > 1.00, the result favors the intervention A. As the CrI 95% does not cross the null value (for OR, the null value is 1.00), this result is statistically significant. The same is observed for A vs C. In the case of B vs D, the CrI 95% does not cross the null value (statistical difference between treatments) and the result 0.56 is < 1.00, which means it favors the intervention in the corresponding column (D).

A	**2.30 (1.50 – 3.00)**	**2.45 (2.00 – 3.08)**	1.50 (0.78 – 2.59)	1.95 (0.98 – 2.65)
	B	0.75 (0.47 – 1.45)	**0.56 (0.35 – 0.89)**	0.99 (0.74 – 1.22)
		C	0.98 (0.55 – 1.99)	1.22 (0.88 – 1.46)
			D	1.01 (0.90 – 1.42)
				E

Figure 5. Example of a consistency table used to report NMA results.
Results are shown as *odds ratio* (OR) along with 95% credible intervals.

In addition to allowing the comparison of several treatments simultaneously, the NMA also estimates which of the intervention is more likely to be the best option for a given outcome, through a ranking analysis. For this, an analysis known as SUCRA (surface under the cumulative ranking curve) can be used. In SUCRA analysis, values close to 1 indicate a greater probability that the treatment is the best and close to 0 indicate a lower probability. For each treatment x of a total of y treatments, the cumulative probability vector ($cum_{x,z}$) must be calculated to be among the z best treatments z = 1, 2, ... y. Hence, the SUCRA value for each treatment can be obtained [29] (see Equation 5):

$$SUCRA_x = \frac{\sum_{z=1}^{y-1} cum_{x,z}}{y-1} \quad (Equation\ 5)$$

In Figure 6 a hypothetical SUCRA analysis is shown. In this example, a ranking was built for the comparison of four interventions: A (blue), B (purple), C (green) and D (yellow).

Ranking	Drug A	Cumulative %	Drug B	Cumulative %	Drug C	Cumulative %	Drug D	Cumulative %
1	70%	70%	1%	1%	15%	15%	14%	14%
2	15%	85%	5%	6%	30%	45%	50%	64%
3	14%	99%	77%	83%	8%	53%	1%	65%
4	1%	100%	17%	100%	47%	100%	35%	100%
SUCRA		85%		30%		38%		48%

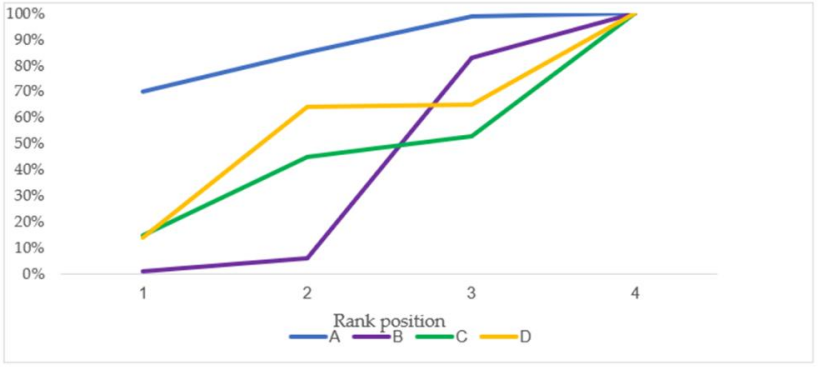


Figure 6. Example of a SUCRA analysis.

In the table, the probabilities of each intervention to be the best, second best, third and fourth, alongside the cumulative probability, are shown. Applying the equation depicted above, a SUCRA value for each one of the interventions was calculated and plotted on the graph. For this hypothetical outcome, the A intervention has the highest probability of being the best one (85%). This approach provides a holistic view of the NMA results and helps to summarize the findings. However, it should be interpreted with caution when no statistical significance is achieved in the comparisons by pairs or when inconsistency is present in the network.

Software for network meta-analysis

Due to the growth and relevance of NMA, a lot of software now support this type of analysis, such as WinBUGS, OpenBUGS, and STATA. R, a frequently used software environment for statistical data analysis, has packages

developed specifically for the conduction of Bayesian NMA, such as GeMTC and BUGSnet. A user-friendly option to conduct NMA that can be used by those unfamiliar with coding is the 'Aggregate Data Drug Information System' (ADDIS), which allows the conduction of consistency and inconsistency analyses through a graphic interface. Access: **addis.drugis.org.** However, if the user wants to perform more advanced analyses and have more flexibility, other options that involve programming language are necessary.

Living systematic review and meta-analysis

Living systematic reviews (LSR), which can include living meta-analysis, is a continuously updated systematic review that aims to incorporate new evidence as soon as it becomes available. This is a reflex of fast pace research observed in some topics nowadays, such as COVID-19. For instance, the COVID-NMA is an international initiative (covid-nma.com) launched on March 2020, led by Cochrane and some other prominent institutions working in conjunction with the World Health Organization (WHO), which promotes up-to-date evidence on clinical studies (peer review and yet not peer reviewed) on COVID-19, including all classes of treatments (antivirals, other antimicrobials, anti-inflammatories, monoclonal antibodies, etc) and focus (curative or preventive therapy, out or inpatients, etc), as well as studies on vaccines. This is a perfect example to illustrate the need for fast and rigorous assessment of primary evidence, as we face a pandemic disease that still has no definitive treatment and urges the need for high quality evidence so trustworthy decisions on where to allocate economic and human resources can be made.

The COVID-NMA group has structured a fast and functional approach to identify any relevant study as soon as it becomes available, extract its data, evaluate its risk of bias and grade the quality of the evidence [30]. LSR are ideal for topics that are priority for decision makers and that still carry uncertainty on the results, i.e., new evidence can affect the current results and alter the conclusions.

The methods of LSR are fundamentally the same of those applied in systematic reviews. However, some additional information is necessary: how often the LSR will be updated (which can be influenced by the quickness of new clinical studies) and when and how the new evidence will be incorporated into the review [18]. LSR are also a way to minimize duplicate and heavy workload, as most part of data will have been already evaluated, just needing the incorporation of the new evidence. Besides that, it is possible to inform the reader on how the new evidence changed (or not) the results compared to the previous analysis.

Some care must be taken concerning the analysis and results of meta-analyses when the review is updated. A study has shown that frequently updating a meta-analysis can increase the probability of type 1 error in 2 to 5 folds, which can exceed the inflation caused by publication bias [31]. This care is especially important if the results of the meta-analyses are being used in a decision-making context, where the effect estimate and its precision are used to formulate recommendations and grade the evidence, and not only as a summary of the evidence at the time of the most recent update. The results of meta-analyses are usually presented along with a 95% confidence interval, considering a p value of 0.05 (alpha). This means that the meta-analysis has a 5% probability of finding a statistically significant result when there is no

difference between groups, which is the type 1 error. As the meta-analysis is recurrently updated, the probability of the type 1 error jointly increases [32].

Other important use of meta-analysis for decision-making is to achieve a level of evidence graded as high, which would mean that the estimate is reliable enough that new studies would not affect the conclusions anymore. In this case, care must be taken with error type 2, which refers to failing to detect a true effect when there is one, especially when the results of the meta-analysis point out to absence of statistical difference between the comparators [32].

Some methods have been developed to minimize types 1 and 2 error: trial sequential analysis, sequential meta-analysis, the Shuster method, and law of the iterated logarithm. The 'trial sequential analysis' penalizes the alpha at each analysis in order to avoid the type 1 error, and, to avoid type 2 error, the sample size is calculated in the same way as the sample size for a clinical trial. In the 'sequential meta-analysis' method, Whitehead's sequential trial boundaries approach to control type 1 error inflation and type 2 error are used. The 'Shuster method' is similar to the trial sequential analysis but uses more conservative boundaries. Finally, the 'law of the iterated logarithm' adjusts the Z statistic, so that the alpha (maximum type 1 error) is kept stable through the updates. For more details on these methods, read the paper by Simmonds et al. (2017) [32].

The success of living systematic reviews and meta-analyses is related to the engagement of researchers in making study data promptly available, so the systematic reviewers do not encounter barriers to access the necessary data. In this way, open, shared, and global science is paramount for the prosperity of initiatives that aim to produce fast and high-quality clinical evidence synthesis.

It is also important to mention that for a LSR to succeed a very qualified team and a sound methodology are necessary, as this is a project that depends on continuous attention and does not have an 'end date'. The reviewers need to estimate how many studies are likely to be published between the searches dates to predict the workload.

Cochrane has published in December 2019 a document called 'Guidance for the production and publication of Cochrane living systematic reviews: Cochrane Reviews in living mode' which brings details on how to conduct an LSR [33].

Model-based meta-analysis

Model-based meta-analysis (MBMA) is a technique that is still in development and consists in incorporating data beyond what is usually included in traditional meta-analyses. A search in PubMed in February 2021 looking for the term retrieved only 108 registries, being 2020 the year with more publication (n=21).

During drug development, some of the objectives are to determine the relation between dose and response and what factors can alter the response. However, usually this type of data is not available for other drugs of interest, that would be the relevant comparators (competitors) of the new drug being developed. In this context, MBMA integrates efficacy and safety data from clinical trials, including preclinical data and predictive biomarkers using pharmacological models of dose/exposure-response, time-response, and/or cross-endpoint relationships. Therefore, MBMA can be applied during drug development in order to estimate the response of the new drug in relation to relevant comparators and prioritize and direct research before more advance

trials (phase III) should be initiated. MBMA can be used to identify which would be the most appropriate comparator and which population could benefit the most with the new drug. Contrary to what is usually seen for clinical trials and traditional meta-analysis, the purpose of MBMA is learning rather than confirming. Using mathematical models, MBMA enables the comparison of technologies that have not been head-to-head compared in clinical trials. Furthermore, MBMA can incorporate longitudinal and dose-response data, proving information that usually are not available with traditional meta-analyses but that are important during drug development [34-36].

As with traditional systematic review and meta-analysis, the conduction of MBMA warrants a very well-defined research question to define its objectives, and a pre-elaborated protocol with all steps for studies and data selection and statistics analyses.

As MBMA is an approach to guide drug development, it can be of interest of pharmaceutical industry. With that in mind, internal and external data should be aggregated into the analysis. Internal data refers to in house clinical trials, with positive or negative results, from which the researchers will have access to patient level data. External data are those available from the scientific literature, such as published clinical trials, including detailed information on trial design, population, and treatments [34]. In this way, the drug being developed can be compared to what is already on the market.

For more details on the MBMA topic, access 'The Handbook of Research Synthesis and Meta-Analysis' [37] and the articles cited in this chapter.

Value of information

Decisions on reimbursement of health technologies are usually made in scenarios of uncertainty, which means that there is always the possibility that the wrong decision is made, i.e., in some cases, the real best intervention may not be chosen because when the decision-making process took place the uncertainty about the results precluded the optimal data analysis. If this is the case, there will be investment and spending on a technology that is suboptimal. For example: imagine that a new drug for treating metastatic breast cancer is available, and the payer would like to know if it is cost-effective to reimburse it in a scenario where another drug already exists. The overall result of the cost-effectiveness analysis showed additional health benefits, such as quality of life and years lived, for a reasonable increase in costs with the new drug. However, the sensitivity analyses showed that when exploring all the range of values that the parameters can adopt (because of parameters uncertainty), in some cases the new drug is not cost-effective. How to decide if the new drug should be reimbursed or not? Is the result of the cost-effectiveness analysis accurate enough or new evidence should be acquired before a decision is made?

The example above is now illustrated using a cost-effectiveness acceptability curve (CEAC), discussed in Chapter 06. This curve shows the probability of each one of the two interventions (old drug and new drug) being cost-effective given a range of different cost-effectiveness thresholds. Consider that for the example being discussed the cost-effectiveness threshold adopted by the payer is $70,000. From the CAEC analysis (Figure 7), it is possible to see that the new drug is cost-effective considering the $70,000 threshold compared to the old drug. However, it has only a 60% probability of being cost-effective. This means that from all the iterations obtained during the probabilistic

analysis, in 60% of the cases the new drug was cost-effective, but in the remaining 40%, it was not. Is a 60% probability enough to guarantee that the new drug is the best option?

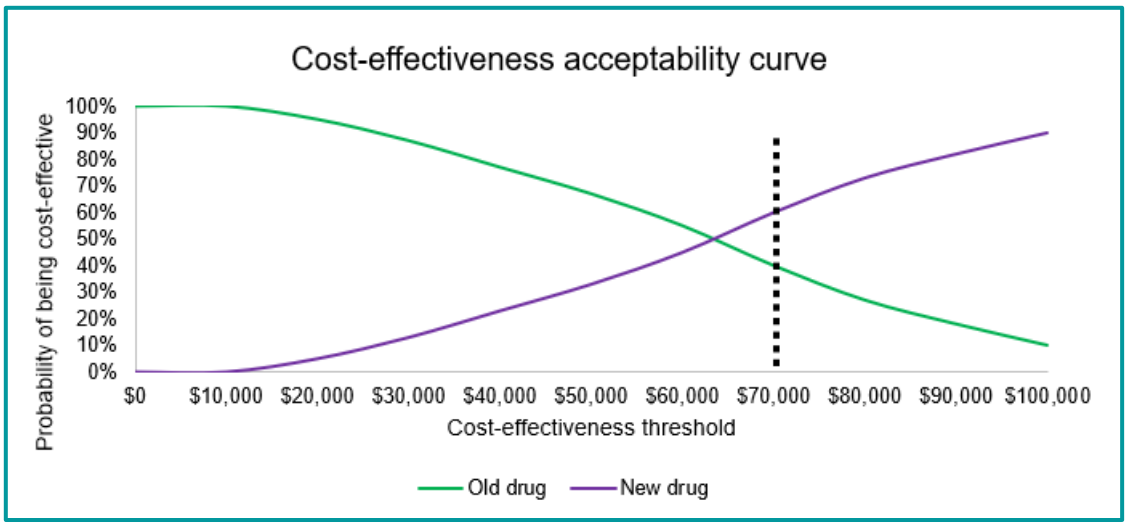


Figure 7. Example of a cost-effectiveness acceptability curve

Expected value of perfect information

The 'expected value of perfect information' (EVPI) analysis can help to solve this problem. At least, it can indicate if the uncertainty of parameters is low enough that a final decision could be made now, or if new evidence is still necessary. EVPI estimates how much the uncertainty that precludes the optimal decision-making costs, which corresponds to the maximum amount to be spent on future research. If the EVPI value is low, it means that it is better to comply with the alternative that the analysis indicates as the most cost-effective. If the EVPI value is high, there is an indication of an environment with high uncertainty that may interfere with the decision, and further research should be conducted to reduce this uncertainty and avoid an erroneous allocation of resources [38,39].

The EVPI estimates depends on the results of the probabilistic analysis (discussed in Chapter 06). In summary, the EVPI corresponds to the difference between the net benefit with perfect information (without uncertainties) and the net benefit of the current scenario (with uncertainties). For each iteration, the most cost-effective alternative is determined, that is, the one that generates the greatest net benefit.

Consider the example given in Figure 7. The new drug has a 60% probability of being cost-effective. But what if the true value of all parameters is in one of the other 40% simulations? How much would be lost because the wrong decision was made?

Each one of the simulations of the probabilistic analysis will have different results. Each simulation will show different costs, different quality of life adjusted years (QALYs), different net monetary benefit (NMB). Some of the iterations will point out that the new drug is the cost-effective option and some others that the old drug is the cost-effective alternative. But which simulation has sampled the real true values? We do not know that. If we had no uncertainty in the parameters, i.e., a perfect information, we would make the right decision and would avoid the expenditure of choosing the wrong option.

In Table 1, ten simulations of the decision problem being discussed (old drug versus new drug) is presented. Of course, in a real analysis, a much higher number of simulations should be performed, but for illustration purposes let us imagine that 10 is a sufficient number of simulations. The net monetary benefit (NMB) will indicate if the new drug is cost-effective in comparison to the old drug given a specific cost-effectiveness threshold, represented by λ (see

Equation 6). In our example, the threshold is U$ 70,000. If the NMB is a positive value, the intervention is cost-effective.

$$NMB = \lambda \times \Delta Effect - \Delta Cost \ (Equation\ 6)$$

The ten simulations in Table 1 show for each of the two interventions the cost, QALY and NMB. Depending on the NMB result, the cost-effective option is determined. After conducting ten iterations, in 6 (60%) cases, the new drug was the cost-effective strategy, whilst in 4 (40%) the old drug was cost-effective. As we do not know which of the simulations represents the true result (all have the same probability of being right), we calculate the average. In this case, the average shows that the new drug is the best option (positive NMB, $4,200).

Table 1. Calculation on the perfect information

Simulat.	Treatment	Cost	QALY	NMB	Cost-effective option	Value of perfect information
1	Old drug New drug	$400,00 $460,00	4.0 5.1	$17,000	New drug	$0
2	Old drug New drug	$410,00 $450,00	4.6 5.5	$23,000	New drug	$0
3	Old drug New drug	$405,00 $480,00	4.2 5.8	$37,000	New drug	$0
4	Old drug New drug	$400,00 $500,00	4.0 5.0	-$30,000	Old drug	$30,000
5	Old drug New drug	$398,00 $448,00	4.5 5.7	$34,000	New drug	$0
6	Old drug New drug	$395,00 $490,00	4.3 5.8	$10,000	New drug	$0
7	Old drug New drug	$380,00 $487,00	4.2 5.3	-$30,000	Old drug	$30,000
8	Old drug New drug	$404,00 $450,00	4.9 5.4	-$11,000	Old drug	$11,000
9	Old drug New drug	$422,00 $467,00	4.7 5.7	$32,000	New drug	$0
10	Old drug New drug	$410,00 $520,00	3.9 4.9	-$40,000	Old drug	$40,000

Simulat.	Treatment	Cost	QALY	NMB	Cost-effective option	Value of perfect information
Average	Old drug New drug	$402,400 $475,200	4.3 5.4	$4,200	New drug	$11,100

Now imagine that we have perfect information available, and we know that the true result is that of simulation number 1. In this case, the decision of incorporating the new drug was right, with a benefit of $17,000, even higher than the average benefit calculated ($4,200). Therefore, the value of the perfect information is 0: the same decision would have been made with or without perfect information.

Now consider that the right simulation is actually the simulation number 4. In this case, we made the wrong decision choosing the new drug, and lost $30,000 of net benefit. If perfect information existed, we would know that the parameters of simulation 4 were the right ones and the old drug would have been chosen, hence the $30,000 would not have been lost. For this simulation, the value of perfect information is $30,000.

The expected value of perfect information (EVPI) is calculated by assessing the value of perfect information for all the iterations, which will be 0 if the simulation result is the same as the average result and will be positive and correspond to the NMB of the other alternative (e.g. $30,000 for simulation number 4) if the simulation result is different from the average result. The value of perfect information for each one of the simulations is shown in the last column of Table 1. The EVPI will be the average of all the perfect information values. In our example, the EVPI is $11,100 per patient. However, the technology will not be used by only one person, so it is important to estimate

the population EVPI. In a simplified manner, considering that 1,000 patients would use the new technology, the population EVPI would be $11,100,000. A more precise calculation should consider factors such as the lifetime of the technology, the period over which the information about the decision would be useful and the incidence and mortality rates of the disease during this period [35-40]. In our example, the population EVPI of $11,100,000 can be interpreted as the cost of the uncertainty and how much should be spent in future research. As this is a high value, it is possible that the payer decides to conduct a new clinical study to reduce uncertainty and minimize the chances of losing money and benefits if the wrong decision is made. But where the new research should be focused? In transition probabilities, costs, quality of life data? The analysis called 'expected value of partially perfect information' (EVPPI) can help to elucidate that.

Expected value of partial perfect information

In order to investigate the source of the uncertainty a little further, the expected value of partial perfect information (EVPPI) analysis can be conducted. It allows to identify where and what type of research should be conducted to minimize the uncertainty. EVPPI presents a theory similar to EVPI, but each parameter is investigated at a time. In this way, it is possible to estimate the parameters whose uncertainty has the greatest influence on the decision problem, and where to direct future research (e.g. RCTs to reduce uncertainty of efficacy data, or quality of life studies to minimize uncertainty about utility data). Although the theory is remarkably like that of the EVPI,

EVPPI is much more computational consuming, which is related to the number of parameters inserted in the model [38,40,41].

It is important to mention that not always a parameter with great uncertainty will result in a high EVPPI. A parameter can be uncertain but, despite the assumed value, the impact in the net benefit difference between the alternatives can be low. For example, the probability for hospitalization in a health condition can be highly uncertain, however, whether the probability is 25% or 75% the net benefit is not affected, and the conclusion of which alternative is cost-effective does not change. Other example is a cost that varies in the same proportion on both the alternatives being compared, not affecting the difference in the net benefit. EVPPI helps identifying which parameters have uncertainty that really affects the results, because considering only the isolated magnitude of uncertainty of the parameters can be misleading [38].

It is important to bear in mind that the sum of individual EVPPIs does not equal the value of the EVPI. Sometimes the isolated parameters do not impact substantially in the net benefit, only impacting when analyzed as a whole group or as subgroups. In this case, analyzing groups of parameters can be a good strategy to identify the relevant uncertainty. For example, it can be interesting to group all the parameters related to quality of life. If there is important uncertainty in this group, the researchers can direct future studies in conducting surveys using, for instance, the EQ-5D questionnaire. It can be also a good strategy to group parameters that are correlated.

Therefore, both EVPI and EVPPI can work together to estimate the value of perfect information, i.e., how much the current uncertainty present in the cost-effectiveness model costs, and whether the available data is good enough

to assure a reliable decision. If the cost of the perfect information is high, EVPPI can be used to identify which parameter or group of parameters should be explored in further research, helping to identify which type of studies are necessary: randomized controlled trials for efficacy parameters, observational studies for natural history parameters, and so on. It is important to highlight that everything discussed here relies on the assumption that high quality evidence has been selected to inform the model. If the uncertainty comes from poor evidence, conducting value of information analysis will not identify the real impactful uncertainty and may lead to erroneous conclusions.

Software for value of information analysis

EVPI is relatively straightforward to calculate and can be performed in simple software as Microsoft Excel. EVPPI is a much more demanding analysis, as inner and outer loops are necessary: first we run a simulation with a defined value of the parameter of interest whilst the other parameters are sampled from their distributions. Following, a new value of the parameter of interest is defined, while the other parameters remain the same. After this process is repeated a sufficient number of times, another set of the remaining parameters are sampled, and new simulations are performed several times. In this case, a more powerful software should be used, such as R. A user friendly and free application to perform value of information analyses is SAVI (Sheffield Accelerated Value of Information) [42], a R Shiny Server application web tool developed by researchers from the University of Sheffield: **savi.shef.ac.uk/SAVI/**. SAVI is also available as an R package for offline analyses: **www.sheffield.ac.uk/polopoly_fs/1.511325!/file/Instructions_for_SAVI-package.txt**. After informing the results of the probabilistic sensitivity analysis, in a matter of

seconds SAVI will provide a cost-effectiveness plane and EVPI and EVPPI results. Other option that is also quite easy to use is TreeAge Pro, which allows for the complete development of the disease model and data analysis, including probabilistic sensitivity analysis, EVPI and EVPPI.

Conclusions

In this chapter, topics that are increasingly present in health decision making process, drug development and research prioritization were discussed. These topics included: network meta-analysis, a technique that enables the comparison of technologies that have not been compared head-to-head; living systematic review and model-based meta-analysis, which are increasingly present in scientific publications; and the concept of value of information analyses, a way to transform the results of probabilistic sensitivity analyses from cost-effectiveness studies in a more tangible information that can be used to identify the source of meaningful uncertainty and help designing future clinical studies. Beyond the topics presented, some other issues that are becoming more relevant each day and that can be of interest for the readers include the 'expected value of sample information' (EVSI) that goes even further in the planning of clinical studies, helping estimate the necessary sample size, and the integration of NMA and value of information analyses.

References

[1] Garg AX, Hackam D, Tonelli M. Systematic review and meta-analysis: When one study is just not enough. Clin J Am Soc Nephrol. 2008;3(1):253–60.
[2] Catalá-López F, Tobías A, Cameron C, Moher D, Hutton B. Network meta-analysis for comparing treatment effects of multiple interventions: an introduction. Rheumatol Int. 2014;34(11):1489–96.
[3] Tonin FS, Rotta I, Mendes AM, Pontarolo R. Network meta-analysis : a technique to gather evidence from direct and indirect comparisons. Pharm Pract (Granada). 2017;15(1):1–11.
[4] Dias S, Ades AE, Welton NJ, Jansen JP SA. Network Meta-Analysis for Decision-Making. Wiley, editor. Hoboken; 2018.
[5] Salanti G, Higgins JPT, Ades AE, Ioannidis JPA. Evaluation of networks of randomized trials. Stat Methods Med Res. 2008;17(3):279–301.
[6] Efthimiou O, Debray TPA, van Valkenhoef G, Trelle S, Panayidou K, Moons KGM, et al. GetReal in network meta-analysis: a review of the methodology. Res Synth Methods. 2016;7(3):236–63.
[7] Salanti G. Indirect and mixed-treatment comparison, network, or multiple-treatments meta-analysis: many names, many benefits, many concerns for the next generation evidence synthesis tool. Res Synth Methods. 2012;3(2):80–97.
[8] Bucher HC, Guyatt GH, Griffith LE, Walter SD. The results of direct and indirect treatment comparisons in meta-analysis of randomized controlled trials. J Clin Epidemiol. 1997;50(6):683–91.
[9] Hasselblad V. Meta-analysis of multitreatment studies. Med Decis Mak. 1998;18(1):37–43.
[10] Tonin FS, Steimbach LM, Mendes AM, Borba HH, Pontarolo R, Fernandez-Llimos F. Mapping the characteristics of network metaanalyses on drug therapy: A systematic review. PLoS One. 2018;13(4):1–16.
[11] Lumley T. Network meta-analysis for indirect treatment comparisons. Stat Med. 2002;21(16):2313–24.
[12] Lu G, Ades AE. Combination of direct and indirect evidence in mixed treatment comparisons. Stat Med. 2004;23(20):3105–24.
[13] Sobieraj DM, Cappelleri JC, Baker WL, Phung OJ, White CM, Coleman CI. Methods used to conduct and report Bayesian mixed treatment comparisons published in the medical literature: A systematic review. BMJ Open. 2013;3(7).
[14] Caldwell DM. An overview of conducting systematic reviews with network meta-analysis. Syst Rev. 2014;3(1):1–4.
[15] Chaimani A, Higgins JPT, Mavridis D, Spyridonos P, Salanti G. Graphical Tools for Network Meta-Analysis in STATA. PLoS One. 2013;8(10).
[16] Chung H, Lumley T. Graphical exploration of network meta-analysis data: The use of multidimensional scaling. Clin Trials. 2008;5(4):301–7.
[17] Rücker G, Schwarzer G. Automated drawing of network plots in network meta-analysis. Res Synth Methods. 2016;7(1):94–107.
[18] Higgins JPT, J T, J C, M C, T L, MJ P, et al. Cochrane Handbook for Systematic Reviews of Interventions version 6.1 (updated September 2020). Cochrane; 2020.

[19] Cipriani A, Higgins J, Geddes JR, Salanti G. Research and Reporting Methods Annals of Internal Medicine Conceptual and Technical Challenges in Network Meta-analysis. Ann Intern Med. 2013;159(8):130–7.
[20] Dias S, Welton NJ, Caldwell DM, Ades AE. Checking consistency in mixed treatment comparison meta-analysis. Stat Med. 2010;29(7–8):932–44.
[21] Lu G, Ades AE. Assessing evidence inconsistency in mixed treatment comparisons. J Am Stat Assoc. 2006;101(474):447–59.
[22] Jansen JP, Naci H. Is network meta-analysis as valid as standard pairwise meta-analysis? It all depends on the distribution of effect modifiers. BMC Med. 2013;11(1).
[23] Borenstein M, Hedges LV, Higgins JP RH. Introduction to Meta-Analysis. Chichester, UK: John Wiley & Sons, Ltd; 2009.
[24] Jansen JP, Crawford B, Bergman G, Stam W. Bayesian meta-analysis of multiple treatment comparisons: An introduction to mixed treatment comparisons. Value Health. 2008;11(5):956–64. A
[25] Shim SR, Kim S-J, Lee J, Gerta Rücker. Network meta-analysis: application and practice using R software. Epidemiol Health. 2019;41:e2019013.
[26] Aickin M. Bayes without priors. J Clin Epidemiol. 2004;57(1):4–13.
[27] Goodman SN. Toward Evidence-Based Medical Statistics. 1: The. Ann Intern Med. 1999;130:995–1004.
[28] Brooks SP, Gelman A, Journal S, Statistics G, Dec N. Interface Foundation of America General Methods for Monitoring Convergence of Iterative Simulations of Mathematical Statistics, and Interface Foundation of America Stable. J Comput Graph Stat. 1998;7(4):434–55.
[29] Salanti G, Ades AE, Ioannidis JPA. Graphical methods and numerical summaries for presenting results from multiple-treatment meta-analysis: An overview and tutorial. J Clin Epidemiol. 2011;64(2):163–71.
[30] Boutron I, Chaimani A, Meerpohl JJ, Hrobjartsson A, Devane D, Rada G, et al. The COVID-NMA Project: Building an Evidence Ecosystem for the COVID-19 Pandemic. Ann Intern Med. 2020;173(12):1015–7.
[31] Borm GF, Donders ART. Updating meta-analyses leads to larger type I errors than publication bias. J Clin Epidemiol. 2009;62(8).
[32] Simmonds M, Salanti G, McKenzie J, Elliott J. Living systematic reviews: 3. Statistical methods for updating meta-analyses. J Clin Epidemiol. 2017;91:38–46.
[33] Brooker J, Synno A, McDonald S, Elliott J, Turner T, Hodder R, et al. Guidance for the production and publication of Cochrane living systematic reviews: Cochrane Reviews in living mode. 2019.
[34] Upreti V V., Venkatakrishnan K. Model-Based Meta-Analysis: Optimizing Research, Development, and Utilization of Therapeutics Using the Totality of Evidence. Clin Pharmacol Ther. 2019;106(5):981–92.
[35] Boucher M, Bennetts M. The many flavors of model-based meta-analysis: Part I - Introduction and landmark data. CPT Pharmacometrics Syst Pharmacol. 2016;5(2):54–64.
[36] Mawdsley D, Bennetts M, Dias S, Boucher M, Welton NJ. Model-based network meta-analysis: A framework for evidence synthesis of clinical trial data. CPT Pharmacometrics Syst Pharmacol. 2016;5(8):393–401.
[37] Cooper H, Hedges L V., Valentine JC. The Handbook of Research Synthesis and Meta-

Analysis. Russell Sage Foundation Publications; 2009.
[38] Briggs A, Claxton K, Sculpher M. Decision modelling for health economic evaluation. Oxford: Oxford University Press; 2011.
[39] Thorn J, Coast J, Andronis L. Interpretation of the Expected Value of Perfect Information and Research Recommendations: A Systematic Review and Empirical Investigation. Med Decis Mak. 2016;36(3):285–95.
[40] Felli JC, Hazen GB. Sensitivity analysis and the expected value of perfect information. Med Decis Mak. 1998;18(1):95–109.
[41] Claxton KP, Sculpher MJ. Using value of information analysis to prioritise health research: Some lessons from recent UK experience. Pharmacoeconomics. 2006;24(11):1055–68.
[42] Strong M, Oakley JE, Brennan A. Estimating multiparameter partial expected value of perfect information from a probabilistic sensitivity analysis sample: A nonparametric regression approach. Med Decis Mak. 2014;34(3):311–26.

ADVANCED STATISTICAL CONCEPTS: NETWORK META-ANALYSES AND VALUE OF INFORMATION

LETICIA P. LEONART

FORMULATION OF FINDINGS AND RECOMMENDATIONS

DANIELA O. DE MELO
PATRICIA M. AGUIAR
TÁCIO DE M. LIMA

Abstract

The wide interest in publishing and disseminating clinical practice guidelines (CPGs) reflects the need for better allocation and rational use of the scarce resources available. CPGs also has a contribution to the management and regulation of health systems, as its implementation is expected to promote better quality and equity in health care and could potentially improve patient outcomes by encouraging evidence-based decision making, influencing choices so that the most cost-effective interventions are applied in the day-to-day health systems and services. In this context, it is important to be aware of the likely facilitating factors and barriers during the CPG development. The process of CPG development can be summarized in several steps, involving different groups and professionals. This chapter will present each of the points, such as the theme and scope choice, guideline working group, conflict of interest, evidence, and recommendations as well as the quality, implementation, adaptation, and up-to-date of CPGs.

Keywords: practice guideline; health plan implementation; quality of health care

FORMULATION OF FINDINGS AND RECOMMENDATIONS

DANIELA O. DE MELO
PATRICIA M. AGUIAR
TÁCIO DE M. LIMA

Introduction

The wide interest in publishing and disseminating clinical practice guidelines (CPGs) reflects the need for better allocation and rational use of the scarce resources available, considering a) the increasing costs with health (due an increase in demand as a consequence of population aging, among other factors, and the availability of more expensive technologies in the market); b) the finding that there are variations in service provision among providers, hospitals and geographical regions; and c) the intrinsic desire of health professionals to offer, and of patients to receive, the best possible care. Thus, CPGs also has a contribution to the management and regulation of health systems, as its implementation is expected to promote better quality and equity in health care and could potentially improve patient outcomes by encouraging evidence-based decision making, influencing choices so that the most cost-effective interventions are applied in the day-to-day health systems and services.

CPGs are statements that assist health professionals in making decisions about a particular clinical situation. However, for CPGs recommendations to be implemented by these professionals, it is necessary to be aware of the likely facilitating factors and barriers during their development. One of the factors that can influence a professional's decision to implement a recommendation is the degree of confidence he has in it; that is, the certainty that following the recommendation will produce the expected improvement in the outcome for its patients. This certainty is not only related to the degree of confidence in the effect size of an intervention in relation to important specific results (evidence), but also covers other issues such as patient preferences and the availability of resources to support the introduction of a new intervention. For this reason, the policy development group must consider both the overall quality of supporting

evidence and the other factors that may influence the strength of the recommendation.

This chapter aims to discuss the main factors related to the development of CPG and recommendations, including the theme and scope choice, guideline working group, conflict of interest, evidence, and recommendations as well as the quality, implementation, adaptation, and up-to-date of CPGs.

Importance of clinical practice guidelines

CPGs are documents with recommendations for health professionals. They can be elaborated and/or published by any type of organization (government, professional societies, group of authors, services, or health systems, being public or private). Their main objective is to make the actions of their professionals or services more predictable, and presumably of higher quality since they are supported on the best evidence available [1-5]. There is no reason to fear that the implementation of CPGs may lead to inflexibility and restriction of the possibility of individualization of care because CPGs are documents with recommendations and not impositions - the final decision is always made by a health professional, being this the difference between CPG and clinical protocol [6].

Among the benefits of implementing CPGs in clinical practice are: a) the potential influence in the adoption or promotion of public policies to ensure access to interventions with evidence of effectiveness, safety and cost effectiveness while promoting disinvestment in low-value technologies; b) the incentive to promote equity and rational use of financial and/or human resources, almost always scarce in health systems; c) the compilation of practical and explicit recommendations, which assist health professionals in their training or

updating; d) potentiality of use for medico-legal protection or for health professionals to have support when dealing with managers; and e) versions addressed to patients to help them in the shared decision making (SDM) process [6]. SDM is an approach that foresees the sharing of leading role in the decision making about care between health professionals and patients [7,8]. It has gained prominence in health care policy especially in situations for which the evidence is scarce or inconclusive [7].

The main criticism to CPGs is that they would be directed to diseases and not to people, which is a problem in view of the aging population and the increasing prevalence of multimorbidity [2,9,10]. There is also concern that recommendations may be influenced by opinions, clinical experience or even conflicts of interest of the elaboration group, being even more serious in areas with scarcity or low quality of evidence to support recommendations [11]. Currently, much has been discussed about the relevance of considering patients' needs and preferences in formulating recommendations since shared decision making is often encouraged. In the same way that CPG can positively influence public policies and support health professionals in auditing situations, deliberation with managers or even legal actions, the dissemination of CPG elaborated without transparency and methodological rigor and/or outdated can be harmful [6].

In the last twenty years, the number of CPGs has increased exponentially while there has been a broad discussion about the need for CPGs to be developed not only on the basis of expert consensus but also as the product of a systematic and transparent literature review process to formulate recommendations based on the best available evidence. But CPGs also consider other factors besides the evidence on the effectiveness and safety of interventions, such as the feasibility

of their implementation, the cost of interventions and/or the budgetary impact of their incorporation, the acceptability by health professionals and patients, and others. Therefore, CPGs are different documents of systematic reviews and, depending on the local context, different drafting groups may formulate different recommendations supported by the same evidence. These documents must be reliable and therefore it is essential that only high-quality CPGs are published, i.e., those prepared with methodological rigor, with transparency and management of potential conflicts of interest, considering the position of stakeholders [12].

Quality of clinical practice guidelines

The Institute of Medicine (IOM) and the Guidelines International Network (G-I-N) have proposed quality standards for CPGs, having in common the fact that quality has been directly associated with the transparency of the document preparation process [1,2]. Several instruments have also been developed to assess the quality of these documents, but The Appraisal of Guidelines for Research and Evaluation (AGREE) is the most widely used internationally, being validated and translated into several languages [13-15]. It is important to emphasize that the AGREE tool is the result of a careful study of the literature on critical evaluation of clinical guidelines and that it mirrors the quality standards established by the IOM and G-I-N, contemplating all the aspects recommended by these two important institutions [16].

The second version of AGREE was published in 2009 (AGREE II) and it is available on the website **http://www.agreetrust.org**

A user's manual is also available on this website plus support for training and conducting appraisal on the platform 'My Agree Plus'. The AGREE II

consists of a 23-item tool, covering six quality domains with a Likert scale from 1 to 7 (1: I totally disagree with 7: I totally agree) for scoring each item. The six domains are:

- Scope and purpose
- Stakeholder involvement
- Development rigor
- Clarity of presentation
- Applicability
- Editorial independence

Each domain receives a score that varies from 0 to 100%, representing how much the CPG report allowed identifying or not the items considered essential. In addition to the six domains, it includes two final questions that are considered the overall evaluation of CPG: a) a judgment of the directive's quality, considering the criteria considered in the evaluation process; and b) whether the evaluator recommends the use of CPG, whether or not it is recommended [15].

The AGREE II manual does not indicate that a domain may be more relevant than others nor does it establish a specific cut-off to determine which CPG has or does not have quality. A recent article by the AGREE group reinforces this idea justifying that main reason for that is that there is no evidence to define it. They suggest that users discuss and decide on threshold take in account each context and being transparent regarding this, if deemed necessary [15]. In studies that evaluate the quality of CPGs, the cut-off most employed have been 50% and 60% and the domains considered most relevant, according to the authors of this type of study, would be the rigor of development (domain 3), editorial independence (domain 6) and applicability (domain 5) [14,17].

DANIELA O. DE MELO
PATRICIA M. AGUIAR
TÁCIO DE M. LIMA

Two metareviews of studies that evaluated the methodological quality of CPGs reported a large proportion of documents with unsatisfactory quality. Although there seems to be an improvement in quality over time for some domains, those domains concerning transparency of the eligibility process, selection, evidence analysis and consequent formulation of recommendations (domain 3), implementation issues (domain 5), and adequate description and management of potential conflicts of interest (domain 6) still require more attention [18,19].

According to a systematic review of 421 CPGs published between 2011 and 2017, containing recommendations for the pharmacological treatment of chronic non-communicable diseases, only 99 (23%) of CPGs presented a score equal to or higher than 60% in domain 3 of AGREE II. The factors associated with the highest score in this area were government involvement in the development of CPGs (either through a program of development of these documents, their funding and/or supervision); reporting by the funding of CPGs; and a greater number of authors [16]. In fact, developing high-quality CPGs requires substantial resources (time, investment and professionals with specific skills and competencies) and depends on a process that involves many methodological and discussion steps.

Development of clinical practice guidelines

The process of CPG development can be summarized in several steps, involving different groups and professionals, as shown in Figure 1.

DANIELA O. DE MELO
PATRICIA M. AGUIAR
TÁCIO DE M. LIMA

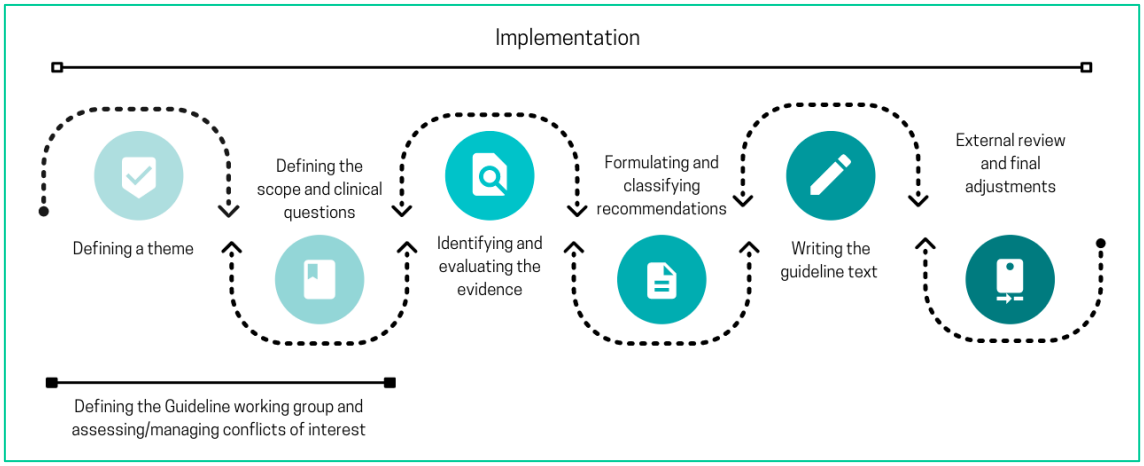


Figure 1. Steps for development of clinical practice guidelines

Defining a theme for CPG

The definition of the CPG theme is frequently performed by those who commission the elaboration of the document - such as the healthcare managers, for example. When defining the theme, it is recommended to follow a prioritization process so that the resources employed in the development of the document can bring the maximum benefit to the health system or service. CPGs should deal mainly with situations for situations where there is uncertainty about best practices and/or potential to improve health outcomes or make better use of resources or to reduce health inequalities [4,5,20].

Recently, a systematic review identified and discussed the different prioritization approaches in the development of CPGs [21]. The involvement of multiple stakeholders and the use of prioritization criteria were the key aspects proposed to be addressed when prioritizing a theme for CPG. The types of stakeholders most cited to be involved in this process were patient representatives, clinicians, experts, along with members of guideline developing organization. In addition, the authors derived a common framework with 20

prioritization criteria clustered in six domains including: disease-related factors, interest, practice, guideline development, potential impact of the intervention, and implementation considerations. The most often reported prioritization criteria were related to the health burden, practice variation, and impact on health outcomes. Finally, the authors considered that the wide variability of prioritization approaches necessitates that researchers develop standardized and validated priority setting tools.

Defining the scope of CPG and developing clinical questions

With the definition of the theme, the next step is the clear scope of the CPG. This is one of the most critical steps of CPG elaboration because it involves structural and management aspects, such as project size and time for execution, as well as clinical aspects (the extent to which it is at this stage that clinical questions must be defined, which may be influenced by characteristics of the disease, number of interventions to be evaluated and even the target user of the CPG) [1,22]. For the scope of the CPG to reflect the real needs of the health system or services, this discussion must include stakeholders - both specialists in the field, methodologists and managers, and patient representatives [2,23].

Few protocols consistently describe steps regarding the key clinical questions for CPG [24]. In this sense, a recent study created a protocol to develop relevant clinical questions for CPG. The authors reviewed the identified ten guideline development manuals and extracted from them instructions for developing clinical questions and established seven steps of the protocol:

- Define the rationale for the guideline
- Use qualitative research methods to determine the initial list of key questions based on the clinical challenges faced by target end-users

- Convert the initial list of questions
- Specify all relevant outcomes for each possible question
- Review and revise draft key questions
- Rate the outcomes in order of importance for clinical decision-making
- Decide on the final list of questions

From there, they developed ten clinical questions related to work-related mental health conditions successfully. Therefore, this protocol can be used for CPG developers to create relevant clinical questions [24].

Guideline working group

A guideline working group must be multidisciplinary and geographically representative, with participation from all relevant professional groups, as well as lay members [20,25]. The size and composition of the working group will depend on some factors, such as the scope, complexity, and the time spend to the guideline development [26]. In addition, the working time of the group varies depending of the type of project; the time spend to developed a new guideline is greater compared to update guideline or minor review [25].

Members of the working group are responsible to advise the organization that ordered the CPG about the scope of CGP and evaluated the evidence synthesis, including the comprehensive search, selection, critical analysis, and advice on the interpretation of results, and the formulation of recommendations. In addition, this group can be evaluated and incorporated suggestions from external reviews [5,27].

It is important the selection of the leader of the working group. Usually, the organization that ordered the CPG select a person with ability to manage

meetings effectively, negotiate with members, and resolve conflict. Moreover, the leader must have knowledge in all areas of the guideline development process to interact with each of the participants of the working group [5,25-27].

All members must make a full commitment to the working group and the tasks involved in guideline development, communicated any concern to the leader, as well as be prepared to consult the group to ensure the widest possible range of views, maintaining confidentiality of discussions conducted within the group. Overall, the meetings of the working group are on average once every two to three months and the subgroups of each specific area can meet more frequently [25].

Conflict of interest

The conflict of interest (COI) is defined as a set of conditions that creates a risk that professional judgment or actions regarding a primary interest (such as patients' welfare or the validity of research) will be unduly influenced (consciously or unconsciously) by a secondary interest (such as financial gain) [5,28].

A Chinese study showed that only 26% of CPG disclosed COIs [29]. In other hand, Saleh el at. reported that 85% of the first and last authors of the CPG published on the American Society of Clinical Oncology's website received payments from industry, whereas 32% did not disclose these payments [30]. Thus, it is important to ensure that conflict of interest will not be a potential source of bias and diminished credibility of the guideline development process [5]. A guideline with conflict of interest can recommends new, expensive, and less effective/safety treatments or products [26].

DANIELA O. DE MELO
PATRICIA M. AGUIAR
TÁCIO DE M. LIMA

All members involved in the guideline development process must complete and sign a declaration of interests form in a proactive, reasoned, transparent and defensible manner before invitations of work meetings and each declaration must be made available to the working group. The leader of the working group must have no direct COI or relevant indirect COIs [31]. Moreover, it is essential to ensure that the majority of working group members have no COI. Any change of COI must during the guideline development process be registered and shared to all members of the groups [27].

Initially, the organization that ordered the CPG develop a clear COI policy and how will be managed, assess the declaration of interests, and determines if a COI exists that can be the risk of adversely affecting the guideline development process [5,26,31]. According to degree of severity of COI, the steering and working group can make decisions, as shown below [5,27,31,34]:

- **Not relevant:** the member can participate of all guideline development process because the potential conflict of interest does not have direct relationship with the topics covered by the proposed guideline (e.g. academy members).
- **Important:** the member can participate of the elaboration process, but will not be able to participate oi the formulation of recommendations (e.g. member that depend economically on the use of technology that can be recommended by guideline; receive research funding from companies with commercial interests in the guideline; has a family member who works for a company that makes a product or technology that can be recommended for use in the guideline; or has current or past involvement in a clinical trial that recommends using a product or technology that can be considered in a guideline recommendation.

- **Very important:** the member must not be included in the guideline development process or must be excluded if the process has started (e.g. member works or has shares in a company that manufactures a product or technology that can be recommended for use in the guideline; has a patent for a product or technology that can be recommended for use in the guideline; be a consultant, paid speaker, or opinion leader for a company or organization with an interest in a specific product related to the guideline; has received any support for scientifical courses, congress travel or similar from a company or organization with an interest in a specific product related to the guideline).

All declarations of interest must be published on the final report of the guideline. The information must describe what the conflicts were, and the strategy used to manage each of them. In addition, it is recommended that this information be included when the guideline is released for public consultation [26,27].

Identifying and evaluating the evidence

After defining clinical questions, the next step is a systematic review to summarize the best available evidence to answer each of the defined clinical questions. For this, it is important that the question is well thought out - in general, using PICO - followed by a well-designed search strategy, study eligibility, and assessment of the quality/risk of bias of individual studies. The production of an evidence-based guideline requires a critical appraisal of the literature relevant to its scope. There are many grading systems, but it is recommended to use GRADE approach (Grading of Recommendations,

Assessment, Development and Evaluation) to assess the quality of the evidence incorporated in CPGs [33] (see Chapter 04). The systematic and transparent approach to making judgments about the quality of evidence helps avoid errors, facilitates critical evaluation of the guidelines, and improves communication of this information to health professionals, the population, and managers.

As explained in Chapter 04, the GRADE system discriminates four categories of quality of evidence: high, moderate, low, and very low. These categories are applied to a body of evidence for each outcome of interest and not to individual studies. The quality assessed reflects the confidence in the effect estimates and whether they are sufficient to support a given recommendation. Starting from an assumed level of high quality for the randomized controlled clinical trials, the quality of the evidence is downgraded due to five factors: risk of bias, unexplained heterogeneity, indirectness of evidence, imprecision of the pooled estimate, and publication bias. On the other hand, observational studies start with low quality, but there are factors that are especially relevant to this type of study and may lead to rating up quality of the evidence, such as magnitude of effect, dose-response gradient, and residual confounding factors [34].

The quality assessment should be detailed using GRADE evidence profile tables, which will support the final decision on the quality of the evidence body for each outcome and make the quality classification explicit and reproducible [35]. For that, you can use the GRADEpro computer software, which is an official GRADE tool available online for free on the website. It is important to note that in the GRADE tables, only outcomes previously judged as critical and important but not critical should be inserted. The GRADE system classifies the relative importance of outcomes in three categories, ranking them on a scale of 1 to 9: critical (9 to 7), important but not critical (6 to 4) and those of limited importance

(3 to 1). Defining the types of outcomes by their relative importance can help to focus attention on those that are considered most important and help to resolve or clarify disagreements [22]. After judging the quality of evidence for each outcome of interest across studies, guideline developers should then come to a final decision regarding the rating of overall quality of body of evidence for the same clinical question, being attributed according to the lowest quality among the critical results [35].

Formulating and classifying recommendations

Having rated of overall quality of body of evidence for the same clinical question, the next step is to determine the direction and strength of recommendations and formulate the recommendations. The broader the scope and target audience, the more complex the decision-making process can be. The formulation of recommendations encompasses a larger number of factors besides scientific evidence during the formulation of recommendations, including normative and technical criteria that require careful evaluation. Thus, it is essential that transparent and systematic processes are established for discussion by the elaborator group, which has driven the development of Evidence to Decision (EtD) frameworks, i.e., tools that propose an organizational structure for this step, guiding the passage of evidence discussion to decision making. The use of a framework is recommended mainly when evidence is scarce or insufficient [36].

The efficacy and cost of the intervention are generally the criteria that are identified as essential in health care decision making, but there are many others, ranging from those most commonly discussed, such as issues involving the feasibility of implementing the recommendation, to more complex ones, such as

potential violation of individual rights or interactions of interventions with other components of a health care system, as well as equity aspects [37,38]. Accountability for reasonableness makes it possible to guide all actors on deliberating on fair decisions, also considering resource constraints. This enables the connection between decision-making in health services and the broader, more fundamental democratic deliberative processes.

- **Can the same evidence generate different recommendations?** Yes! Guidelines are different of systematic reviews. When making recommendations, there is no direct application of the evidence but an interpretation, which involves other aspects such as patient's values and preferences, and available resources. A guideline uses information from evidence review to make recommendations for a specific context, always linking the strength of recommendation to the quality of evidence. Therefore, different guideline developers may generate conflicting recommendations, even if they follow systematic and transparent development processes [39,40].
- **One of the most emblematic cases of guidelines with conflicting recommendations has been with regard to the diagnosis and therapeutic goals for hypertension.** In this case, the question does not involve the local context but the weight each guideline gave when considering the SPRINT study in the recommendations. From the SPRINT results, the 2017 American College of Cardiology/American Heart Association (ACC/AHA) guideline has recommended stricter parameters than the ones adopted [41,42].

During the course of the DECIDE project, the GRADE working group developed EtD frameworks to systematize different types of health decisions, including clinical recommendations [43]. Web platforms for structuring the assessment of the certainty of evidence (GRADEpro) and for formulating recommendations (GRADEpro Guideline Development Tool) were also developed, with the purpose of making it easier to summarize and present the

information used in health care decision making based on the methods proposed by this group. At the webise **https://guidelines.gradepro.org/search** you can consult DCs prepared using the GRADE EtD.

According to the GRADE approach, the direction of the recommendations is defined by the balance between desirable results (benefits) and undesirable results (damages) for all critical and important but not critical outcomes of a given strategy, in relation to the comparator. The balance between desirable and undesirable consequences is also called net benefit and involves judging the relative importance of these consequences which is affected mainly by the patient values and preferences. In addition, the strength of a recommendation can be understood as the confident of the balance between desirable and undesirable consequences, that is, if the guideline panel is highly confident of the balance of the consequences, they make a strong recommendation and if they are less confident, they offer a weak recommendation (also known as conditional, discretionary, or qualified). Thus, the GRADE approach classifies the recommendations into four categories: strong or weak recommendations for or against a particular strategy [44]. The term 'conditional recommendation' has replaced 'weak recommendation' to improve understanding that its implementation depend on circumstances such as patient values, resource availability or other contextual considerations. Conditional recommendations signal situations for which evidence is scarce or insufficient, or when the evidence suggests that there is no advantage with one of the evaluated alternatives. In these cases, it is expected that the shared decision process will be implemented [44,45].

Although the World Health Organization (WHO) employs the GRADE method in assessing the certainty of evidence, it has proposed its own EtD

framework, WHO-INTEGRATE. It is a framework rooted in global health norms and values, with the scope focused on an integral understanding of health problems. While GRADE EtD is broadly congruent with WHO norms and values, concerns were raised that the decision criteria within GRADE EtD may not be sufficiently complete and useful for decisions about complex interventions and/or the complex systems in which they are implemented; that it did not sufficiently present the central role of the social and economic determinants of health; and regarding consistency in the application of GRADE EtD frameworks within WHO guideline development processes [38].

The formulation and judgement of the strength of recommendations is a complex process and requires a carefully selected and role-minded guideline development group. Recently, the GRADE group published a support tool for participants in guideline development groups, based on the assumption that there is a need for orientation and training of members who often have no previous experience with CD development or teamwork processes [46]. A recent study identified challenges experienced guideline developers and others using GRADE in public health-related contexts [47] - which shows that using this system is not an easy task. Five priority issues can pose challenges for GRADE users: incorporating the perspectives of diverse stakeholders; selecting and prioritizing health and 'non-health' outcomes; interpreting outcomes and identifying a threshold for decision-making; assessing certainty of evidence from diverse sources, including non-randomized studies; and addressing implications for decision makers, including concerns about conditional recommendations. The authors illustrate the challenges with examples from CPGs, identifying gaps, and planned the further elaboration of GRADE guidance to address these topics.

DANIELA O. DE MELO
PATRICIA M. AGUIAR
TÁCIO DE M. LIMA

Writing the guideline text

The working group must write a draft guideline, containing the consensual final result and recommendations made, and a detailed description of the entire process of guideline development. The final size of the guideline will depend essentially on the number of questions to be answered and the volume of evidence available on the subject, as well as the complexity of it. Guidelines with many clinical issues can be subdivided in chapters.

Rosenfeld & Shiffman proposed tips for writing guidelines as: do not write a chapter or review article but the length should be modest; define words, phrases or actions with specificity to ensure that the reader understands the recommendation; start with the particular key actions statement and after offer some justification with a few paragraphs summarizing the evidence discussing any potential risks, harms, and costs related to the particular recommendation; end with any suggestions for future research on this topic based on evidence gaps [40]. The Essential Reporting Items for Practice Guidelines in Healthcare (RIGHT) can support guideline developers, providing a guide regading how each item should be clearly presented and sufficiently detailed somewhere in the guideline. RIGHT is a checklist with 22 items considered essential for good reporting of practice guidelines, encompassing the following domains: basic information (items 1 to 4), background (items 5 to 9), evidence (items 10 to 12), recommendations (items 13 to 15), review and quality assurance (items 16 and 17), funding and declaration and management of interests (items 18 and 19), and other information (items 20 to 22) [48].

Recommendations are the core components of guidelines and many users read only the recommendations instead of the full text of guideline. So, writing the recommendations is one of the most important steps in developing a clinical

guideline and key recommendations should be easily identifiable, preferably on first page of the guideline [49]. The wording must be concise, unambiguous and easy to translate into clinical practice, focusing on the action that needs to be taken and/or emphasizing the need of shared decision. Whenever possible, start the recommendation with a verb [20]. Studies showed that the clarity and specificity of recommendation could significantly influence decision making among health professionals [50,51]. The working group of RIGHT is developing one extension of RIGHT for reporting recommendations (RIGHT-R). So, a special research team have systematically analyzed the recommendations in order to provide an initial and important insight into the items to be considered for the eventual checklist of reporting recommendations as the suggestion on recommendations should be listed in tables, companied by strength of recommendation and quality of evidence [49].

The technical language of guidelines can difficult its understanding. Therefore, derivative documents with other versions of guideline in a more easily understandable and usable format for patients and the public should be developed. Patient or public versions of guidelines can increase patient confidence and facilitate shared-decision process. In addition, algorithm map, summary tables of drugs and any other potential material also can be helpful in making the guideline more understandable and practical [52].

External review and final adjustments

External reviewers may provide useful feedback on clarity of text and/or how easily recommendations may be adopted in health services, for example. They should comprise representants of all relevant stakeholders including organizations, patients or their representatives and scientific and clinical experts.

Furthermore, inviting commentary from key stakeholders can influence the guideline acceptance and increase engagement of more professionals [2,53]. External review can be performed in two ways: peer review (reviewers are invited considering their ability to contribute) or public consultation (publishing the document and letting it open to comment from any interested party) [54].

Preferably, reviewers should provide clear directives for their criticism, including the requirement of evidence and citations as substantiation. The development group should consider the potential necessity for revisions, and an additional post review meeting, when planning the original timeline, and to adopt a systematic process for responding to reviewer comments. The public availability of such information is important to transparency, so a brief summary of the external review process can be provided [2,54].

Implementation and dissemination of practice guidelines

The implementation of guidelines is a big challenge. Several studies show that applicability (the domain of AGREE tool for implementation issues) is always the domain with lower score [16,18,19]. The difficulty in implementing guideline recommendations in clinical practice can be explained by several factors, such as: a) knowledge or attitude of health professionals; b) opposition from key stakeholders or professional skepticism about the relevance of the document; c) lack of ready access to CPGs or lack of understanding about CPG; d) lack of clarity on operational guidelines or roles and responsibilities for implementation; e) patient noncompliance; f) lack of institutional infrastructure, heavy workload, social norms, limited resources, and coordination at the system level; g) unwillingness to change practices [55-57]. Knowing this difficulty, efforts have been intensified to ensure that evidence is effectively understood and

implemented in health care practices. Maybe the biggest challenge is likely to be building a bridge among researchers, policymakers, implementers, and target users for planning or evaluating implementation efforts [58,59].

To implementation be successful, questions of feasibility must permeate the entire process of guideline development and recommendation formulation, comprising the identification of barriers/facilitators, selection or development of tools or strategies according to the issue of the guideline, its target population, and available resources in order to optimize the implementation process [4,60].

According to the WHO, the basic steps for implementing a guideline can be summarized as [5]:

- Analyze the local needs and priorities (current practice)
- Identify all potential barriers and facilitating factors
- Determine the available resources
- Design and implementation strategy (adopting the recommendations and using the context for the proposed changes

Importantly, the implementation should be customized to the scenario in which it is desired to implement the guideline [60,61].

A recently published review reported that even countries with established guideline program seldom include implementation strategies in their documents – 15 of the 20 guidelines analyzed identified barriers to guideline implementation but none did provide any guidance on how to identify and solve the barriers on daily life of the services [62] Another study already reported that the process of defining strategies based on relevant barriers did not change over time even in the face of the publication of several models, theories, taxonomies, and frameworks aimed at improving implementation [63]. But, conducting audits to

assess whether recommendations have been followed with subsequent feedback to staff is a widely used strategy based on various theories of behavior change [63].

Since 2009, the National Institute for Health and Care Excellence (NICE) was responsible for developing and maintaining quality indicators within the Quality Outcomes Framework. Then, NICE guidelines almost always include strategies and tools that support its implementation or even measures of health outcomes to evaluate whether guideline implementation has improved the quality of care.64 In 2015, a working group of the Guideline International Network (G-I-N) published a set of standards for reporting on guideline-based performance measures, including quality indicators [65]. However, there is not yet a gold standard for developing quality indicators based on guideline recommendations, and care should be taken when proposing that evaluations of health services or professional performance be linked to compliance with guideline recommendations [66]. Conducting studies to evaluate guideline implementation is challenging. While there has been rapid progress in the development of frameworks, theory, concepts, terminology, measures, and reporting standards, it is still common for these studies to have important limitations mainly because lack standard terminology, for example. Thus, Wolfenden et al have proposed a guide to assist researchers in planning and conducting this type of study [67].

Adapting guidelines

The development of high-quality guidelines requires human and financial resources and a lot of time. It is estimated that the direct costs of guideline development can be around $200,000 per guideline in the United States [68-70].

At the same time, most organizations are under pressure to produce more documents in less time at the lowest possible cost [71]. Among the alternatives to the development of a de novo guideline ('from scratch') are the adoption, contextualization, adaptation of recommendations of available guidelines, as long as they are current and systematically developed [72].

- **ADOPTION:** it is possible to 'adopt' a guideline and implement its recommendations fully, without any changes. But this is unusual because it depends on a great deal of similarity between the local contexts where the 'parent' guideline was developed and where it is intended to be implemented.
- **CONTEXTUALIZATION:** in this case, despite the decision to adopt a "parent" guideline, additional information is required to support the effective implementation of the recommendations in the new context. In general, implementation issues such as local workforce, training, health systems, equipment, and/or access to services need to be addressed.
- **ADAPTION:** When it is not possible to find one guideline that meets all the demands, i.e. answers all the clinical questions defined in the scope, a possible way forward is to adapt recommendations from multiple guidelines. It may involve additional work to seek local research, or obtain local consensus, on how best to adapt the recommendations.

Source: adapted from Dizon et al., 2016.

Adapting a guideline appears attractive because it can be seen as a way of reducing development costs and avoiding duplication of effort and funding. However, there is limited evidence showing that guideline adaptation could saves time or money compared to developing one new guideline [73,74].

Adaption of guideline comprises a systematic process to the endorsement and/or modification of one or more guidelines 'parentals' for application in a different setting. In fact, the process of guideline adaptation demands as much methodological rigor and transparency as the elaboration of a de novo guideline. We can highlight similarities between the development of an original guideline

and adaptation of guideline such as the need to respect the principles of evidence-based health care and the adoption of consistent and reliable methods to review the available evidence; a participatory approach with the involvement of all stakeholders and the consideration of the local context to increase the chances that the recommendations will be implementable [74-76].

There are at least eight frameworks to guide a guideline adaptation process: ADAPTE; The Adapted ADAPTE by the Alexandria Center for Evidence-Based Clinical Practice Guidelines (EBCPG); Alberta Ambassador Program; GRADE-ADOLOPMET; Making GRADE the Irresistible Choice (MAGIC); RAPADAPTE; Royal College of Nursing (RCN); Systematic Guideline Review Method (SGR). The Adapte and GRADE-Adolopment methods have stood out, one for being the pioneering model and most employed, and the other for structuring the process based on the GRADE method, allowing it to encompass adopted, adapted or newly developed recommendations [74,75,77].

Keeping a guideline up-to-date

Just as important as developing a guideline is keeping it up to date. The speed at which new evidence emerges that may generate the need to update a guideline is very variable depending on the clinical condition and care setting, so there is no consensus on how long a guideline is considered current. For example, during the pandemic of COVID-19, there was a need for rapid development of guidelines and constant updating as new evidence came to light, since after days or weeks the document could already be out of date. In general, guidelines developing programs endorse three years as a reasonable time period to review their documents, however the most groups manage to do every five years [72,78,79].

DANIELA O. DE MELO
PATRICIA M. AGUIAR
TÁCIO DE M. LIMA

Clinical practice guidelines become outdated as new evidence is published and require periodic reassessment to remain valid. The principles for updating CD can be complex, especially resource-intensive, and include three main steps [78-81]:

- Identify new evidence
- Assess whether the new evidence has a potential impact on the current guideline
- Revise and modify the guideline

Preferably, the monitoring of the need for updating and the process of updating guidelines should be carried out by the same group responsible for drafting the first version of the document. The guidelines' methodological manuals include very little guidance on how to revise and update the guidelines. But the CheckUp tool compiles a 16-item checklist that assesses the reporting of the guideline update process [80].

Conclusions

Healthcare professionals and managers are daily surrounded by health information of different sources and CPGs of several organizations which are often of non-reliable and of poor quality. This chapter highlights the importance of scientific rigor for the development or adaptation of CPGs and their implementation in clinical routine, which should also consider healthcare professionals experiences, the availability of resources and patients' preferences. The summarized discussion of these topics provided in this chapter can allow researchers and healthcare professionals to critically read CPGs as well as to develop and report CPGs with high-quality and transparency.

References

[1] Qaseem A, Forland F, Macbeth F, Ollenschläger G, Phillips S, van der Wees P. Board of Trustees of the Guidelines International Network. Guidelines International Network: toward international standards for clinical practice guidelines. Ann Intern Med. 2012;156(7):525-31.

[2] Institute of Medicine. 2011. Clinical practice guidelines we can trust. Washington, DC: The National Academies Press.

[3] Shiffman RN, Shekelle P, Overhage JM, Slutsky J, Grimshaw J, Deshpande AM. Standardized reporting of clinical practice guidelines: a proposal from the Conference on Guideline Standardization. Ann Intern Med. 2003;139(6):493-8.

[4] Schünemann HJ, Wiercioch W, Etxeandia I, Falavigna M, Santesso N, Mustafa R et al. Guidelines 2.0: systematic development of a comprehensive checklist for a successful guideline enterprise. CMAJ. 2014;186(3):E123-42.

[5] World Health Organization. WHO Handbook for Guideline Development. 2nd ed. Geneva, Switzerland: WHO Press, 2014.

[6] Woolf SH, Grol R, Hutchinson A, Eccles M, Grimshaw J. Clinical guidelines: potential benefits, limitations, and harms of clinical guidelines. BMJ. 1999;318(7182):527-30.

[7] Elwyn G, Laitner S, Coulter A, Walker E, Watson P, Thomson R. Implementing shared decision making in the NHS. BMJ. 2010;341:c5146.

[8] Senate and House of Representatives. Patient Protection and Affordable Care Act. HR 3590. Washington: 2010.

[9] Tinetti ME, Bogardus ST Jr, Agostini JV. Potential pitfalls of disease-specific guidelines for patients with multiple conditions. N Engl J Med. 2004;351(27):2870-4.

[10] Mutasingwa DR, Ge H, Upshur RE. How applicable are clinical practice guidelines to elderly patients with comorbidities? Can Fam Physician. 2011;57(7):e253-62.

[11] Steel N, Abdelhamid A, Stokes T, Edwards H, Fleetcroft R, Howe A, et al. A review of clinical practice guidelines found that they were often based on evidence of uncertain relevance to primary care patients. J Clin Epidemiol. 2014;67(11):1251-7.

[12] Incze M, Ross JS. On the Need for (Only) High-Quality Clinical Practice Guidelines. JAMA Intern Med. 2019;179(4):561.

[13] Siering U, Eikermann M, Hausner E, Hoffmann-Eßer W, Neugebauer EA. Appraisal tools for clinical practice guidelines: a systematic review. PLoS One. 2013;8:e82915.

[14] Hoffmann-Eßer W, Siering U, Neugebauer EAM, Lampert U, Eikermann M. Systematic review of current guideline appraisals performed with the Appraisal of Guidelines for Research & Evaluation II instrument-a third of AGREE II users apply a cut-off for guideline quality. J Clin Epidemiol. 2018;95:120-127.

[15] Brouwers MC, Spithoff K, Lavis J, Kho ME, Makarski J, Florez ID. What to do with all the AGREEs? The AGREE portfolio of tools to support the guideline enterprise. J Clin Epidemiol. 2020;125:191-197.

[16] Molino CGRC, Leite-Santos NC, Gabriel FC, Wainberg SK, Vasconcelos LP, Mantovani-Silva RA, et al. Chronic Diseases and Informed Decisions (CHRONIDE) Group. Factors Associated With High-Quality Guidelines for the Pharmacologic Management of Chronic Diseases in Primary Care: A Systematic Review. JAMA Intern Med. 2019;179(4):553-560.

[17] Hoffmann-Eßer W, Siering U, Neugebauer EA, Brockhaus AC, Lampert U, Eikermann M. Guideline appraisal with AGREE II: Systematic review of the current evidence on how users handle the 2 overall assessments. PLoS One. 2017;12(3):e0174831.
[18] Armstrong JJ, Goldfarb AM, Instrum RS, MacDermid JC. Improvement evident but still necessary in clinical practice guideline quality: a systematic review. J Clin Epidemiol. 2017;81:13-21.
[19] Alonso-Coello P, Irfan A, Solà I, Gich I, Delgado-Noguera M, Rigau D, et al. The quality of clinical practice guidelines over the last two decades: a systematic review of guideline appraisal studies. Qual Saf Health Care. 2010;19(6):e58.
[20] National Institute for Health and Care Excellence (NICE). Developing NICE guidelines: the manual. United Kingdom: NICE, 2014.
[21] El-Harakeh A, Morsi RZ, Fadlallah R, Bou-Karroum L, Lotfi T, Akl EA. Prioritization approaches in the development of health practice guidelines: a systematic review. BMC Health Serv Res. 2019;19(1):692.
[22] Guyatt GH, Oxman AD, Kunz R, Atkins D, Brozek J, Vist G, et al. GRADE Guidelines: 2. Framing the question and deciding on important outcomes. J Clin Epidemiol. 2011;64:395–400.
[23] Armstrong MJ, Mullins CD, Gronseth GS, Gagliardi AR. Impact of patient involvement on clinical practice guideline development: a parallel group study. Implement Sci. 201816;13(1):55.
[24] Fadlallah R, El-Harakeh A, Bou-Karroum L, Lotfi T, El-Jardali F, Hishi L, et al. A common framework of steps and criteria for prioritizing topics for evidence syntheses: a systematic review. J Clin Epidemiol. 2020;120:67-85.
[25] Scottish Intercollegiate Guidelines Network (SIGN). A guideline developer's handbook. Edinburgh: SIGN; 2019.
[26] National Health and Medical Research Council (NHMRC). Guidelines for Guidelines. Canberra: NHMRC, 2016.
[27] Brasil. Ministério da Saúde. Secretaria de Ciência, Tecnologia e Insumos Estratégicos. Departamento de Gestão e Incorporação de Tecnologias em Saúde. Diretrizes metodológicas: elaboração de diretrizes clínicas. Brasília: Ministério da Saúde, 2016.
[28] Macbeth F, Webster A, Foxlee R, Smith G, Loudon K, Soares-Weiser K. Cochrane Conflict of Interest Policy for Cochrane Library Content. 2020.
[29] Wang Q, Huang C, Li J, Jiang T, Lv P, Zhao M, et al. A survey of reporting status of conflict of interest and economic evidence in Chinese clinical practice guidelines published in 2017. CJEBM. 2018;4:379-87.
[30] Saleh RR, Majeed H, Tibau A, Booth CM, Amir E. Undisclosed financial conflicts of interest among authors of American Society of Clinical Oncology clinical practice guidelines. Cancer. 2019;125(22):4069-4075.
[31] Traversy G, Barnieh L, Akl EA, Allan GM, Brouwers M, Ganache I, et al.. Managing conflicts of interest in the development of health guidelines. CMAJ. 2021;193(2):E49-E54.
[32] Schünemann HJ, Al-Ansary LA, Forland F, Kersten S, Komulainen J, Kopp IB, et al. Board of Trustees of the Guidelines International Network. Guidelines International Network: Principles for Disclosure of Interests and Management of Conflicts in Guidelines. Ann Intern Med. 2015;163(7):548-53.

[33] Baker A, Young K, Potter J, Madan I. A review of grading systems for evidence-based guidelines produced by medical specialties. Clin Med (Lond). 2010;10(4):358-63.
[34] Balshem H, Helfand M, Schünemann HJ, Oxman AD, Kunz R, Brozek J et al. GRADE guidelines: 3. Rating the quality of evidence. J Clin Epidemiol. 2011;64(4):401-6.
[35] Guyatt G, Oxman AD, Akl EA, Kunz R, Vist G, Brozek J, et al. GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. J Clin Epidemiol. 2011;64(4):383-94.
[36] Alonso-Coello P, Schünemann H, Moberg J, Brignardello-etersen R, Akl E, Davoli M et al. GRADE Evidence to Decision (EtD) frameworks: a systematic and transparent approach to making well informed healthcare choices. 1: Introduction BMJ 2016; 353:i2016
[37] Dukhanin V, Searle A, Zwerling A, Dowdy DW, Taylor HA, Merritt MW. Integrating social justice concerns into economic evaluation for healthcare and public health: a systematic review. Soc Sci Med. 2018;198:27–35.
[38] Rehfuess EA, Stratil JM, Scheel IB, Portela A, Norris SL, Baltussen R. The WHO-INTEGRATE evidence to decision framework version 1.0: integrating WHO norms and values and a complexity perspective. BMJ Glob Health. 2019;4:e000844
[39] Mancia G, Zanchetti A. Evidence-based medicine: an educational instrument or a standard for implementation? J Hypertens 1999; 17:15091510.
[40] Rosenfeld RM, Shiffman RN, Robertson P. Department of Otolaryngology State University of New York Downstate. Clinical Practice Guideline Development Manual, Third Edition: a quality-driven approach for translating evidence into action. Otolaryngol Head Neck Surg. 2013;148(1 Suppl):S1-55.

[41] Steichen O. The burden of conflicting guidelines. J Hypertens. 2020;38(10):1945-1947.
[42] Leite-Santos NC, de Melo DO, Mantovani-Silva RA, Gabriel FC, Fornasari GS, Dórea EL, et al. Guidelines for hypertension management in primary care: is local adaptation possible? J Hypertens 2020; 38:20592073.
[43] Alonso-Coello P, Oxman AD, Moberg J, Brignardello-Petersen R, Akl EA, Davoli M, et al. GRADE Working Group. GRADE Evidence to Decision (EtD) frameworks: a systematic and transparent approach to making well informed healthcare choices. 2: Clinical practice guidelines. BMJ. 2016;353:i2089.
[44] Andrews J, Guyatt G, Oxman AD, Alderson P, Dahm P, Falck-Ytter Y, et al. GRADE guidelines: 14. Going from evidence to recommendations: the significance and presentation of recommendations. J Clin Epidemiol. 2013;66(7):719-25.
[45] Shaker MS, Oppenheimer J, Wallace DV, Golden DBK, Lang DM; Joint Task Force for Allergy Practice Parameters; Contributors, Lang ES, Bernstein JA, Campbell RL, Chu D, Dinakar C, Ellis AK, Greenhawt M, Horner C, Lieberman JA, Rank MA, Stukus DR, Wang J. Making the GRADE in anaphylaxis management: Toward recommendations integrating values, preferences, context, and shared decision making. Ann Allergy Asthma Immunol. 2020;124(6):526-535.e2.
[46] Piggott T, Baldeh T, Akl EA, Junek M, Wiercioch W, Schneider R, et al. Supporting effective participation in health guideline panels: The Guideline Participant Tool. J Clin Epidemiol. 2020:S0895-4356(20)31100-8.
[47] Boon MH, Thomson H, Shaw B, Akl EA, Lhachimi SK, López-Alcalde J, et al. GRADE Working Group. Challenges in applying the GRADE approach in public health guidelines

and systematic reviews: A concept paper from the GRADE Public Health Group. J Clin Epidemiol. 2021:S0895-4356(21)00003-2.

[48] Chen Y, Yang K, Marušic A, Qaseem A, Meerpohl JJ, Flottorp S, etl al. RIGHT (Reporting Items for Practice Guidelines in Healthcare) Working Group. A Reporting Tool for Practice Guidelines in Health Care: The RIGHT Statement. Ann Intern Med. 2017;166(2):128-132.

[49] Yang N, Yu Y, Zhang A, Estill J, Wang X, Zheng M, et al. Reporting, presentation and wording of recommendations in clinical practice guideline for gout: a systematic analysis. BMJ Open. 2019;9(1):e024315.

[50] Shekelle PG, Kravitz RL, Beart J, Marger M, Wang M, Lee M. Are nonspecific practice guidelines potentially harmful? A randomized comparison of the effect of nonspecific versus specific guidelines on physician decision making. Health Serv Res. 2000;34(7):1429-48.

[51] Michie S, Johnston M. Changing clinical behaviour by making guidelines specific. BMJ 2004;328:343–5.

[52] Wang X, Chen Y, Akl EA, Tokalić R, Marušić A, Qaseem A, et al. RIGHT working group. The reporting checklist for public versions of guidelines: RIGHT-PVG. Implement Sci. 2021;16(1):10.

[53] Shekelle, PG, Schünemann H, Woolf SH, Eccles M, Grimshaw J. 2010. State of the art of CPG development and best practice standards. In: Committee on Standards for Trustworthy Clinical Practice Guidelines commissioned paper.

[54] Shekelle P, Woolf S, Grimshaw JM, Schünemann HJ, Eccles MP. Developing clinical practice guidelines: reviewing, reporting, and publishing guidelines; updating guidelines; and the emerging issues of enhancing guideline implementability and accounting for comorbid conditions in guideline development. Implement Sci. 2012;7:62.

[55] Cabana MD, Rand CS, Powe NR, Wu AW, Wilson MH, Abboud PA et al. Why don't physicians follow clinical practice guidelines? A framework for improvement. JAMA. 1999;282(15):1458-65.

[56] Francke AL, Smit MC, de Veer AJ, Mistiaen P. Factors influencing the implementation of clinical guidelines for health care professionals: a systematic meta-review. BMC Med Inform Decis Mak. 2008;8:38.

[57] Grimshaw J, Eccles M, Thomas R, MacLennan G, Ramsay C, Fraser C, et al. Toward evidence-based quality improvement. Evidence (and its limitations) of the effectiveness of guideline dissemination and implementation strategies 1966-1998. J Gen Intern Med. 2006;21 Suppl 2(Suppl 2):S14-20.

[58] Grimshaw JM, Eccles MP, Lavis JN, Hill SJ, Squires JE. Knowledge translation of research findings. Implement Sci. 2012;7:50.

[59] Bullock HL, Lavis JN, Wilson MG, Mulvale G, Miatello A. Understanding the implementation of evidence-informed policies and practices from a policy perspective: a critical interpretive synthesis. Implement Sci. 2021:16.

[60] Gagliardi AR, Marshall C, Huckson S, James R, Moore V. Developing a checklist for guideline implementation planning: review and synthesis of guideline development and implementation advice. Implement Sci. 2015;10:19.

[61] Fischer F, Lange K, Klose K, Greiner W, Kraemer A. Barriers and Strategies in Guideline Implementation-A Scoping Review. Healthcare (Basel). 2016;4(3):36.

[62] de Vasconcelos LP, Melo DO, Stein AT, de Carvalho HB. Even High-Quality CPGs Seldom Include Implementation Strategies. Front Pharmacol. 2021;11:593894.
[63] Gagliardi AR, Alhabib S. Members of Guidelines International Network Implementation Working Group. Trends in guideline implementation: a scoping systematic review. Implement Sci. 2015;10:54.
[64] Bennett B, Coventry E, Greenway N, Minchin M. The NICE process for developing quality standards and indicators. Z Evid Fortbild Qual Gesundhwes. 2014;108(8-9):481-6.
[65] Nothacker M, Stokes T, Shaw B, Lindsay P, Sipilä R, Follmann M, et al. Guidelines International Network (G-I-N) Performance Measures Working Group. Reporting standards for guideline-based performance measures. Implement Sci. 2016;11:6.
[66] Kötter T, Blozik E, Scherer M. Methods for the guideline-based development of quality indicators--a systematic review. Implement Sci. 2012;7:21.
[67] Wolfenden L, Foy R, Presseau J, Grimshaw JM, Ivers NM, Powell BJ, et al. Designing and undertaking randomised implementation trials: guide for researchers. BMJ. 2021;372:m3721.
[68] Shekelle PG, Woolf SH, Eccles M, Grimshaw J. Clinical guidelines: developing guidelines. BMJ. 1999;318(7183):593-6.
[69] Fervers B, Burgers JS, Voellinger R, Brouwers M, Browman GP, Graham ID, et al. ADAPTE Collaboration. Guideline adaptation: an approach to enhance efficiency in guideline development and improve utilisation. BMJ Qual Saf. 2011;20(3):228-36.
[70] Ransohoff DF, Pignone M, Sox HC. How to decide whether a clinical practice guideline is trustworthy. JAMA. 2013;309:139–40.

[71] Browman GP, Somerfield MR, Lyman GH, Brouwers MC. When is good, good enough? Methodological pragmatism for sustainable guideline development. Implement Sci. 2015;10:28.
[72] Dizon JM, Machingaidze S, Grimmer K. To adopt, to adapt, or to contextualise? The big question in clinical practice guideline development. BMC Res Notes. 2016;9(1):442.
[73] Burgers JS, Anzueto A, Black PN, Cruz AA, Fervers B, Graham ID, etl al. ATS/ERS Ad Hoc Committee on Integrating and Coordinating Efforts in COPD Guideline Development. Adaptation, evaluation, and updating of guidelines: article 14 in Integrating and coordinating efforts in COPD guideline development. An official ATS/ERS workshop report. Proc Am Thorac Soc. 2012;9(5):304-10.
[74] Darzi A, Abou-Jaoude EA, Agarwal A, Lakis C, Wiercioch W, Santesso N, et al. A methodological survey identified eight proposed frameworks for the adaptation of health related guidelines. J Clin Epidemiol. 2017;86:3-10.
[75] ADAPTE Collaboration. Manual for Guideline Adaptation. 2007 Available from: http://www.g-i-n.net/document-store/working-groups-documents/adaptation/adapte-manual-for-guideline.pdf
[76] Wang Z, Norris SL, Bero L. The advantages and limitations of guideline adaptation frameworks. Implement Sci. 2018;13(1):72.
[77] Schünemann HJ, Wiercioch W, Brozek J, Etxeandia-Ikobaltzeta I, Mustafa RA, Manja V, et al. GRADE Evidence to Decision (EtD) frameworks for adoption, adaptation, and de novo development of trustworthy recommendations: GRADE-ADOLOPMENT. J Clin Epidemiol. 2017;81:101-110.

[78] Shekelle PG. Updating practice guidelines. JAMA. 2014;311(20):2072–2073.
[79] Martínez García L, Sanabria AJ, García Alvarez E, Trujillo-Martín MM, Etxeandia-Ikobaltzeta I, Kotzeva A, et al. Updating Guidelines Working Group. The validity of recommendations from clinical guidelines: a survival analysis. CMAJ. 2014;186(16):1211-9.
[80] Vernooij RW, Alonso-Coello P, Brouwers M, Martínez García L. CheckUp Panel. Reporting Items for Updated Clinical Guidelines: Checklist for the Reporting of Updated Guidelines (CheckUp). PLoS Med. 2017;14(1):e1002207.
[81] Agbassi C, Messersmith H, McNair S, Brouwers M. Priority-based initiative for updating existing evidence-based clinical practice guidelines: the results of two iterations. J Clin Epidemiol. 2014;67(12):1335–1342.

Since available resources are limited, delivering health services involves making decisions. Decisions are required on what health technologies (e.g. drugs, devices, vaccines, procedures) should be offered, the way the health system is organized, and how the interventions should be provided in order to achieve an optimal health gain. At the same time, a balance between stakeholders' expectations (e.g. innovators, manufacturers, clinicians, society) should be maintained. Providing reliable inputs for the decision-makers is highly dependent of interaction, division of labor and cooperation between healthcare professionals, researchers, and the political environment. Decisions must be made on an evidence-based foundation where all relevant circumstances and consequences are systematically illustrated by means of scientific methods. This book provides a basic framework for understanding the major concepts and methods available in the evidence-based practice field and how they can be used for Health Technology Assessment (HTA).