




**ADVERTIMENT.** L'accés als continguts d'aquesta tesi queda condicionat a l'acceptació de les condicions d'ús establertes per la següent llicència Creative Commons:  [http://cat.creativecommons.org/?page\\_id=184](http://cat.creativecommons.org/?page_id=184)

**ADVERTENCIA.** El acceso a los contenidos de esta tesis queda condicionado a la aceptación de las condiciones de uso establecidas por la siguiente licencia Creative Commons:  <http://es.creativecommons.org/blog/licencias/>

**WARNING.** The access to the contents of this doctoral thesis it is limited to the acceptance of the use conditions set by the following Creative Commons license:  <https://creativecommons.org/licenses/?lang=en>



# **Universitat Autònoma de Barcelona**

## **Facultat de Medicina**

Doctorado en Metodología de la Investigación Biomédica y Salud Pública

---

### **TESIS DOCTORAL**

### **INCORPORATING DECISION ANALYSIS MODELS IN THE DEVELOPMENT OF HEALTH RECOMMENDATIONS.**

---

#### **Doctorando**

Carlos Gilberto Canelo Aybar

#### **Director**

Pablo Alonso Coello

#### **Tutor**

Xavier Bonfill Cosp

Barcelona, mayo 2022

## ACKNOWLEDGMENTS

I would like to thank Pablo Alonso-Coello, as well as Xavier Bonfill, for their guidance and constant support throughout this thesis

To all my colleagues from Cochrane Iberoamerican for their friendship, anecdotes and great collaborative work.

To all the researchers that collaborated during the development of each project

Finally, to my family and friends for their always encouraging words

“Doubt is one of the names of intelligence”

Jorge Luis Borges



## INDEX

<b>1. SUMMARY</b>	6
1.1 Abstract	6
1.2 Resúmen	8
1.3 Resumeixen	10
<b>2. INTRODUCTION</b>	13
2.1 Clinical Practice Guidelines	13
2.2 Decision analysis models	16
2.3 Decision models and synthesis of evidence	21
2.4 The GRADE approach and certainty of modelling evidence	25
<b>3. JUSTIFICATION</b>	32
<b>4. OBJECTIVES</b>	35
<b>5. METHODS</b>	37
5.1 First study: Benefits and harms of annual, biennial, or triennial breast cancer mammography screening for women at average risk of breast cancer: a systematic review for the European Commission Initiative on Breast Cancer (ECIBC)	37
5.2 Second study: Recommendations from the European Commission Initiative on Breast Cancer for multigene testing to guide the use of adjuvant chemotherapy in patients with early breast cancer, hormone receptor positive, HER-2 negative	39
5.3 Third study: GRADE Guidelines 30: the GRADE approach to assessing the certainty of modeled evidence—An overview in the context of health decision-making	41
<b>6. RESULTS</b>	43
6.1 First study: Benefits and harms of annual, biennial, or triennial breast cancer mammography screening for women at average risk of breast cancer: a systematic review for the European Commission Initiative on Breast Cancer (ECIBC)	45
6.2 Second study: Recommendations from the European Commission Initiative on Breast Cancer for multigene testing to guide the use of adjuvant chemotherapy in patients with early breast cancer, hormone receptor positive, HER-2 negative	65
6.3 Third study: GRADE Guidelines 30: the GRADE approach to assessing the certainty of modeled evidence—An overview in the context of health decision-making	77
<b>7. DISCUSSION</b>	108
7.1 Main findings	108
7.2 Our results in the context of previous research	114
7.3 Limitations and strengths	117
7.4 Implications for practice and research	118
<b>8. CONCLUSIONS</b>	122
<b>9. REFERENCES</b>	125
<b>10. APENDIX</b>	133

---

## **SUMMARY**

---

## 1. SUMMARY

---

### 1.1 Abstract

#### Background

Decision analysis models are mathematical frameworks representing variables and their interrelationships, to describe observed phenomena or predicting events. Models may improve decision making, by projecting interventions to life-time horizon or predicting the effect of alternative ways of delivering an intervention. Although there are a few relevant examples of incorporating models in clinical practice guidelines (CPGs), the methods are still underdeveloped in areas, such as the assessment of the certainty of evidence or how to integrate this type of evidence within the CPG development process.

#### Objectives

To develop methods for incorporating models into CPG development. The design of each study addresses different relevant aspects such as: 1) the integration of empirical and modelling evidence to inform the effectiveness of a public health intervention, 2) the use of a model in a guideline panel to assist the recommendation formulation process, 3) the development of guidance to assess the certainty of evidence of models within the context of systematic reviews and CPG.

#### Methods

The thesis contains three studies with different methodological designs:

1) A systematic review of effects: We searched PubMed, EMBASE, and the Cochrane Library to identify RCTs, observational or modelling studies, comparing desirable (i.e. deaths averted) and undesirable (i.e. overdiagnosis) effects from annual, biennial, or triennial breast cancer (BC) mammography screening. We assessed the certainty of the evidence adapting the GRADE approach.

2) A systematic review of effects and use of a model by a guideline panel: we search for RCTs or cohort studies to assess the value of multigene tests to assist adjuvant chemotherapy decisions in early BC. Then, we develop a decision tree model to estimate the downstream consequences of testing patients with multigene tests. A multidisciplinary guideline panel developed recommendations informed by the model estimations.

iii) Development of a GRADE guidance: a workshop with experts from different fields who participated in specific tasks and in a large group discussion session to inform the development of an approach to assess the certainty of evidence.

## **Results**

The first study identified one RCT, 11 modelling and 13 observational studies. The balance of effects probably favours biennial screening in women 50–69. In younger women, annual screening may have a less favourable balance, in women aged 70–74 years longer screening intervals may be more favourable. The overall certainty of the evidence was very low. We included models to complement the gaps in the empirical evidence and presented this work on the workshop during the third study of this thesis.

The second study included five studies for two types of multigene tests (four RCTs and one pooled analysis of observational studies). We showed that modelling on different treatment scenarios the number of chemotherapies avoided by using the 21-RS test would be from more than 600 to about 200, while using the 70-GS test would result in an avoidance of chemotherapy in about 230 women out of 1,000. The guideline panel issued recommendations using this evidence.

In the third study, participants agreed that the GRADE approach to assess the certainty of evidence also applies when assessing the certainty of evidence from models. Guidance to use the GRADE approach to modelling evidence was developed, along with a framework to consider this type of evidence over the CPG development process.

## **Conclusion**

This thesis provides new knowledge on how to incorporate evidence from models in health decision-making, including real examples, a framework, and guidance on how to assess the certainty of evidence of this type of evidence. Future areas of research include the developing of more detailed methods for assessing specific GRADE domains, and improve the presentation formats to adequacy display modelling evidence research.

## **1.2 Resumen**

### **Introducción**

Los modelos de análisis de decisión son marcos matemáticos que representan variables y sus interrelaciones para describir fenómenos observados o predecir eventos. Los modelos pueden mejorar la toma de decisiones proyectando una intervención a un horizonte de tiempo de vida o prediciendo el efecto de formas alternativas de brindar una intervención. Aunque hay algunos ejemplos en la literatura sobre la incorporación de modelos en guías de práctica clínica (GPC), los métodos están aún poco desarrollados en áreas como la evaluación de la certeza de evidencia o cómo integrarlos en el desarrollo de GPC.

### **Objetivos**

Desarrollar métodos para incorporar modelos en el desarrollo de GPC. El diseño de cada estudio aborda diferentes aspectos: 1) integrar evidencia empírica y de modelos para informar la efectividad de una intervención de salud pública, 2) el uso de un modelo en un panel de GPC durante el proceso de formulación de recomendaciones, 3) desarrollar una guía para evaluar la certeza de evidencia de modelos en el contexto de revisiones sistemáticas o GPC.

### **Métodos**

La tesis contiene tres estudios con diferentes diseños:

- 1) Una revisión sistemática de efectividad: se buscó en PubMed, EMBASE y la Biblioteca Cochrane para identificar ECA, estudios observacionales o modelos, que compararan los efectos deseables (ej. muertes evitadas) e indeseables (ej. sobrediagnóstico) del cribado anual, bienal o trienal de cáncer de mama (CM) con mamografía. Evaluamos la certeza de la evidencia adaptando el sistema GRADE
- 2) Una revisión sistemática de efectividad y uso de un modelo en un panel de guías: buscamos ECA o estudios de cohortes para evaluar pruebas multigénicas para informar las decisiones de quimioterapia adyuvante en el CM temprano. Desarrollamos un modelo de árbol de decisión para estimar las consecuencias de evaluar a pacientes con pruebas multigénicas. Un panel multidisciplinario formulo recomendaciones basadas en estimaciones del modelo.
- iii) Desarrollo de una guía GRADE: un taller con expertos de diferentes campos que participaron en tareas específicas y en una sesión general de discusión informo el desarrollo de una guía para evaluar la certeza de la evidencia.

## **Resultados**

El primer estudio identificó un ECA, 11 modelos y 13 estudios observacionales. El balance de efectos favorecería el cribado bienal en mujeres de 50-69 años. En mujeres más jóvenes, el cribado anual tendría un balance menos favorable, en mujeres de 70-74 años intervalos de cribado más largos sería más favorables. La certeza de la evidencia fue muy baja. Incluimos modelos para complementar los vacíos en la evidencia empírica y presentamos este trabajo en el taller del tercer estudio.

El segundo estudio incluyó cinco estudios para dos tipos de pruebas multigénicas (cuatro ECA y un análisis de estudios observacionales). En el modelo con diferentes escenarios de tratamiento, la cantidad de quimioterapias evitadas con la prueba 21-RS sería de más de 600 a aproximadamente 200, mientras con la prueba 70-GS se evitaría la quimioterapia en aproximadamente 230 mujeres por 1.000. El panel de la guía emitió recomendaciones usando esta evidencia.

En el tercer estudio, los participantes consideraron que el enfoque GRADE para evaluar la certeza de la evidencia es aplicable a la evidencia de modelos. Se desarrolló una guía para usar el sistema GRADE en modelos, y un marco para considerar este tipo de evidencia durante el desarrollo de GPC.

## **Conclusión**

Esta tesis proporciona nuevos conocimientos sobre cómo incorporar evidencia de modelos en la toma de decisiones en salud, incluidos ejemplos reales, un marco y una guía sobre cómo evaluar la certeza de este tipo de evidencia. Futuras áreas de investigación incluyen el desarrollo de métodos detallados para evaluar dominios específicos de GRADE y mejorar los formatos para presentar la evidencia proveniente de modelos.

## **1.3 Resumeixen**

### **Introducció**

Els models d'anàlisi de decisió són marcs matemàtics que representen variables i les seves interrelacions per a descriure fenòmens observats o predir esdeveniments. Els models poden millorar la presa de decisions projectant una intervenció a un horitzó de temps de vida o predient l'efecte de formes alternatives de brindar una intervenció. Encara que hi ha alguns exemples en la literatura sobre la incorporació de models en guies de pràctica clíniques (GPC), els mètodes estan encara poc desenvolupats en àrees com l'avaluació de la certesa d'evidència o com integrar-los en el desenvolupament de GPC.

### **Objectius**

Desenvolupar mètodes per a incorporar models en el desenvolupament de GPC. El disseny de cada estudi aborda diferents aspectes: 1) integrar evidència empírica i de models per a informar l'efectivitat d'una intervenció de salut pública, 2) l'ús d'un model en un panell de GPC durant el procés de formulació de recomanacions, 3) desenvolupar una guia per a avaluar la certesa d'evidència de models en el context de revisions sistemàtiques o GPC.

### **Mètodes**

La tesi conté tres estudis amb diferents dissenys:

- 1) Una revisió sistemàtica d'efectivitat: es va buscar en PubMed, EMBASE i la Biblioteca Cochrane per a identificar ECA, estudis observacionals o models, que comparessin els efectes desitjables (ex. morts evitades) i indesitjables (ex. sobrediagnòstic) del garbellat anual, biennal o triennal de càncer de mama (CM) amb mamografia. Avaluem la certesa de l'evidència adaptant el sistema GRADE
- 2) Una revisió sistemàtica d'efectivitat i ús d'un model en un panell de guies: busquem ECA o estudis de cohorts per a avaluar proves multigèniques per a informar les decisions de quimioteràpia adjuvant en el CM primerenc. Desenvolupem un model d'arbre de decisió per a estimar les conseqüències d'avaluar a pacients amb proves multigèniques. Un panell multidisciplinari formulo recomanacions basades en estimacions del model.
- iii) Desenvolupament d'una guia GRADE: un taller amb experts de diferents camps que van participar en tasques específiques i en una sessió general de discussió informo el desenvolupament d'una guia per a avaluar la certesa de l'evidència.

### **Resultats**

El primer estudi va identificar un ECA, 11 models i 13 estudis observacionals. El balanç d'efectes afavoriria el garbellat biennal en dones de 50-69 anys. En dones més joves, el garbellat anual tindria un balanç menys favorable, en dones de 70-74 anys intervals de garbellat més llargs seria més favorables. La certesa de l'evidència va ser molt baixa. Incloem models per a complementar els buits en l'evidència empírica i presentem aquest treball en el taller del tercer estudi.

El segon estudi va incloure cinc estudis per a dos tipus de proves multigèniques (quatre ECA i una anàlisi d'estudis observacionals). En el model amb diferents escenaris de tractament, la quantitat de quimioteràpies evitades amb la prova 21-RS seriosa de més de 600 a aproximadament 200, mentre amb la prova 70-GS s'evitaria la quimioteràpia en aproximadament 230 dones per 1.000. El panell de la guia va emetre recomanacions usant aquesta evidència.

En el tercer estudi, els participants van considerar que l'enfocament GRADE per a avaluar la certesa de l'evidència és aplicable a l'evidència de models. Es va desenvolupar una guia per a usar el sistema GRADE en models, i un marc per a considerar aquest tipus d'evidència durant el desenvolupament de GPC.

### **Conclusió**

Aquesta tesi proporciona nous coneixements sobre com incorporar evidència de models en la presa de decisions en salut, inclosos exemples reals, un marc i una guia sobre com avaluar la certesa d'aquesta mena d'evidència. Futures àrees de recerca inclouen el desenvolupament de mètodes detallats per a avaluar dominis específics de GRADE i millorar els formats per a presentar l'evidència provinent de models.



---

## INTRODUCTION

---

## 2. INTRODUCTION

---

Decision analytical models (aka. mathematical models) have been used in public health to assist decision making for a long time ago. First description dates back to 1760, when Daniel Bernoulli developed a model simulate smallpox transmission and the potential impact of control measures.<sup>1</sup> Subsequently, in 1906, William Hamer developed a measles transmission model<sup>1</sup> and two years later Ronald Ross presented a model of malaria transmission.<sup>2</sup>

In recent years, the number of mathematical modelling publications has increased steeply, along with the complexity of clinical and public health interventions, and the needs for timely decisions by policy makers. Noteworthy, mathematical modelling studies are not restricted to infectious diseases field, they address a wide range of questions as exemplified by recent clinical guidelines in the field of cancer screening, issued by important international institutions.

Below, I will describe the “state of the art” of clinical practice guidelines (CPG) and the use of decision analytical models (hereafter “models”) in the context of decision making and formulation of recommendations during CPG development; the role of modelling in economic evaluation is well recognized in guideline development, and will therefore not be covered throughout this work.

### 2.1 Clinical Practice Guidelines

The constant grow of health literature makes unfeasible for clinicians or healthcare policy makers to keep themselves updated. For example, the number of randomized controlled trials (RCT) published in MEDLINE grew from 5,000 during the period from 1978-1985<sup>3</sup> to around 45,000 registered RCTs only in the year 2021.<sup>4</sup> In addition, the identified literature may have methodological limitations, or be not applicable to the target populations or setting of interest. Thus, clinicians may become increasingly overwhelmed by a vast volume of evidence of uncertain value, without the required skills to appraise credibility.

Clinical Practice Guidelines (CPG) are statements intended to provide a systematic aid to clinicians, through the complex process of medical decisions.<sup>3</sup> When rigorously developed, using a transparent process that combines scientific evidence, clinician experiential knowledge, and patient values, CPGs have the potential to improve healthcare decision making, and enhance healthcare quality and outcomes.<sup>5</sup>

In 2011, the Institute of Medicine (IoM), now the National Academy of Medicine, defined clinical practice guidelines as, “*statements that include recommendations intended to*

*optimize patient care that are informed by a systematic review of evidence and an assessment of the benefits and harms of alternative care options”*.<sup>3</sup> Following this definition, a trustworthy guideline should fulfil the following requirements:

- Be based on a systematic review of the existing evidence;
- Be developed by a knowledgeable, multidisciplinary panel of experts and representatives from key affected groups;
- Consider important patient subgroups and patient preferences, as appropriate;
- Be based on an explicit and transparent process that minimizes distortions, biases, and conflicts of interest;
- Provide a clear explanation of the logical relationships between alternative care options and health outcomes, and provide ratings of both the quality of evidence and the strength of the recommendations; and
- Be reconsidered and revised as appropriate when important new evidence warrants modifications of recommendations.<sup>3</sup>

Since the publication of the IoM report, the number of associations dedicated to CPG initiatives have undergone a remarkable expansion globally. Initially, large guideline development organizations at a national level appeared, such as the National Institute for Health (NICE), the Scottish Intercollegiate Guidelines Network (SIGN) in the United Kingdom. Subsequently, several of those organizations agreed to create the Guidelines International Network (GIN), a worldwide scientific association, whose member (individuals or institutions) are involved in development or implementation of evidence-based guidelines. This network nowadays, comprises 115 organizations, representing about 47 countries from all continents.<sup>6</sup>

Alongside the expansion of CPG dedicated organization, the methods for developing recommendations have also made relevant progress over time. In 2011, the Grading of Recommendation, Assessment, Development, and Evaluation (GRADE) working group described the GRADE approach, which proposed an structured system for rating quality (certainty) of evidence in systematic reviews and guidelines, and for grading the strength of recommendations in CPGs.<sup>7</sup> Later in 2016, the GRADE Working Group, in the context of a European funded project called DECIDE (Developing and Evaluating Communication Strategies to Support Informed Decisions and Practice Based on Evidence) developed the Evidence to Decision (EtD) frameworks, to support the process of moving from evidence to recommendations. This process includes, considering aspects such as the magnitude of

desirable and undesirable effects, the balance of effects, values and preferences, use of resources, or equity.<sup>8,9</sup> The GRADE working group has also outlined the main stages in the process of developing CPG (Table 1).<sup>10</sup>

Other approaches under development include, methods to use resources efficiently, such as adaptation and updating of CPG, building on existing guidelines or provide more local recommendations. Approaches like ADAPTE provide detailed guidance on how to modify guidelines produced in one setting for use in a different setting.<sup>11</sup> The GRADE working group developed the “ADOLOPMENT” approach, combining advantages of adoption, adaptation, and de novo guideline development, building on the EtD framework to allow formulation of recommendations for a specific setting.<sup>12</sup> The updating of CPGs in the context of emergence of new evidence has also made significant progress, noteworthy efforts are the development of tools like UpPriority, to prioritise clinical questions for updating<sup>13</sup>, and CheckUp, to evaluate the completeness of reporting in updated guidelines<sup>14</sup>

**Table 1.** Steps to develop a clinical practice guideline

- |  |
|--|
| <ul style="list-style-type: none"> <li>• Establish the guideline panel</li> <li>• Define the scope of the guidelines</li> <li>• Prioritize the problems</li> <li>• Formulate the clinical questions</li> <li>• Value the relative importance of outcomes</li> <li>• Identify the existing evidence for every clinical question</li> <li>• Grade the quality of existing evidence for each outcome separately</li> <li>• Determine the overall quality of available evidence across outcomes</li> <li>• Decide on the balance between desirable and undesirable consequences</li> <li>• Decide on the strength of recommendation</li> <li>• Formulate the recommendation reflecting its strength</li> </ul> |
|--|

Adapted from: Brozek et al. (2009)<sup>10</sup>

The final aim of implementation of CPGs is to promote high-value interventions most relevant to practitioners, patients, and the society as a whole after consideration of the desirable and undesirable effects. However, in practice, findings from systematic reviews may not directly apply, being sparse or not available to the scenarios of interest due to factors such as the complexity of interventions or the horizon time.<sup>15</sup> For example, cancer screening guideline panels may need to assess not only whether they should or not recommend screening, but also the age at which to start or stop screening, the intervals of testing, and the confirmation methods.<sup>16</sup> Thus, evidence directly addressing these types of scenarios might be unfeasible or even unethical to produce.

Another example of complex scenarios for CPG development is, how to account for patients who have multiple medical conditions. Boyd et al, assessed the applicability of guidelines to a hypothetical 79 year old woman with five chronic conditions: osteoporosis, osteoarthritis, diabetes, hypertension, and chronic obstructive pulmonary disease, and noted that most did not discuss recommendations for management in patients with comorbidities.<sup>17</sup> One strategy to cover complex scenarios is the incorporation of decision analysis models evidence (modelling evidence). This approach has been implemented during the COVID-19 pandemic to develop timely evaluate non-pharmaceutical interventions, which underlines the need to develop methods for the incorporation of this type of evidence.<sup>18, 19</sup>

## 2.2 Decision analysis models

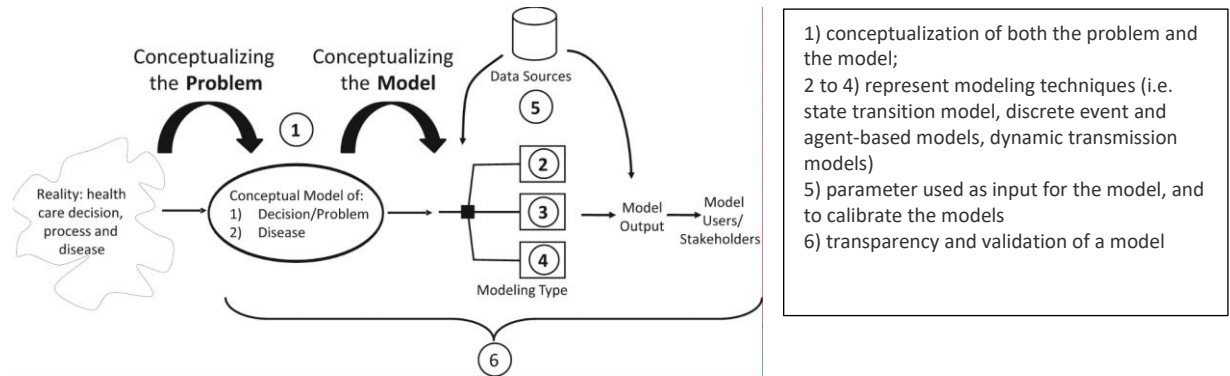
Researchers have used the term model to describe a variety of concepts, but most agree that given the complexity of healthcare, decision makers often rely on some sort of a modelling to answer health-related questions.<sup>20</sup> Overall, models might vary in their structure and degree of complexity. A very simple model may be any calculation to estimate a single variable not directly measured. For example, the population impact of an intervention, estimated as the product of their relative effect (informed by a trial), multiplied by the baseline risk of the population of interest (informed by a cohort study).<sup>21</sup> On the other end of the spectrum, we may find more complex models, such as system dynamics models, used to predict infectious disease transmission, which might require considerable computing.

This thesis will focus on decision models defined as “mathematical framework representing variables and their interrelationships, to describe observed phenomena or predict future events”,<sup>22</sup> excluding statistical models used to estimate the associations between measured variables and their outcomes (e.g., logistic regression models). Practically, all problems can be represented by any model, although some methods are preferable for particular scenarios; for example, a Markov model for chronic diseases over a lifetime horizon, or dynamic models to evaluate vaccine effectiveness.

The modelling process should start by the problem and model conceptualization (Figure 1). Problem conceptualization includes consultation with experts to ensure that the model adequately addresses the decision problem and perspective of analysis, and reflects the strategies of interest, the target population, the time horizon and the relevant resources and health outcomes.<sup>23</sup> The availability of data may constrain the model development, but should not limit the initial discussion of the problem, which must incorporate features of the disease and its outcomes for even though data may be poor or unavailable.<sup>23</sup> In such the

latter scenarios, sensitivity analyses should be conducted on model parameters for which no data exists, to explore their influence on the results.

**Figure 1.** Model and problem conceptualization\*



\*Adapted from: Roberts et al (2012)<sup>23</sup>

During model conceptualization, the components of the problem are represented using a particular analytic method. The conceptual representation will usually influence the decision of which type of modelling approach would best represent the phenomenon or decision problem, under consideration. The choice of a modelling approach is crucial since it can affect the validity of the results.<sup>23</sup> Among the characteristics that should be taken into consideration, to decide which method is most appropriate are: whether it should represent individuals or groups, whether there will be interactions among individuals, the time horizon, whether time will be continuous or discrete, or if the events would occur more than once per individual.<sup>24</sup>

The most common model types among the several techniques available are decision trees, state-transition models, discrete event simulation (DES), agent-based simulation, and dynamic transmission models.<sup>24, 25</sup> Decision trees are useful for problems with short time horizons where estimation of outcomes is straightforward. State-transition models are useful for problems with longer time frames, or when probabilities vary over time.<sup>24</sup> DES are useful for representing what happens to individuals, particularly when there are resource constraints or interactions among individuals.<sup>24</sup> Dynamic transmission models are useful when interactions occurring between groups may have an impact on the results (*Table 2*).<sup>24</sup>

**Table 2. Types of modelling approaches**

Approaches	Description	Areas of application
Decision tree	<p>Decision trees arrange events from left to right through three components in a chronological order:</p> <ul style="list-style-type: none"> <li>. A decision node which indicates competing alternatives</li> <li>. A chance node representing consequences of a given decision (mutually exclusive)</li> <li>. A terminal node, showing value of a branch</li> </ul> <p>Branches connect the nodes and represent the pathways through the tree</p> <p>The expected costs and/or effects associated with each strategy are estimated by weighted averaging of the values of all branches emanating from a decision node</p>	<p>Decision trees are easy to interpret, if the number of branches is kept low</p> <p>A common usage is when the disease is acute and events are not recurrent</p>
Markov cohort model	<p>Markov cohort models simulates the movement of patients through health states over time according to specific transition probabilities</p> <p>The basic Markov cohort model relies on the Markovian assumption that transition probabilities depend only on the current state and not on any previous health states</p> <p>Typically, the entire cohort will enter the model at the same time, although it can be distributed among various states</p> <p>Costs and health outcomes are determined by health states, and overall costs and QALYs are estimated by adding the cycle sums over the time horizon</p>	<p>Markov models provide a more manageable representation if the time horizon is long or if events recur</p>
Markov microsimulation	<p>Markov microsimulation simulates individual patients over time. This approach is capable of storing the history of each individual patient over the course of the model, thus transition rates may be conditional on previous and existing risk factors or events</p> <p>Transitions occur only once per cycle, similar to the Markov cohort model</p> <p>Following the simulation, each patient has their own respective costs and outcomes. Overall costs and QALYs can then be calculated as the average from a large number of simulated patients</p>	<p>Markov microsimulations model are preferred when the representation of all aspects of the decision problem would lead to an unmanageable number of health states</p>
Discrete event simulation	<p>Discrete event simulation describes the progression of entities (individuals), rather than a fixed time, time is continuous with patient progression sampled according to parametric or empirical time-to-event distributions. Individuals may either be simulated one-by-one or simultaneously</p> <p>Consequences, such as costs and effects, can be attributed to anything that is sensible, such as events, time with a particular condition or simply having a particular patient attribute within the model</p>	<p>Discrete event simulation has greater flexibility due to incorporating time as a continuous instead of fixed intervals. it is useful for settings of constrained resources (e.g., number of hospital beds) or process-driven situations (e.g., waitlists)</p>

System dynamics	System dynamics describes a system through feedback loops and flows. The causal loop diagram provides a qualitative visualization of a system. Its basic building block is the feedback loop, describing movement (i.e., flow) from one pool eventually returning in some form to its origin. Movement between stocks is defined by the rate of flow, dictated by differential equation, and time changes continuously Costs and outcomes may be linked to the time spent in a particular stock or by the flow between stocks	Application on infectious diseases where differential equations are taken from mathematical models of infectious disease epidemiology
Compartmental models	The population is divided into various compartments, representing their average state. Individuals within a single compartment are considered homogeneous. Most commonly, it contains compartments of the population whom are at different stages of the illness	Models the transmission and epidemiology of infectious disease (e.g., susceptible-infectious-recovered)

Adapted from: Tsoi et al. (2015)<sup>25</sup>

The development of a decision model requires the synthesis all of the relevant literature that pertains to the question, and that is included in the structure of the model, including parameters for the natural history of (or risk of) a disease, effectiveness and risks of alternative interventions, and health-related quality of life.<sup>20, 23</sup> Thus, modelling development often relies on much of the same information typically provided by systematic reviews, but it usually needs to be supplemented by clinically reasonable assumptions, where data may be limited or non-existent.<sup>20</sup>

Decision models are an important tool for assessing complex public health or clinical policies, and may improve the quality of health care decision making. Authors have identified areas where models can be specially relevant, such as: i) projecting out beyond the observed time horizon of an interventions, ii) extrapolate the effects to population subgroups not addressed in the available research evidence, iii) incorporate data from multiple sources, iv) evaluate relevant comparators that have not been assessed in empirical studies, and v) extrapolate intermediate outcome measures (e.g., disease free survival) to long-term or patient-centred outcomes (i.e. quality-adjusted life) (Table 3).<sup>15, 16, 20</sup>



**Table 3.** Potential application of models

Scenarios	Examples of modelling studies
The long-term effectiveness of an intervention is unclear	Life time effect on decompensated cirrhosis of obeticholic acid (a selective farnesoid X receptor agonist) as second line treatment in primary biliary cholangitis <sup>26</sup>
The outcomes of an intervention in routine care settings are unclear	Outcomes of medical management of asymptomatic patients with carotid artery stenosis typically excluded from clinical trials <sup>27</sup>
The comparative effectiveness of different interventions overall or in subgroups of patients is unclear	Comparative effectiveness of different statins and statin doses in patient groups with varying baseline cardiovascular risk <sup>28</sup>
The overall effect of an intervention at the population level, including direct and indirect effects is unknown	Effects of different vaccination strategies with serogroup C meningococcal conjugate vaccines on meningococcal carriage and disease <sup>29</sup>

Adapted from: Egger et al. (2018)<sup>30</sup>

As a summary, models, when developed in a transparent and rigorous way, may provide a systematic and explicit way to examine a decision process when there are limitations in the evidence or when there are multiple sources of evidence that require synthesis. Previous publications and task forces have described and provided guidance on good practice during modelling development, to ensure it is done appropriately including domains such as: structure, disease states, or cycle length (Table 4).<sup>24, 31</sup>

**Table 4.** Minimal criteria for a high-quality decision model

Dimensions of Quality	Attributes of “Good Practice”
Structure	Model structure should be consistent with the decision problem. The structure should be dictated by theory of disease, and not by data availability
Disease states	Model should reflect the time dependence of the disease process. States should reflect the underlying biological process of the disease and the impact of intervention. The number of states should be manageable, reflect all important aspects of disease, and not be omitted on the basis of lack of data
Options	Options and strategies should not be limited by constraints of currently accepted clinical practice A balance is needed between full range of options and keeping decision problem manageable
Time horizon	The time horizon should be sufficient to capture all important health outcomes. Lifetime time horizons will be appropriate for many models; shorter time horizons can be justified according to the disease process and effect of interventions
Cycle length (if relevant)	The length of a cycle should be the minimum interval over which pathology and/or symptoms in patients is expected to alter

Data identification	<p>“Best available” data should be referred to as “optimal available” data as it is an empirical question whether acquiring all existing evidence is a good use of resources</p> <p>Models can be used to undertake formal value of information analysis to determine the optimal data to incorporate.</p> <p>Analyst should make clear all low-cost sources have been searched for the appropriate parameter values</p> <p>Methods used for parameter identification when no data are identified should be fully detailed</p>
Data incorporation	<p>The process of data incorporation should follow accepted methods of epidemiology and statistics</p> <p>Different sources of uncertainty should be distinguished (uncertainty, heterogeneity, first- and second-order uncertainty).</p> <p>Interval rates should be translated into transition probabilities using appropriate formula</p> <p>Models should use half-cycle correction</p>
Internal consistency	The model should be checked and tested by the analyst (debugging)
External consistency	If possible, the model outputs should be compared to the results from relevant primary research studies (not used to inform model inputs)

---

Adapted from Sculpher et al. (2000)<sup>31</sup>

### 2.3 Decision models and synthesis of evidence

The use of decision models for has increased over time in the medical literature. Petitti et al reported from MEDLINE search an increase, from approximately 20 decision models’ studies published in 1980, to approximately 250 in 1997.<sup>32</sup> Another overview between 2005 and 2009 identified 1,773 articles, published between 2005 to 2009 that included the use of a decision model to assess clinical outcomes, comparing two or more strategies; 70% of them were related to treatment (pharmaceutical of procedures), 12% related to prevention and evaluation of vaccines, and 18% assessed either screening or diagnostic interventions.<sup>20</sup>

Regarding systematic reviews and health technology assessment reports. In 2009, the Agency for Healthcare Research and Quality (AHRQ) have published 11 reports (from a total of 193) that used models to support their conclusions, most of them modelled diagnostic tests or screening strategies along with subsequent treatments.<sup>33</sup> Luhn et al, searched systematic reviews of complete health economic evaluations, published between 2015 to 2017 in Medline and other economic databases, identifying 202 reviews; among them 181 included trial and model base evaluations, while 10 reviews included only model base evaluation.<sup>34</sup>

To evaluate the “state of the art” of systematic reviews (SRs) including modelling evidence, we conducted an overview of SRs (non-published data).<sup>35</sup> We identified reviews that included only modelling studies (i.e. the reviews excluded empirical studies), to inform

either the effectiveness or cost-effectiveness of any type of interventions (i.e. pharmacological, screening), published between 2018 to 2021. We identified 17 reviews, the majority from US and Europe; 35% of reviews addressed screening or prevention intervention, and 29% were related to cancer diseases (Table 5).<sup>35</sup>

The introduction of models in the field of CPG has been relatively slowly, although some relevant examples are described in the literature. For example, Egger et al reviewed 185 WHO guidelines approved from 2007 to 2015, 42 (23%) referred to modelling studies, but were rarely formally assessed as part of the body of evidence, and there was no description of quality criteria for this type of evidence.<sup>30</sup> The U.S. Preventive Services Task Force (USPSTF) has informed their screening recommendations with model results, usually involving several models, such as two models for colorectal cancer screening, five for lung cancer screening, and six for breast cancer screening.<sup>16</sup> Some of this examples are described below:

- Colorectal Cancer (CRC) Screening: the USPSTF recommendations on CRC screening were informed by 2 models, to calculate the number of life-years gained (measure of benefits), and the number of diagnostic colonoscopies (measure of harms and resource use).<sup>36</sup> CRC screening using faecal occult blood tests (FOBTs) reduces colorectal cancer deaths, but new FOBT tests such as Hemoccult SENSА and immunochemical tests are available. There are no clinical trials for these newer tests although estimates of their diagnostic performance have been published. The model evidence supported a 10-year screening interval for colonoscopy and a 1-year interval for high-sensitivity FOBTs.<sup>37</sup>
- Tuberculosis (TB) prevention: the WHO guideline development group for TB infection control, assessed systematic reviews, which included mathematical modelling studies.<sup>38</sup> One modelling study, estimated the effects of several control measures on the spread of extensively drug resistant (XDR) TB in a community in South Africa (which are ethically unfeasible to assess through RCTs). Compared with natural ventilation, the mechanical ventilatory systems would reduce XDR-TB cases by 12% (range 10%-25%), the use of respiratory masks by health workers would prevent 2% of all TB cases, and two-thirds of XDR cases in hospital staff.<sup>39</sup> The guideline development group considered the evidence for the use of ventilation systems and particulate respirators as of low quality, but suggested that these interventions are favourable for TB infection control.<sup>38</sup>

**Table 5.** Systematic reviews including only decision models from 2018 to 2021 (n=17)

Author	Year	Country	Disease	Age population	Type or intervention	# Model type	List of model type	Time Horizon
Castro	2018	Brasil	Hepatitis C	More and equal than 50 years	Pharmacologic	3	Markov, discrete event simulation, and Monte Carlo simulation model	3 months to lifetime
Leal	2018	UK	Prediabetes	NR	Preventions	4	Markov model, microsimulation, decision tree, and other	3 years to lifetime
Anothainsintawe e	2019	Thailand	Rabies	NR	Preventions	3	Decision tree model, dynamic transmission model, and simulation model	1 to 12 years
Chen	2019	China	Cardiovascular disease	NR	Digital health	2	Markov model and decision tree	0 years to lifetime
Kibret	2019	Italy	Prostate cancer	NR	External beam radiation	2	Markov and discrete event simulation	10 years to lifetime
Szilberhom	2019	Hungary	Hepatitis C	NR	Direct acting antiviral agents	6	Markov model, deterministic sensitivity analysis, decision tree, individual sampling model, discrete individual simulation – discrete event simulation in discrete time, and discrete time individual event history model	5 years to lifetime
Abreha	2019	Italy	Prostate cancer	More and equal tan 65 years	External beam radiation	2	Markov model and discrete event simulation model	10 years to lifetime
Mendivil	2019	Switzerland	Colorectal cancer	NR	Screening	4	State-transition modelling, microsimulation modelling, decision analytic model, and Archimedes model	20 years to lifetime
Jiang	2019	China	Cardiovascular disease	NR	Treatment or management	2	Markov model and decision tree	90 days to lifetime
Ran	2019	Germany	Colorectal cancer	More and equal tan 50 years	Screening	2	Markov and microsimulation model	20 years to lifetime
Canakis	2020	USA	Gastric cancer	More than 30 years	Upper endoscopy	1	Markov model	7 years to lifetime
Henrique	2020	Brazil	Schizophrenia	NR	Treatment or management	4	Markov model, decision tree, discrete event simulation, and Monte Carlo micro simulation	NR

Kemmak	2020	Iran	Venous Thromboembolism	NR	Treatment or management	3	Decision tree model, markov model, and decision analytical model	3 months to 5 years
Yao	2020	China	Inflammatory Bowel Disease	NR	Treatment or management	3	Markov model, discrete event model, and decision tree model	1 to 10 years
Trieu	2021	Australia	Osteoarthritis	More and equal than 65 years	Computer navigation	1	Markov model	120 months to 20 years
Hodkinson	2021	Australia	Herpes zoster infection	NR	Preventions	1	Dynamic transmission model	25 years to lifetime
Khan	2021	Germany	Breast cancer	More and equal than 40 years	Screening	4	Markov model, microsimulation, discrete event simulation model, and life table model	40 years to lifetime

From Canelo-Aybar et al (2022, unpublished data).<sup>35</sup> NR: not reported

- Cervical cancer screening: a WHO guideline for cervical cancer screening developed a model to inform their recommendations. The model estimated the proportions of TP, TN, FP and FN findings for each of the screening tests (VIA, HPV and cytology) based on the test-accuracy estimates and the pre-test probability of having cervical intraepithelial neoplasia.<sup>40</sup> Then, they calculated the probability of developing any of the critical outcomes for decision-making based on the treatment they may receive and the estimates of efficacy and potential complications of the different treatments (cryotherapy, CKC and LEEP).<sup>40</sup> Finally, they suggested to screen (test and treat) with an HPV strategy over cytology followed by colposcopy.

As those examples' underlines, findings from systematic reviews may not apply directly to the guideline development setting. Habbema et al based in the experience by the USPSTF described some scenarios where models can bridge the gap between empirical evidence and the guideline setting (Table 6).<sup>16</sup> The WHO conducted a workshop with experts in the CPG and modelling field and proposing similar scenarios of how to use models appropriately: 1) in the absence of empirical data directly addressing the question of interest, for example in the context of public health programmes where RCTs are rarely available. 2) where immediate action is needed but little direct empirical evidence is available, for example in the Ebola, or Zika epidemics. 3) with a systematic and transparent approach to identifying existing models that may be relevant, and careful consideration of commissioning new models.<sup>30</sup>

**Table 6.** Areas where models can bridge the gap between primary evidence and guideline development

- Applying new information on disease risk, prognosis, medical technologies, and treatments to estimate changes in health outcomes.
- Exploring the full array of alternative intervention strategies.
- Assessing important benefits and harms over the lifetime of the population.
- Making appropriate assumptions about attributes of the target population and health care setting for the guideline conditions.

Adapted from Habbema et al. (2014)<sup>16</sup>

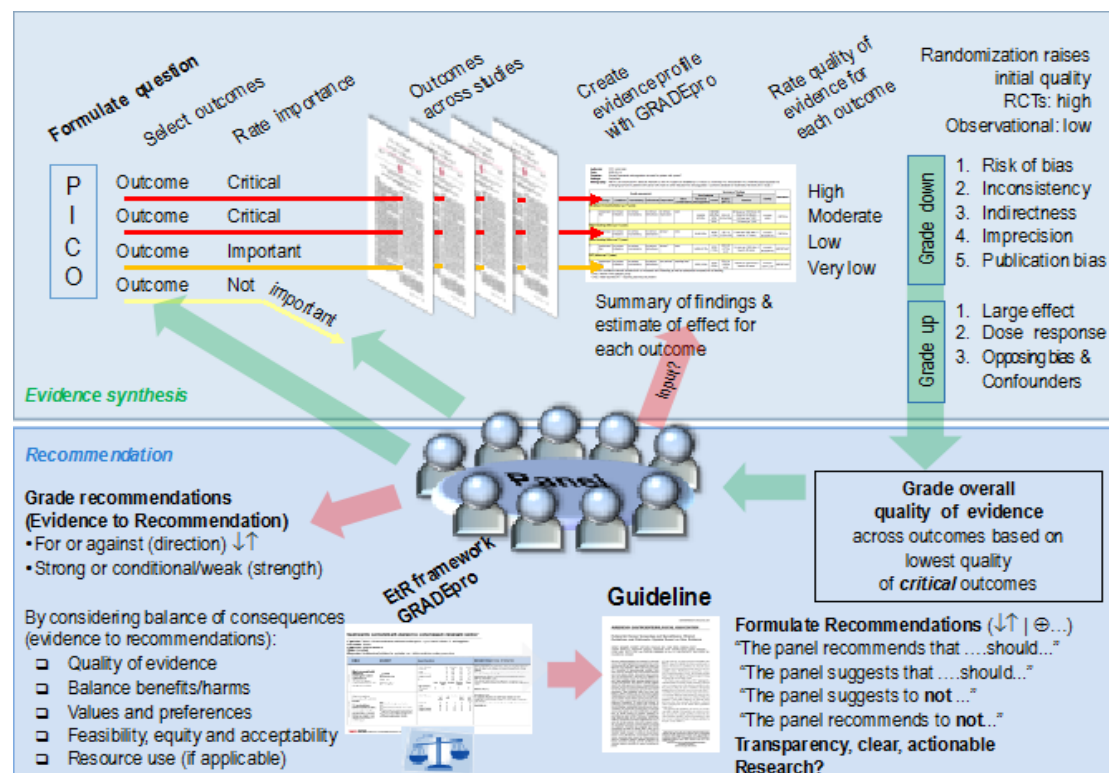
## 2.4 The GRADE approach and certainty of modelling evidence

The GRADE Working Group was established in the year 2000 and continues as a community striving to create systematic and transparent approach for assessing and communicating the certainty of the available evidence used in making recommendations in health care—related

disciplines.<sup>7</sup> The GRADE Working Group now includes over 600 members from 40 countries. GRADE is widely used internationally by over 110 organizations to address topics related to clinical medicine, public health, coverage decisions, health policy, and environmental health, examples include the Cochrane Collaboration, the World Health Organization (WHO) and international societies such as the European Respiratory Society (ERS) or Infectious Disease Society of America (IDSA).<sup>21</sup>

The GRADE approach offers a transparent and structured process for developing and presenting evidence summaries for systematic reviews and guidelines in health care and for carrying out the steps involved in developing recommendations (Figure 2).<sup>7, 10</sup> Some of the steps the GRADE approach specifies includes: framing questions, choosing outcomes of interest and rating their importance, evaluating the certainty of evidence, and incorporating evidence on aspects such as the balance of effects, values and preferences, resource use for equity when developing recommendations.<sup>9</sup> To support guideline developers, the GRADE working group has also develop a check list to guide the overall development process.<sup>41</sup>

**Figure 2.** Outlined of the process of reviewing the evidence and developing recommendations using the GRADE approach



Adapted from: Schünemann et al. (2013)<sup>42</sup>

The GRADE approach considers four levels for the certainty of evidence (Table 7) from very low to high, each of them have a different implication for our confidence on the estimates of

effect.<sup>43</sup> The domains to evaluate and rating the certainty of evidence includes: the risk of bias, directness of evidence, precision of an estimate, consistency of estimates across studies, risk of bias related to selective reporting and factor to increase our confidence. Specific approaches to the concepts may differ depending on the nature of the body of evidence, but they usually follow the concepts described below:

- Risk of bias: the certainty is lower in the estimated effect if the studies had inherent limitations in the design or conduct of the study.
- Imprecision: the certainty would be lower if the clinical decision is likely to be different if the true effect was at the upper versus the lower end of the confidence interval. The certainty may be rated down if the effect estimate comes from only one or two small studies or if there are few events.
- Inconsistency the certainty would be higher when several studies show consistent effects; when assessing whether or not rated down for inconsistency, reviewers should inspect the similarity of point estimates and the overlap of their confidence intervals.
- Indirectness: the certainty would be higher when studies directly compare the interventions of interest in the population of interest, and report the outcome(s) critical for decision-making.
- Publication bias: this domain requires making inferences about missing evidence, some statistical methods are helpful in detecting publication bias. Publication bias is typically more common with observational data and when most studies are funded by industry.
- increases confidence in the evidence: in rare circumstances, certainty in the evidence can be rated up (Table 2) for example: i) when there is a very large magnitude of effect, ii) when there is a clear dose-response gradient, or iii) when residual confounding is likely to decrease rather than increase the magnitude of effect.



**Table 7.** GRADE certainty of evidence levels and meaning

Certainty	Definition
Very low	We have very little confidence in the effect estimate: The true effect is likely to be substantially different from the estimate of effect
Low	Our confidence in the effect estimate is limited: The true effect may be substantially different from the estimate of the effect
Moderate	We are moderately confident 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
High	We are very confident that the true effect lies close to that of the estimate of the effect

Adapted from: Balshem et al. (2011)<sup>43</sup>

The strength of a recommendation reflects the extent to which a guideline panel is confident that desirable effects of an intervention outweigh undesirable effects (or vice versa) in the patients for whom the recommendation is intended. The GRADE approach defines two categories of the strength of a recommendation (strong and weak).<sup>8</sup> A guideline panel would issue a strong recommendation if they are certain about the various factors that influence the strength of a recommendation such as a clear balance towards either the desirable (to recommend an action) or undesirable effects (to recommend against an action).<sup>7-9</sup> Otherwise, if a guideline panel is uncertain whether the balance of effects or the various factors that influence the strength of a recommendation (Table 8), a guideline panel would like to make a weak recommendation.

**Table 8.** Criteria that contribute to the strength of a recommendation

Factors
Is the problem a priority?
How substantial are the desirable anticipated effects?
How substantial are the undesirable anticipated effects?
What is the overall certainty of the evidence of effects?
Is there important uncertainty about or variability in how much people value the main outcomes?
Does the balance between desirable and undesirable effects favour the intervention or the comparison?
How large are the resource requirements (costs)?
What is the certainty of the evidence of resource requirements (costs)?
Does the cost effectiveness of the intervention favour the intervention or the comparison?

What would be the impact on health equity?

Is the intervention acceptable to key stakeholders?

Is the intervention feasible to implement?

Adapted from: Alonso-Coello et al. (2016)<sup>8,9</sup>

A major barrier for the incorporation of evidence from modelling studies into guidelines development is the perceived complexity of the methods to construct and analyse these studies, that there are no widely agreed approaches to the evaluation of the certainty of estimates from modelling studies, and their integration with primary data to inform guidelines recommendations.<sup>30</sup> These statements are also reflected on our findings (unpublished data) on the quality of SRs including only modelling studies from 2018 to 2021, as only two reviews (12%) assessed the quality of studies with an appropriate tool (i.e. Philips<sup>44</sup> or Jaime Caro tools<sup>45</sup>) while most reviews did not assess the quality or used inappropriate tools (Table 8) and none of them assessed or made statement about the certainty of evidences of the estimates.<sup>46</sup>

**Table 9.** Characteristics of included systematic reviews (N=17)

Characteristics	N (%)
<b>Protocol registration</b>	
No	13 (76.47)
Yes	4 (23.53)
<b>Use of method review guidelines</b>	
Not reported	13 (76.47)
Cochrane guideline	1 (5.88)
Other	3 (17.65)
<b>Use of reporting guideline</b>	
No or it was done with an inappropriate tool	7 (41.18)
PRISMA	8 (47.06)
CHEERS	2 (11.76)
<b>Use of quality assessment method</b>	
No or it was done with an inappropriate tool	15 (88.24)
Yes, with an appropriate tool	2 (11.76)
<b>Use of certainty of evidence assessment</b>	
No or it was done with an inappropriate tool	17 (100.00)
Yes, with an appropriate tool	0 (0.00)
<b>Number of databases included</b>	5.06 ± 2.36
<b>Language restriction</b>	
No	7 (41.18)
Yes	10 (58.82)
<b>Duplicate screening</b>	
No	13 (76.47)
Yes	4 (23.53)
<b>Duplicate full-text selection</b>	
No	13 (76.47)

Yes	4 (23.53)
<b>Disagreement</b>	
Not reported	7 (41.18)
Third reviser	6 (35.29)
Consensus	4 (23.53)
<b>PRISMA flowchart</b>	
No	5 (29.41)
Yes	12 (70.59)
<b>Number of model studies included</b>	18.29 ± 15.32
<b>Type of data synthesis</b>	
Quantitative	2 (5.88)
Qualitative	16 (94.12)
<b>Number of types of models included</b>	2.82 ± 1.43

From Canelo-Aybar et al (unpublished data)<sup>46</sup>

International organization have also recognized the need for developing standard methods to incorporate modelling evidence to guidelines and in particular to assess the certainty of evidence.<sup>15, 16, 20, 30</sup> The WHO after a workshop and survey to experts in the field, proposed that guidelines groups might include modelling studies as an additional study category, in addition to RCTs and observational studies currently defined in GRADE approach domains for rating the certainty of evidence should be tailored to modelling studies by adding or omitting questions and developing specific guidance (Table 10).<sup>30</sup>

**Table 10.** WHO recommendation on how to adapt the GRADE approach to modelling evidence

- The certainty of the evidence for modelling studies should be assessed and presented separately in summaries of the evidence (GRADE evidence profiles), and classified as high, moderate, low, or very low certainty.
- GRADE dimensions of certainty (imprecision, indirectness, inconsistency and publication bias) and the criteria defined for their assessment are also relevant to modelling studies.
- For modelling studies, the concept of the 'credibility' of the model, which takes the structure of the model, input data, dimensions of uncertainty, as well as transparency and validation into account, is more appropriate than 'study limitations' or 'risk of bias'.
- When summarizing the evidence, a distinction should be made between observed and modelled outcomes.

Adapted from: Egger et al. (2018)<sup>30</sup>

---

## JUSTIFICATION

---

### 3. JUSTIFICATION

---

Decision analysis models can be broadly defined as “mathematical framework representing variables and their interrelationships to describe observed phenomena or predict future events”.<sup>22</sup> In general, models might vary in their structure and degree of complexity. For example, a simple model may be any calculation to estimate a single variable not directly measured, as when we are interested on the population impact of an intervention which are usually estimated as the product of their relative effect (informed by a trial), multiplied by the baseline risk of the population of interest (informed by a cohort study).<sup>21</sup>

Decision models are important tools for assessing clinical policies and may improve the quality of health decision making.<sup>20</sup> Some areas where models can be specially relevant includes: i) projecting out beyond the observed time horizon of an interventions, ii) extrapolate the effects to population subgroups not included within a study, iii) incorporate data from multiple sources, iv) evaluate relevant comparators that have not been assessed in empirical studies, v) extrapolate intermediate outcome measures (e.g., disease free survival) to long-term or patient-centred outcomes (i.e. quality-adjusted life).<sup>16, 20, 30</sup>

Although the relevance of modelling studies to bridge the gap between evidence and guideline settings,<sup>16</sup> they are not routinely incorporated in CPG development. Egger et al reviewed 185 WHO guidelines approved from 2007 to 2015, 42 (23%) referred to modelling studies, but these studies were rarely formally assessed as part of the body of evidence and there was no description of quality criteria for this type of evidence.<sup>30</sup> Thus, a major barrier for the incorporation of evidence from modelling studies into guidelines development is the perceived complexity of the methods to analyse these studies as there are no widely agreed approaches for evaluation of the certainty of estimates from modelling studies, and their integration with primary data to inform guidelines recommendations.<sup>30</sup>

The GRADE approach offers a transparent and structured process for developing and presenting evidence summaries for systematic reviews and guidelines in health care and for carrying out the steps involved in developing recommendations.<sup>7, 10</sup> Some of the steps the GRADE approach specifies includes: framing questions, choosing outcomes of interest and rating their importance, evaluating the certainty of evidence, and incorporating evidence with considerations of the balance of effects, values and preferences of patients and society or resource use to arrive at recommendations.<sup>10</sup>

In general, the GRADE approach may be applicable irrespective of health discipline as It has been applied (with specific guidance and modifications) to rating the certainty of evidence

for management interventions, diagnostic tests,<sup>47, 48</sup> prognosis studies,<sup>49</sup> animal studies,<sup>50</sup> use of resources and cost-effectiveness evaluations,<sup>51</sup> and values and preferences.<sup>52, 53</sup> Institutions like the WHO have underlined the urgent need for developing methodologies on how the results of modelling studies should be included in the process of developing recommendations, the evaluation of certainty of modelling estimates, and on their integration to inform guidelines and recommendations.<sup>30</sup>

---

## **OBJECTIVES**

---

## 4. OBJECTIVES

---

### **General objectives**

To develop methods to improve the use of modelling evidence during the development of health care decision making.

### **Specific objectives**

- i) To provide insights in how to integrate empirical and modelling evidence to inform the effectiveness of a public health intervention in the context of clinical guideline development.
- ii) To describe how a model can be developed and used to complement the evidence from empirical studies and assist a guideline panel in the recommendation formulation process.
- iii) To provide guidance on how to assess the certainty of evidence of models estimates using the GRADE approach and describe a framework of how to incorporate the modelling evidence in the guideline development process.



---

## **METHODS**

---

## 5. METHODS

---

This doctoral thesis is organized in the form of a compendium of publications. Therefore, the methods described are those corresponding to each of the studies carried out. The design of each study was determined in order to provide experiences in the incorporation of the results (evidence) from models for the formulation of health recommendations, as well as the development of methods for the evaluation of the certainty in the evidence of this type of studies.

### **5.1 First study: Benefits and harms of annual, biennial, or triennial breast cancer mammography screening for women at average risk of breast cancer: a systematic review for the European Commission Initiative on Breast Cancer (ECIBC)**

#### **Design**

A systematic review, integrating empirical studies (randomized clinical trials and observational studies) along with modelling studies to inform clinical outcome from multiple strategies of population mammography screening for breast cancer (BC).

This systematic review informed the recommendations about mammography screening intervals for women of average breast cancer risk<sup>54, 55</sup> of the European Guidelines on Breast Cancer Screening and Diagnosis launched by European Commission Initiative on Breast Cancer (ECIBC) (the publication containing all recommendation available in the *appendix 1*)

#### **Structured question and outcome prioritization**

The Guideline Development Group, including multidisciplinary experts on BC screening, framed the clinical question as “*Should an annual, biennial or triennial screening frequency be used for screening asymptomatic women?*”. The review focused on the three age subgroups for which the European Guidelines previously issued recommendations for screening (45 to 49, 50 to 69, and 70 to 74 years old).

#### **Data sources and searches**

We searched MEDLINE (via PubMed), EMBASE (via Ovid) and CENTRAL (via The Cochrane Library) databases using pre-defined algorithms for individual studies up to April 2020.

#### **Study selection**

We included studies of the following designs:

- I. Randomized clinical trials (RCTs),

- II. Observational studies such as cohorts, time trend (before-after), or analysis of population surveillance registries, and
- III. Decision analytic models including at least two screening intervals in one of the age groups of interest

We excluded studies of women at high risk for BC, i.e. having known susceptibility gene mutations (BRCA1/BRCA2), a history of previous BC, exposure to chest irradiation or having a direct family member with breast cancer.

### **Data extraction and risk of bias assessment**

We extracted the study's details on design, patient population (simulated), setting, screening method, follow-up, mammography intervals and results. We assessed the risk of bias (or credibility for modelling studies) with the following tools:

- (I) Cochrane Risk of Bias Assessment tool for RCTs <sup>56</sup>
- (II) The Risk of Bias in Non-randomised Studies of Intervention (ROBINS-I) for observational studies <sup>57</sup>
- (III) The Questionnaire to Assess Relevance and Credibility of Modelling Studies (the ISPOR-AMCP-NPC Good Practice Task Force) for modelling studies<sup>45</sup>

### **Data analysis**

We summarized the results narratively, and we did not attempt to conduct a meta-analysis for empirical studies because there were not enough studies across age groups to be meaningful or because several publications reported the same population data at overlapping time periods.

Modelling studies reported the incremental number of events for each screening interval compared to a non-screening scenario. For some studies, we calculated the number of events by subtracting overlapping age groups (i.e. to obtain events in annual screening in women 45 to 49 years old, we subtracted the estimates in women 50 to 69 from the larger group of 45 to 69). Across the different studies, we presented the range of the absolute difference of events per each pairwise screening interval comparison.

### **Certainty of the evidence**

We rated the certainty of the evidence, as high, moderate, low or very low, for each outcome based on the standard GRADE approach for RCTs and observational studies.<sup>58, 59</sup>

To apply the GRADE approach to modelling studies, we considered the certainty would depart from the lowest certainty of the bodies of evidence that informed the main inputs in the model. We used the credibility and relevance items from the ISPOR-AMCP-NPC tool to inform the judgments for the risk of bias and indirectness domains.<sup>45</sup>

## **5.2 Second study: Recommendations from the European Commission Initiative on Breast Cancer for multigene testing to guide the use of adjuvant chemotherapy in patients with early breast cancer, hormone receptor positive, HER-2 negative**

### **Design**

A systematic review of the clinical impact of using of multigene tests to guide the decision to provide adjuvant chemotherapy on women with early breast cancer. After, we identified relevant gaps in the available evidence for relevant outcomes, the panel member of the Guideline Development Group decided to develop a decision-tree model to estimate the downstream consequences of the multigene test (also described in the publication).

This systematic review and the decision tree model informed the discussion on the guideline panel for issuing the recommendations about multigene testing for women early breast cancer at diagnosis,<sup>54, 55</sup> from the European Guidelines on Breast Cancer Screening and Diagnosis launched by European Commission Initiative on Breast Cancer (ECIBC).

### **Structured question and outcome prioritization**

The clinical question was framed as: *“Should multigene tests be used in patients who have HoR-positive, HER-2 negative, lymph node-negative or up to 3 lymph nodes-positive invasive breast cancer to guide the use of adjuvant chemotherapy”*.

### **Data sources and searches**

We searched MEDLINE (via PubMed), EMBASE (via Ovid) and CENTRAL (via The Cochrane Library) databases using pre-defined algorithms for individual studies up to April 2020 up to October 2018.

### **Study selection**

We included studies of the following designs:

- (I) Randomized controlled trials (RCT) and,
- (II) cohort studies (including pooled analyses of studies), either from prospective or retrospective analysis, of stored specimen samples

Studies must have applied any of the four tests as predictive markers for guiding the use of adjuvant chemotherapy (Supplementary Fig. 1). A predictive marker identifies the differential benefit of a treatment based on the marker status. Thus, we included the following assessment approaches: a) *Marker-based strategy design*: patients are assigned to a treatment arm depending on whether they received treatment, b) *Treatment interaction design*: patients are divided into groups based on the marker status (i.e. high and low marker status), the predictive value is assessed by observing the relative efficacy of treatment differences between marker status and treatment assignments.

### **Data extraction and risk of bias assessment**

Two reviewers independently assessed risk of bias and extracted the following information: study design, inclusion and exclusion criteria, number of patients, age, participants' characteristics and prioritized outcomes.

The risk of bias of the included RCTs was assessed using the Cochrane Risk of Bias tool for randomized trials.<sup>56</sup> Cohort studies were assessed with the "Risk Of Bias In Non-randomized Studies - of Interventions-I" (ROBINS-I) tool.<sup>57</sup>

### **Data analysis**

Descriptive statistics were used to summarize the characteristics of the included patients across studies. The effect measures for prioritized outcomes and their corresponding 95% confidence intervals (CIs) were reported as presented in individual studies.

### **Development of a *de novo* model**

We develop a *deterministic decision tree model* to estimate the downstream consequences of testing patients with the multigene tests versus different scenarios of usual care. The model complemented the empirical evidence for only the two multigene tests for which predictive evidence was identified. We provided different scenarios (as sensitivity analysis), and did not include discounting to the clinical effects.

### **Certainty of the evidence**

The certainty of evidence per outcome and overall certainty was rated using the GRADE approach. For each recommendation, the GDG received a Summary of Findings (SoF) table and a first draft of an evidence to decision framework (EtD).<sup>9</sup> We did not assess the certainty from the model developed for the panel meeting.

### **5.3 Third study: GRADE Guidelines 30: the GRADE approach to assessing the certainty of modeled evidence—An overview in the context of health decision-making**

#### **Design**

Development of a methodological guideline based on an iterative consultation process to expert in multiple fields related to modelling evidence (decision analysis models as well as other types of mathematical models like toxicology or environmental models).

#### **Development process**

On May 15 and 16, 2017, the GRADE modelling project group lead a workshop in Hamilton, Ontario, Canada, to develop common principles for the application of the GRADE assessment of certainty of evidence to modelled outputs. The National Toxicology Program of the Department of Health and Human Services in the United States of America and the MacGRADE Center in the Department of Health Research Methods, Evidence, and Impact at McMaster University sponsored the workshop which was co-organized by MacGRADE Center and ICF International.

Workshop participants were selected to ensure a broad representation of all modelling related fields. Participants had expertise in modelling in the context of clinical practice guidelines, public health, environmental health, dose—response modelling, physiologically based pharmacokinetic (PBPK) modelling, environmental chemistry, physical/chemical property prediction, evidence integration, infectious disease, computational toxicology, exposure modelling, prognostic modelling, diagnostic modelling, cost-effectiveness modelling, biostatistics, and health ethics.

Participants addressed specific tasks in small groups and large group discussion sessions and agreed on key principles both during the workshop and through written documents. In summary, the workshop participants suggested an approach to incorporate model outputs in health-related decision-making and the principles to assess the certainty of evidence for modelling evidence.

---

## RESULTS

---

## 6. RESULTS

---

The results for this thesis are those corresponding to each study published on peer-reviews journal of high impact. In brief, our findings are organized in the following thematic sequence.

### First study.

- Reference: Canelo-Aybar C, Posso M, Montero N, Solà I, Saz-Parkinson Z, Duffy SW, Follmann M, Gräwingholt A, Giorgi Rossi P, Alonso-Coello P. *Benefits and harms of annual, biennial, or triennial breast cancer mammography screening for women at average risk of breast cancer: a systematic review for the European Commission Initiative on Breast Cancer (ECIBC)*. Br J Cancer. 2022 Mar;126(4):673-688
- Author position: First author and co-corresponding
- Journal: British Journal of Cancer
- Scimago Journal Ranking: Q1
- 5-year impact factor: 7.57

### Second study

- Reference: Giorgi Rossi P\*, Lebeau A\*, Canelo-Aybar C, Saz-Parkinson Z, Quinn C, Langendam M, McGarrigle H, Warman S, Rigau D, Alonso-Coello P, Broeders M, Graewingholt A, Posso M, Duffy S, Schünemann HJ; ECIBC Contributor Group. *Recommendations from the European Commission Initiative on Breast Cancer for multigene testing to guide the use of adjuvant chemotherapy in patients with early breast cancer, hormone receptor positive, HER-2 negative*. Br J Cancer. 2021 Apr;124(9):1503-1512. \*Co-first author
- Author position: Second author
- Journal: British Journal of Cancer
- Scimago Journal Ranking: Q1
- 5-year impact factor: 7.57

### Third Study.

- Reference: Brozek JL\*, Canelo-Aybar C\*, Akl EA, Bowen JM, Bucher J, Chiu WA, Cronin M, Djulbegovic B, Falavigna M, Guyatt GH, Gordon AA, Hilton Boon M, Hutubessy RCW, Joore MA, Katikireddi V, LaKind J, Langendam M, Manja V, Magnuson K, Mathioudakis AG, Meerpohl J, Mertz D, Mezencev R, Morgan R, Morgano GP, Mustafa R, O'Flaherty M, Patlewicz G, Riva JJ, Posso M, Rooney A, Schlosser PM, Schwartz L, Shemilt I, Tarride JE,



Thayer KA, Tsaioun K, Vale L, Wambaugh J, Wignall J, Williams A, Xie F, Zhang Y, Schünemann HJ; GRADE Working Group. *GRADE Guidelines 30: the GRADE approach to assessing the certainty of modeled evidence-An overview in the context of health decision-making*. J Clin Epidemiol. 2021 Jan;129:138-150. \*Co-first author

- Author position: First author
- Journal: Journal of Clinical Epidemiology
- Scimago Journal Ranking: Q1
- 5-year impact factor: 7.30

### **6.1 First study: Benefits and harms of annual, biennial, or triennial breast cancer mammography screening for women at average risk of breast cancer: a systematic review for the European Commission Initiative on Breast Cancer (ECIBC)<sup>46</sup>**

We included evidence from 25 studies (27 publications) comprising one RCT,<sup>60, 61</sup> 13 observational studies<sup>62-75</sup> and 11 modelling studies.<sup>76-86</sup> Our finding suggested that in women of average BC risk, screening intervals may have different trade-offs between benefits and harms across age group. For example, among women 50 to 69 years old, compared to biennial screening, annual screening may have additional benefits which should be balanced against an important increase in false-positive results; whereas among women aged 70 to 74 longer screening intervals (i.e. triennial) probably obtain a more favourable overall balance of benefits and harms than in other age groups.

#### **Studies' characteristics**

We identified only one RCT conducted between 1989 and 1996 in the United Kingdom which randomly allocated 99,389 women aged 50 to 62 to either annual or triennial screening.<sup>61</sup> From the 11 observational studies, nine studies performed a secondary analysis from surveillance mammography registries which were linked to the Surveillance, Epidemiology, and End Results (SEER) pathology registries;<sup>62, 64, 66, 68, 70, 71, 73</sup> one quasi-experimental study included women aged 40-49 invited for mammography screening every year or every 3 years;<sup>72</sup> one study compared two time periods, before and after a change from annual to biennial mammography for women aged 50 to 79;<sup>63</sup> and two studies included women from screening programs provided through medical centres from the US.

Six of studies implemented microsimulation models developed within the Cancer Intervention and Surveillance Modelling Network (CISNET) collaboration varying in the structures and assumptions.<sup>87</sup> (Model D: Dana-Farber,<sup>88</sup> Model E: Erasmus,<sup>89</sup> Model GE: Georgetown-Einstein,<sup>90</sup> Model M: MD Anderson,<sup>91</sup> Model S: Stanford,<sup>92</sup> and Model W: Wisconsin-Harvard).<sup>93</sup> The remaining four modelling studies implemented non-individual models. One transition model evaluated annual versus biennial screening intervals in Japan.<sup>80</sup> One Markov model assessed breast cancer deaths averted and overdiagnosis due to screening for women in the United Kingdom,<sup>77</sup> and another study applied the model developed by Preston to estimate radiation related events<sup>82</sup>. We obtained non-publicly available data of a transition modelling study simulating an Spanish cohort of women.<sup>84, 94</sup>

#### **Benefits and harms in women aged 45 to 49**

##### *Observational studies*

Evidence was available only for women 40 to 49 years. One study suggested an increase in the risk of BC mortality in annual versus triennial screening (RR 1.14; 95%CI 0.59 to 2.19)<sup>72</sup> while the odds of advanced BC stage (IIB-IV) at diagnosis may be higher in women exposed to biennial screening compared to annual screening (OR 1.17; 95%CI 0.93 to 1.46)<sup>70</sup> The 10-year probability of false positive biopsy recommendation was 11.4% (95%CI 10.5%-12.4%) with annual screening, 5.9% (95%CI 5.6%-6.2%) with biennial screening, and 3.9% (95%CI 3.7%-4.1%) with triennial screening.<sup>71</sup>

#### *Modelling studies*

One study implementing six models, estimated a median of 30 more deaths averted and 480 additional QALYs per 100,000 women undergoing annual screening compared to biennial screening in the US population;<sup>76</sup> overdiagnosis was higher with annual screening compared to biennial screening.<sup>76</sup> Another modelling study assessed the risk of radiation induced adverse events, estimating 14 more induced BC and 2 more deaths per 100,000 women with annual screening compared to biennial screening.<sup>78</sup>

### **Benefits and harms in women aged 50 to 69**

#### *Randomized Clinical Trials*

In the UKCCR study, over a median of 162 months of follow-up, annual screening may decrease the risk of BC mortality compared to triennial screening (RR = 0.89, 95% CI 0.73–1.07).<sup>60</sup>

#### *Observational Studies*

One study comparing the period before and after mammography screening changed from annual to biennial found there may be little to no difference in mortality (MR 1.06; 95%CI 0.76, 1.46) or interval cancer (RR 0.98; 95%CI 0.90-1.06) between the two-time periods.<sup>63</sup> Miglioretti et al. found there may be no difference in the risk of advanced BC stage (IIB-IV) in the age groups 50-69 and 60-69 with annual versus biennial screening.<sup>70</sup>

The 10-year probability of a false positive result was 55.2% (95%CI 54.8%-55.7%) with annual screening, 35.4% (95%CI 35.0%-35.7%) with biennial screening, and 24.8% (95%CI 24.5%-25.2%) with triennial screening.<sup>71</sup> The cumulative 10-year probability of having a false positive biopsy recommendation was 9.7% (95%CI 9.3%-10.1%) with annual screening, 5.4% (95%CI 5.2%-5.6%) with biennial screening, and 3.7% (95%CI 3.6%-3.9%) with triennial screening.<sup>71</sup>

#### *Modelling studies*

One Canadian modelling study estimated that the BC deaths averted for annually, biennially or triennially screening compared to no screening would be 740, 520 and 400, respectively.<sup>83</sup> In another study, the number of BC deaths averted per 100,000 screened women with scattered fibroglandular breast density, was 690, 520 and 400 for annual, biennial and triennial screening<sup>79</sup> and the number of QALYs gained was 6,000, 4,700 and 3,600, respectively.<sup>79</sup> A microsimulation model for the German population found consistent result to the aforementioned studies.<sup>86</sup>

The estimated overdiagnosis was greater with more frequent screening intervals, a microsimulation model study estimated 2,900, 2,000 and 1,600 for annual, biennial and triennial screening compared to no screening per 100,000 women.<sup>79</sup> A microsimulation model estimated 27 radiation induced BC cases with biennial screening and 49 with annual screening.<sup>78</sup> The attributed number of radiation related deaths was four with biennial screening and seven with annual screening.<sup>78</sup>

### **Benefits and harms in women aged 70 to 74**

#### *Observational studies*

Three studies provided data on advanced BC stage (IIB-IV) at diagnosis but for different age ranges (i.e. 66 to 89,<sup>62</sup> 70 to 85<sup>70</sup> and 70 to 89 years<sup>73</sup>). For women 70 to 85 the odds of stage IIB-IV were no different among those exposed to biennial or annual screening (OR 0.98 95%CI 0.76-1.27).<sup>70</sup> The 10-year cumulative probability of false positive results for women between the ages of 75 to 89 may be higher with annual screening (47%, 95%CI 44.9% to 49.5%) compared to biennial screening (26.6%, 95%CI 25.7% to 27.5%),<sup>62</sup> a similar trend for false positive biopsy recommendations was reported.<sup>62</sup>

#### *Modelling studies*

The estimated difference for BC deaths between the different intervals might be small. A microsimulation model estimated the number of BC deaths averted for annual, biennial and triennial screening to be 100, 90 and 80, respectively, compared to no screening per 100,000 screened Canadian women.<sup>83</sup> Only one non-individual based model estimated overdiagnosis for this age group and it showed a small increasing trend with shorter screening intervals from 193 for triennial screening to 269 for annual screening.<sup>94</sup>

### **Risk of bias and certainty of the evidence**

Our GRADE assessment for modelling studies departed from low certainty after considering some methodological limitations from the input evidence (i.e. indirectness due

mammography sensitivity estimated from BCSC registries including women from wider age groups than our clinical question<sup>95</sup>) and the credibility assessment of the development of the included models which was limited due to suboptimal reporting. We had concerns about indirectness given that most models used observational data from the US to inform their input parameters (i.e. radiation induced BC), and because in one modelling study, data was only available by levels of breast density (i.e. scattered fibro glandular density.<sup>79</sup>

## ARTICLE OPEN



## Epidemiology

## Benefits and harms of annual, biennial, or triennial breast cancer mammography screening for women at average risk of breast cancer: a systematic review for the European Commission Initiative on Breast Cancer (ECIBC)

Carlos Canelo-Aybar<sup>1,2,3,4</sup>, Margarita Posso<sup>1,4</sup>, Nadia Montero<sup>1</sup>, Ivan Solà<sup>1,2</sup>, Zuleika Saz-Parkinson<sup>5,6</sup>, Stephen W. Duffy<sup>7</sup>, Markus Follmann<sup>8</sup>, Axel Grawingholt<sup>9</sup>, Paolo Giorgi Rossi<sup>9</sup> and Pablo Alonso-Coello<sup>1,2</sup>

© European Union 2021

**BACKGROUND:** Although mammography screening is recommended in most European countries, the balance between the benefits and harms of different screening intervals is still a matter of debate. This review informed the European Commission Initiative on Breast Cancer (ECIBC) recommendations.

**METHODS:** We searched PubMed, EMBASE, and the Cochrane Library to identify RCTs, observational or modelling studies, comparing desirable (BC deaths averted, QALYs, BC stage, interval cancer) and undesirable (overdiagnosis, false positive related, radiation related) effects from annual, biennial, or triennial mammography screening in women of average risk for BC. We assessed the certainty of the evidence using the GRADE approach.

**RESULTS:** We included one RCT, 13 observational, and 11 modelling studies. In women 50–69, annual compared to biennial screening may have small additional benefits but an important increase in false positive results; triennial compared to biennial screening may have smaller benefits while avoiding some harms. In younger women (aged 45–49), annual compared to biennial screening had a smaller gain in benefits and larger harms, showing a less favourable balance in this age group than in women 50–69. In women 70–74, there were fewer additional harms and similar benefits with shorter screening intervals. The overall certainty of the evidence for each of these comparisons was very low.

**CONCLUSIONS:** In women of average BC risk, screening intervals have different trade-offs for each age group. The balance probably favours biennial screening in women 50–69. In younger women, annual screening may have a less favourable balance, while in women aged 70–74 years longer screening intervals may be more favourable.

*British Journal of Cancer* (2022) 126:673–688; <https://doi.org/10.1038/s41416-021-01521-8>

## INTRODUCTION

Breast cancer (BC) is the second most prevalent cancer in the world and the most frequent among women [1]. In the European Union, 404,920 women were diagnosed with BC and 98,755 women died during 2018 [2]. Despite these high rates, the mortality risk of BC has decreased over the last decades due to improvements in treatment, services quality, and to early diagnosis linked to the implementation of population-based screening programmes [3]. However, there is still ongoing research and debate on how to best implement BC screening

programmes, including which is the optimal mammography screening interval.

Published recommendations on mammography screening frequencies vary among organisations. The National Health Service Breast Screening Program (NHSBSP) of the United Kingdom, recommends screening every 3 years to women aged 50–70 (47–73 in England) [4]. The United States Prevention Services Task Force (USPSTF) recommends biennial mammography for women aged 50–74 and making a case by case decision for women in their 40s [5]. The American Cancer Society recommends annual

<sup>1</sup>Iberoamerican Cochrane Centre - Department of Clinical Epidemiology and Public Health, Biomedical Research Institute Sant Pau (IB-Sant Pau), Barcelona, Spain. <sup>2</sup>CIER de Epidemiología y Salud Pública (CIERESP), Madrid, Spain. <sup>3</sup>Department of Paediatrics, Obstetrics and Gynaecology, Preventive Medicine, and Public Health PhD Programme in Methodology of Biomedical Research and Public Health, Universitat Autònoma de Barcelona, Bellaterra, Barcelona, Spain. <sup>4</sup>Department of Epidemiology and Evaluation, IMM (Hospital del Mar Medical Research Institute), Barcelona, Spain. <sup>5</sup>European Commission, Joint Research Centre (JRC), Ispra, Italy. <sup>6</sup>Wolfson Institute of Preventive Medicine, Queen Mary University of London, London, UK. <sup>7</sup>German Cancer Society, Berlin, Germany. <sup>8</sup>Radiologie am Theater, Padernborn, Germany. <sup>9</sup>Epidemiology Unit, Azienda Unità Sanitaria Locale - IROCC di Reggio Emilia, Reggio Emilia, Italy. <sup>✉</sup>email: Carlos.canelo-aybar@gmail.com; Zuleika.SAZ-PARKINSON@ec.europa.eu

Received: 9 October 2020 Revised: 20 June 2021 Accepted: 30 July 2021  
Published online: 26 November 2021

Published on Behalf of CRUK



screening between the ages of 45 and 54 (with the option of starting annual screening between 40 and 44), and screening every two years from age 55 or continue annually if the woman is in good health and expected to live ten more years [6].

Previous studies have suggested that the balance between benefits and harms for different screening intervals might vary depending on the age subgroup. A modelling study found that for every 1000 women aged 50–74, biennial screening avoided seven BC deaths, while annual screening had similar benefits but caused more harms [7]. Observational data from the US Breast Cancer Surveillance Consortium (BCSC) registries, observed that premenopausal women undergoing biennial screening had more BC lesions with less favourable prognostic characteristics compared to those having annual screening [8].

In 2015, the European Commission Initiative on Breast Cancer (ECIBC) was launched to develop the European Guidelines on Breast Cancer Screening and Diagnosis. This article describes the systematic review that informed the recommendations about mammography screening intervals for women of average breast cancer risk in three separate age subgroups [9, 10]. During the guideline development process [9], the Guidelines Development Group (GDG) made detailed considerations about the balance between desirable and undesirable effects [9], values and preferences, equity, acceptability and feasibility; these considerations are described in the published methodology and summary of recommendations [9, 10] (<https://healthcare-quality.jrc.ec.europa.eu/european-breast-cancer-guidelines/screening-ages-and-frequencies>).

## METHODS

### Structured question and outcome prioritisation

The clinical question 'Should an annual, biennial or triennial screening frequency be used for screening asymptomatic women?' was prioritised by the GDG (Box 1: Structured clinical question). This review focused on the three age subgroups for which the European Guidelines previously issued recommendations for screening (45–49, 50–69, and 70–74 years old). The GDG prioritised the outcomes using a 1–9 scale (7–9 critical; 4–6 important; 1–3 of limited importance) [11].

### Data sources and searches

We initially searched MEDLINE (via PubMed, October 2016), EMBASE (via Ovid, October 2016) and CENTRAL (via The Cochrane Library, October 2016) databases using pre-defined algorithms for individual studies. We updated our initial search in MEDLINE (via PubMed) and EMBASE (via Ovid) in April 2020 (Supplementary Table S1: Protocol Systematic Review, Supplementary Table S2: Search strategy).

### Study selection

We included studies published in English of the following designs: (i) randomised clinical trials (RCTs), (ii) observational studies such as cohort,

time trend (before-after), or analysis of population surveillance registries, and (iii) decision analytic models (hereafter referred to as modelling studies) (Supplementary Tables S3a and S3b). All studies included at least two screening intervals in one of the age groups of interest; screening intervals from observational studies should have been defined based on at least two examinations prior to diagnosis; modelling studies should have assumed 100% adherence to the screening programmes and applied no discounting to the effects. Due to sparse empirical evidence in the 45–49 age subgroup, we included RCTs and observational studies that recruited women from 40 to 49.

We excluded studies of women at high risk for breast cancer, i.e. having known susceptibility gene mutations (BRCA1/BRCA2), a history of previous breast cancer or lobular neoplasia, exposure to chest irradiation (other than diagnostic imaging over that anatomical area) or having a direct family member with breast cancer.

Pairs of reviewers (CCA, MP), after calibration, assessed eligibility and reviewed the full text of the selected references. Discrepancies were resolved either by consensus or with the help of a third reviewer.

### Data extraction and risk of bias assessment

CCA and MP independently extracted details of the study design, patient population, setting, screening method, follow-up, mammography intervals and results. If needed, we requested additional data from the authors. We assessed the risk of bias (or credibility for modelling studies) with the following tools: (i) for RCTs with the Cochrane Risk of Bias Assessment tool [12] (ii) for observational studies with the Risk of Bias in Non-randomised Studies of Intervention (ROBINS-I) [13], (iii) for modelling studies with the Questionnaire to Assess Relevance and Credibility of Modelling Studies (the ISPOR-AMCP-NPC Good Practice Task Force) [14].

### Data analysis

We prioritised observational studies reporting the longest observation time when different studies used the same surveillance or clinical registries from an identical population covering overlapping time periods. We prioritised the more direct evidence for a European population of average risk when data was stratified by women's characteristics (i.e. white women instead of other ethnic groups).

Modelling studies reported the incremental number of events for each screening interval compared to a non-screening scenario. For some studies, we calculated the number of events by subtracting overlapping age groups (i.e. to obtain events in annual screening in women 45–49 years old, we subtracted the estimates in women 50–69 from the larger group of 45–69). We used the estimates for women with scattered fibroglandular breast density when they were only reported by breast density categories. Across the different studies, we presented the range of the absolute difference of events per each pairwise screening interval comparison.

We did not attempt to conduct a meta-analysis of relative risks (RR) or odds ratios (OR) from empirical studies because there were not enough studies across age groups to be meaningful or because several publications reported the same population data at overlapping time periods.

Box 1. PICO structured clinical question

Population	Intervention	Comparison	Outcomes
Women who are at average risk of breast cancer: • 45–49 years • 50–69 years • 70–74 years	Annual, biennial or triennial mammography screening (film or digital)	Another interval (annual, biennial or triennial)	1. Breast cancer mortality 2. Incidence of interval cancer 3. Stage of breast cancer 4. Radiation induced breast cancers 5. Deaths due to radiation induced breast cancers 6. Quality of life 7. False positive related adverse effects 8. Overdiagnosis

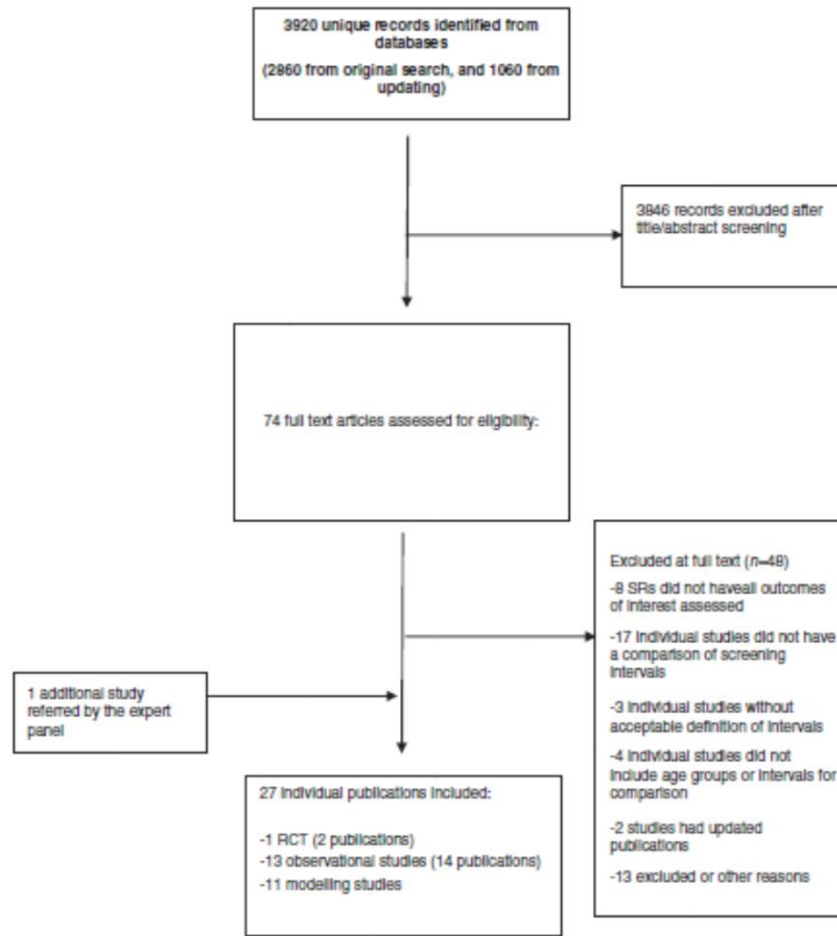


Fig. 1 PRISMA flow diagram of literature search and selection.

#### Certainty of the evidence

We rated the certainty of the evidence, as high, moderate, low or very low, for each outcome based on the standard GRADE approach for RCTs and observational studies [15, 16]. To apply the GRADE approach to modelling studies, we considered the certainty would depart from the lowest certainty of the bodies of evidence that informed the main inputs in the model. We used the credibility and relevance items from the ISPOR-AMCP-NPC tool to inform the judgments for the risk of bias and indirectness domains.

As is customary in systematic reviews, we adopted a partially contextualised approach to rate the certainty of evidence, this means that for a point (or range) estimate of a single outcome we assessed our certainty that the true effects lie within the boundaries of what we consider a trivial, small, medium or large effect without considering the evidence from other outcomes [17]. During the development of recommendations, guideline panel members might consider our results using a contextualised approach which means considering the evidence from other critical outcomes (i.e. whether the benefits are consistent across outcomes) when rating the certainty for a single outcome [17].

#### RESULTS

##### Search results

We included 22 studies from 2860 unique citations in our initial evidence synthesis in October 2016 which was used to develop the ECIBC recommendations. After the updated search in April, 2020, we included 3 additional studies comprising a total of 25 studies (from 27 publications) during both periods: one RCT [18, 19], 11 modelling studies [7, 20–29] and 13 observational studies (Fig. 1) [8, 30–42]. The list of excluded studies and reasons for exclusion are described in Supplementary Table S4.

##### Studies' characteristics

We provide here a summary of the study design, and the main results for only the three age groups of interest. When there is empirical data (from observational or RCTs) we rely primarily on those estimates instead of simulated number of events from modelling studies. To interpret the modelling estimated events,



we must consider that they represent the estimated events for a cohort of individuals from the time of screening until death or during the individual's lifetime (or other given time point). The estimated 10-year probability of false positive or false biopsy recommendation in the observational studies were estimated using a previously described statistical model [43]. A detailed reporting of the results from studies covering larger age groups (i.e. 66–74 years) can be found in Supplementary Table S3a, b.

The only available RCT was conducted between 1989 and 1996, the United Kingdom Co-ordinating Committee on Cancer Research (UKCCCR) trial of Breast Screening Frequency and randomly allocated 99,389 women aged 50–62 to either annual or triennial screening [19]. Of the women originally invited to either of the screening arms, 38,492 (77%) attended triennial screening and 37,530 (76%) attended annual screening. The primary end point was predicted mortality based on two validated risk-models. However, as the UKCCCR published observed data for survival up to the end of 2006, we reported these estimates in our assessment [18].

Nine studies performed analysis from surveillance systems data of the United States which differed in the time periods covered and the age group of the women included. Eight studies used national Breast Cancer Surveillance Consortium (BCSC) mammography registries which were linked to the Surveillance, Epidemiology, and End Results (SEER) pathology registries [8, 30, 32, 34, 37, 38, 40]. One study used the Vermont Breast Cancer Surveillance System (VBCSS) from the state of Vermont [33]. The studies included two types of analysis: first a case series of invasive BC that were used to evaluate the association between screening intervals and adverse tumour characteristics, and secondly, they estimated the 10-year cumulative probabilities of false positive results and false positive biopsy recommendations (Table 1) [43].

A quasi-experimental study included women aged 40–49 who were invited to attend a screening programme in Finland. Those women born in an even calendar year were invited for mammography screening every year, while those born in an odd calendar year were invited to screening every 3 years [39]. One study conducted a comparative analysis of two time periods in British Columbia-Canada, before and after 1997, year when the Screening Mammography Program of British Columbia (SMPBC) changed its policy from annual to biennial mammography for women aged 50–79 [31].

Two studies included women from screening programmes at medical centers from the US. The first performed a retrospective analysis of data from women who chose to attend either annual or biennial mammography examinations in a screening programme of the University of California San Francisco Medical Center [35]. The second study was a retrospective cohort of women without previous diagnosis of BC who attended a routine screening examination at Columbia University Medical Center in New York; the screening interval was defined using the time elapsed since their previous exam according to their electronic clinical records [42] (Table 1).

Six studies used microsimulation models developed within the Cancer Intervention and Surveillance Modelling Network (CISNET) collaboration: Model D (Dana-Farber) [44], Model E (Erasmus) [45], Model GE (Georgetown-Einstein) [46], Model M (MD Anderson) [47], Model S (Stanford) [48], and Model W (Wisconsin-Harvard) [49]. Each of these models has its own characteristics which are described elsewhere [50], they vary in the model structures and assumptions such as factors conditioning screen detection, individual risk factors or allowing spontaneous regression of ductal carcinoma in-situ (DCIS) lesions [51]. Four studies assessed mammography screening intervals for the U.S. population reporting the median estimates from two to six models [7, 21, 22, 24]. Two studies simulated screening for a Canadian population based on an adaptation of Model W [26, 28]. One

microsimulation study projected adverse events related to radiation exposure from mammography exams in women 50–74 years of age (Table 2) [21]. One additional study adapted a microsimulation Markov model to the German context to assess annual, biennial, and triennial routine screening in women aged 50–69 [29].

The remaining four modelling studies implemented non-individual models. One transition model evaluated annual versus biennial screening intervals in Japan [23]. One Markov model assessed breast cancer deaths averted and overdiagnosis due to screening for women in the United Kingdom [20], and another study applied the model developed by Preston to estimate radiation related events [25]. We obtained non-publicly available data of a transition modelling study for a Spanish cohort described elsewhere (Table 2) [27, 52].

#### Benefits and harms in women aged 45–49 (Tables 3/4)

**Observational studies.** A Finnish study suggested an increase in the risk of BC mortality in annual versus triennial screening (incidence RR 1.14; 95%CI 0.59–2.19) although the estimate was very uncertain [39]. The odds of advanced breast cancer stage (IIb–IV) may be higher in women with a history of biennial screening compared to annual screening (OR 1.17; 95%CI 0.93–1.46) among incident breast cancers from US registries [37].

In women of normal weight, the 10-year probability of false positive results was 11.2% (95%CI 9.8–12.8%) with annual screening and 6.0% (95%CI 5.4–6.6%) with biennial screening [32]. The probability of a false positive biopsy recommendation was 11.4% (95%CI 10.5–12.4%) with annual screening, 5.9% (95%CI 5.6–6.2%) with biennial screening, and 3.9% (95%CI 3.7–4.1%) with triennial screening among white women [38].

Moreover, indirect evidence from the wider age group of women (40–79) suggested that the incidence of interval cancers may be lower among annually screened (0.07%) compared to biennially screened (0.15%) women, but it was very uncertain given the small number of events [35].

**Modelling studies.** One study estimated, across six microsimulation models, a median of 30 more deaths averted per 100,000 women undergoing annual screening compared to biennial screening in the US population [7], while the median number of additional QALYs gained with annual screening was 480 more compared to biennial screening [7]. In the same modelling study, the overdiagnosis estimation was higher with annual screening compared to biennial screening [7]. One modelling study assessed the risk of radiation induced adverse events in this age group and found that annual screening yielded 14 more induced BC and 2 more deaths per 100,000 screened women compared to biennial screening [21].

#### Benefits and harms in women aged 50–69 (Tables 3/4)

**Randomised clinical trials.** Duffy et al. reported in the UKCCCR study, over a median of 162 months of follow-up, that annual screening may decrease the risk of BC mortality compared to triennial screening among attenders to the prevalent screening (RR = 0.89, 95% CI 0.73–1.07) [18]. Moreover, there was a small difference in the size of the tumour at diagnosis, with a major proportion of them being 10mm or smaller in the annual screening group compared to the triennial group (25% vs. 19%) [18, 19].

**Observational studies.** One study in a province of Canada comparing the period before and after mammography screening changed from annual to biennial found there may be little to no difference in mortality (MR 1.06; 95%CI 0.76, 1.46) or interval cancer (RR 0.98; 95%CI 0.90–1.06) between the two-time periods [31].

Miglioretti et al. found there may be no difference in the risk of advanced BC stage (IIb–IV) in the age groups 50–69 (adjusted RR 0.98;

Table 1. Characteristics of the clinical trials, and observational studies identified in the literature search.

Author, year	Country, period	Design/screening intervals	No enrolled	Inclusion/exclusion criteria	Age ranges, (years)	Outcomes of interest
<i>Randomized clinical trial</i>						
BSFTG, 2002 [19], Duffy, 2008 [18]	United Kingdom 1989–2006	Attender women to prevalent screening at NHS breast screening program invited to conventional (3 years) interval or to three annual screenings.	-Annual: 37,530. -Triennial: 38,492.	I: women attending prevalent screening. E: women with BC diagnosed prior to the trial.	50–62	-BC mortality. -Interval cancer.
<i>Observational studies</i>						
Braithwaite, 2012	United States 1999–2006	Data from five BCSC registries (those matched to Medicare claims) linked to SEER programs/tumor registries. -BC case series: women with incident DCIS or invasive BC; interval groups defined as 1 (9–18 months) or 2 (19–30 months) years, based on the two most recent mammograms prior diagnosis. -FP-cohort all first and subsequent screening mammography from 1999 to 2006, without BC diagnosis after 1 year of last examination.	-Annual (BC cases): 1227. -Biennial (BC cases): 453. -FP-Cohort: 137,949.	I: women with at least two mammograms.	66–89	Reported by co-morbidity score categories: -Stage of BC (IB+). -FP results. -FP recommendations.
Coldman, 2008 [31]	Canada	Pre-post screening policy changing evaluation. Data from the SMBC linked to VSA registries over two time periods. From 1988 to 1997 women 50–79 years recommended annual screening. 1998–2005 changed to a biennial recommendation.	-Annual (before July 1996): 152,226 -Biennial (July 1996 or after): 184,764	E: women with a prior BC diagnosis not eligible to attend SMBC.	50–79	-BC mortality. -Interval cancers.
Dittus, 2013 [32]	United States 1994–2008	Data from seven BCSC registries linked to SEER programs/tumor registries. -BC case series: women with incident BC; interval groups defined as 1 (9–18 months) or 2 (19–30 months) years based on the time between two most recent mammograms prior diagnosis. -FP-cohort all screening mammography from 1994 to 2008, without BC diagnosis after 1 year of last examination.	-Annual (BC cases): 2766. -Biennial (BC cases): 1666. -FP-Cohort: 555,343	I: women with at least two mammograms before BC diagnosis. E: history of BC diagnosis, reporting hormone therapy use.	40–74	Reported by BMI categories: -Stage of BC (IB+). -FP results. -FP recommendations.
Gail, 2007 [33]	United States 1994–2002	Data from the VBCSS which collects information from patients, radiologists and hospital pathology. -BC case series: women with incident BC; interval groups defined as 1 (0.75–1.49 years), 2 (1.5–2.49 years) years based on the time between two most recent mammograms prior diagnosis.	-Annual (BC cases): 1236 -Biennial (BC cases): 439	I: women with at least two mammograms before BC diagnosis. E: intervals of less than 273 days between mammograms. History of BC diagnosis.	>40 years	



Table 1 continued

Author, year	Country, period	Design/screening intervals	No enrolled	Inclusion/exclusion criteria	Age ranges (years)	Outcomes of interest
Hubbard, 2011 [34]	United States, 1994–2006	Data from seven BCSC registries linked to SEER programs/tumor registries. -BC case series: women with incident invasive BC; interval groups defined as 1 (9–18 months) or 2 (19–30 months) years based on the time between two most recent mammograms prior diagnosis. -FP-cohort: screening mammograms from 1994 to 2004 or 2007 (depending on the registry).	-Annual (BC case): 36,445 -Biennial (BC case): 27,775 -FP-Cohort: 169,456	I: women with at least two mammograms. E: women with BC at or after 60 years.	40–59	-BC stage (IIB+). -FP results. -BC recommendations.
Hunt, 1999 [35]	United States 1985–1997	Retrospective analysis from prospectively collected data of women that choose annual or biennial screening mammography performed by University of California San Francisco Medical Center at screening mammography mobile van.	-Annual: 19,905 -Biennial: 4306	I: previous normal screening mammography, asymptomatic physician referred women from six contiguous counties. E: –	40–79	-Interval cancers
Kořilowski, 2013 [8]	United States 1994–2008	Data from BCSC registries linked to SEER programs/tumor registries. -BC case series: women with incident DCIS or invasive BC; interval groups defined as 1 (0.75–1.49 years) or 2 (1.5–2.49 years) years based on the time between two most recent mammograms prior diagnosis. -FP-cohort: all first and subsequent screening mammography from 1994 to 2008, without BC diagnosis after 1 year of last examination.	-Annual (BC case): 7039 -Biennial (BC case): 3476 -Triennial (BC case): 959 -FP-Cohort: 922,624	I: women with diagnosis of incident invasive or DCIS BC and at least 2 prior mammograms. E: History of BC diagnosis.	40–74	Reported by breast density categories: -BC stage(IIB+). -FP results. -BC recommendations.
Parvinen, 2011 [39] Kern, 1997 [36]	Finland 1987–2003	Population based, quasi-experimental. Mailed screening invitation to women aged 50 or more at biennial interval. Women less than 50 years, and born even-numbered year invited to annual screening and those born odd-numbered year were invited to triennial screening.	-Triennial: 6,926 -Annual: 7839	I: – E: –	40–49	-BC mortality. -Interval cancer.
McGuinness, 2018 [42]	United States 2014–2015	Retrospective cohort. Women were approached during routine screening mammography at Columbia University Medical Center in New York. Annual interval from 9 to 18 months, biennial from 18 to 30 months. More than 913 days (3+ years) were considered non-compliant. Less than 274 days was considered recall imaging.	-Annual: 1399 -Biennial: 335	I: no previous diagnosis of BC, age ≥18 years	<50 years ≥50 years	-FP results.
Miglioretti, 2015 [37]	United States 1996–2012	Data from seven BCSC registries linked to SEER programs/tumor registries. -BC case series: women with incident	-Annual (BC case): 12,070	I: women with at least two mammograms.	40–85	Reported by menopausal and HT use categories:

Table 1 continued

Author, year	Country, period	Design/screening intervals	No enrolled	Inclusion/exclusion criteria	Age ranges, (years)	Outcomes of interest
O'Meara, 2013 [38]	United States 1994–2008	DCIS or invasive BC; intervals groups defined as 1 (11–14 months) or 2 (23–26 months) years based on the time between two most recent mammograms prior diagnosis. Data from seven BCSC registries linked to SEER programs/tumor registries. -BC case series: women with incident DCIS or invasive BC; interval groups defined as 1 (9–18 months) or 2 (>18–30 months), or 3 (>30–42 months) years based on the time between two most recent mammograms prior diagnosis. -FP cohort all screening mammography from 1994 to 2008, without BC diagnosis after 1 year of last examination.	-Annual (BC case): 8876 -Biennial (BC case): 3370 -Annual (BC case): 4265 -Triennial (BC case): 1255 -FP cohort: 1,276,312	I: women with at least two mammograms.	40–74	Reported by breast density and ethnic categories: -BC stage (I–IV). -Interval cancer. -FP results. -BC recommendations.
Sanderson, 2015 [41]	United States 1995–2000	Data from Medicare claims and SEER. -BC case series: women with incident BC; groups defined based on mammography screening periodicity over the 4 years prior to diagnosis as (a) no or irregular mammography screening, (b) biennial mammography, and (c) annual mammography.	-Irregular (BC case): 29,712 -Biennial (BC case): 11,227 -Annual (BC case): 23,355	I: non-Hispanic white or black ethnicity, complete Medicare coverage during 4-year before BC diagnosis; primary BC diagnosed between 69 to 84 years. E: BC diagnosed by autopsy or death certificate, stage IV cancer.	69–84	Reported by ethnicity category (non-Hispanic white or black women). -Mortality among BC cases (according to screening interval history).
White, 2004 [40]	United States 1996–2001	Data from seven BCSC registries linked to SEER programs/tumor registries. -BC series: women with incident DCIS or invasive BC; intervals groups defined as 1 (9–18 months) or 2 (>18–30 months) years based on the time between two most recent mammograms prior diagnosis.	-Annual (BC case): 5400 -Biennial (BC case): 2440	I: women with diagnosis of invasive or DCIS BC and at least two prior mammograms. E: history of BC.	40–69	Reported by breast density: -BC stage (I–IV). -Interval cancer.

SEER: Surveillance Epidemiology and End Results; BCSC: Breast Cancer Surveillance Consortium; VBCSS: Vermont Breast Cancer Surveillance System; DCIS: ductal carcinoma in situ; BC: breast cancer; MTS: National Health Services; FP: false positive; Bx: biopsy; I: inclusion criteria; E: exclusion criteria.

Table 2. Characteristics of decision analysis studies identified.

Author, year	Modelled population	Design/screening interval	Strategies	Parameters	Years of screening	Outcomes of interest
Non-individual based models						
Gurney, 2014 [20]	United Kingdom	Markov cohort simulation model. Healthy, preclinical non-progressive in situ, preclinical progressive in situ, preclinical invasive, diagnosed in situ, and diagnosed invasive breast cancer by NPI category, death from BC, and death from other causes.	Six strategies defined by: -Starting age: 40, 47, 50 years. -Interval: triennial, annual, and hybrid (annual for 40–47, biennial thereafter). -Screening scenarios: stopped at 70 or 73 years.	-Q1: up to 85 years old. -Q2: on uptake, sensitivity, and exposure time. -Q3: difference in the cumulative incidence of invasive in situ cancer. -Q4: NHS breast cancer program.	40–73	-Breast cancer deaths averted -Overdiagnosis -QALYs
Tsurumaki, 2015 [28]	Japan, United States	Transition cohort model to simulate impact of screening for the Japanese and US population. The source of age distributions were data from the Japanese Breast Cancer Society and the BCSC and National Cancer Data Base respectively.	Twelve strategies defined by: -Starting age: 40, 50 years -Interval: annual, biennial -Screening stopping at 65, 74, and 79 years.	-Q1: NE -Q2: NE -Q3: mortality rate of undetected BC -Q4: SE: 81.3%; 57.9%; 44.4% (Japan)	40–79	-Breast cancer deaths averted -PP results
Vilagines, 2014 [27]*	Spain	Stochastic transition model. Estimation of the Lee and Zelen model to estimate incidence and prevalence.	Twenty strategies defined by: -Starting age: 40, 45, 50. -Interval: annual, biennial -Screening stopping at 65, 70, 74 and 79.	-Q1: born from 1946 to 1993 -Q2: time horizon was 40–79 years -Q3: NE -Q4: not provided by authors -Q5: SE: 0.55 for 40–45 years, 0.70 for 45–50 years, 0.75 for 50–55 years and 0.80 for >70 years -Q6: include anxiety and PP results	40–79	-Breast cancer deaths averted -Overdiagnosis -QALYs -PP results -Design breast bc
Yaffe, 2011 [23]	Canada	Model by Preston (exact absolute risk of radiation induced BC). Applied to Canadian population of 2002. Digital mammography.	Six strategies defined by: -Starting age: 40, 50 -Interval: annual, biennial (hybrid annually in 40s, biennial thereafter) -Screening stopping at 45, 59 years.	-Q1: screening began up to 109 years -Q2: using relative model instead of absolute model, biennial years, survival rates.	40–74	- Radiation induced BC - Radiation induced BC deaths
Individual based models						
Amick, 2019 [29]	Germany N = 3,000,000 women	A microsimulation-Markov model included 6 health states: healthy (no breast cancer); ductal carcinoma in situ (DCIS); localized, regional, or distant invasive breast cancer; and death.	Three regular screening strategies and additional strategies based on individual risk assessment (not shown) -Interval: annual, biennial, triennial -Starting age: 30 years -Screening stopping: 69 years	-Q1: from age of 30, until the end of life or 100 years. -Q2: specific treatment based on hormone receptors -Q3: digital mammography sensitivity based in BCSC -Q4: unknown and probabilistic sensitivity analysis (eg, DCIS incidence, invasive cancer incidence, invasive cancer mortality)	-QALY -Bopsy after false positive screening	
Mayes-Baird, 2016 [7]	United States N = 1,000 women	Six micro simulation models developed within the CONET collaboration: model D, model E, model G, model M, model S and model W. Updating of models include 1) portrayal of molecular subtypes based on BR and HBB status, current population incidence, digital screening, and uptake therapies.	Eight strategies defined by: -Starting age: 40, 45, or 50 years -Interval: annual, biennial and hybrid (annual in 40s, biennial thereafter). All strategies stop screening at 74.	-Q1: born in 1970 of average-risk and average breast density -Q2: from age 25 years until death or age 100 -Q3: models assume proportions of DCIS non-progressive, models M and W assumed some non-progressive invasive hormone receptors -Q4: specific treatment based on hormone receptors -Q5: digital mammography sensitivity based in BCSC -Q6: SE: 0.72 (95%CI 0.65–0.75)	40–74	Reported as median across models: -BC deaths averted -BC deaths averted -Overdiagnosis -QALYs -PP results -Design breast bc
Mijlmeester, 2016 [21]	United States N = 100,000 women	Two micro simulation modeling approaches for digital mammography, MCA/MR/DA model and a new model for radiation exposure (which accounts for repeated mammography or radiation exposure and BS). Based on radiation induced BC using the results from Preston.	Eight strategies defined by: -Starting age: 40, 45, or 50 years -Interval: annual, biennial and hybrid (annual in 40s, biennial thereafter). All strategies stop screening at 74.	-Q1: born in 1970 of average-risk and average breast density -Q2: from age 25 years until death or age 100 -Q3: models assume proportions of DCIS non-progressive, models M and W assumed some non-progressive invasive hormone receptors -Q4: specific treatment based on hormone receptors -Q5: digital mammography sensitivity based in BCSC -Q6: SE: 0.72 (95%CI 0.65–0.75)	40–74	-BC deaths averted - Radiation induced BC - Radiation induced BC deaths
Mitmann, 2018 [28]**	Canada N = 2,000,000 women	One modified microsimulation, from the perspective of the Ontario public health care system, and one microsimulation, from the perspective of the Ontario public health care system. Based on the results from the CONET collaboration. Data were modified against US data and modified against Canadian data.	Eleven screening scenarios: -Annual, biennial, triennial, and hybrid (annual in 40s, biennial thereafter). All strategies stop screening at 74.	-Q1: specific treatment based on individual risk assessment -Q2: MCA model based in 1960 validated against US data and modified against Canadian data.	40–74	-QALYs

Table 2 continued

Author, year	Modelled population	Design/screening interval	Strategies	Parameters	Years of screening	Outcomes of interest
Trentmann-Oliver, 2016 [22]	United States N = 1000	Populations were stratified by the lives of women at 6-month intervals.	Screening stopping at: 40, 69 or 74 years.  Six screening scenarios: -Annual, biennial, or triennial digital mammography. -Starting age: 50 or 65 (adjusted biennial from 50 to 64). -Stopping age: 74.	-Q1: lifetime horizon -Q2: 1000 women screened for key resources in one-way analysis -MA3: digital mammography sensitivity based in BCC.  -IC: born in 1970 -O7: from age 25 years until death or age 100 -O5: models assume proportions of DCIS non-pregnant models W assumed biennial from 50 to 64 -T1: specific treatment based in hormone receptors -MA3: digital mammography accuracy based in BCC.	40–74	Reported as median across models (justified by BO) -BC deaths averted. -Overdiagnosis. -QALYs (includes quality-adjusted life expectancy) -pp results -Design breast box.
Van Ravenna, 2012 [24]	United States N = 1000 women	Four micro simulation models developed within the CSNET collaboration: model D, model E, model G6 and model W. The models were based on a cohort of women 50 to 74 with biennial screening for women 50 to 74. Model G6 was based on a cohort of women 50 to 74 with screening interval and biennial and screening method (film and digital).	Five screening scenarios -Interval: annual and biennial -film or digital mammography -Age group: 40–49 years -All effects were assumed incremental effects compared to 50–74 screening.	-T1: specific treatment based in hormone receptors -IC: born in 1960 of average-risk -MA3: digital and film mammography accuracy based in BCC.	40–49	-BC deaths averted. -pp results.
Yaffe, 2013 [26]	Canadian, N = 2,000,000 women	One model from the CSNET collaboration (model W), adapted to the Canadian context. Treatment effectiveness was implemented on a control case model. The model allowed different proportion of hormone receptors	Eleven screening scenarios: -Interval: annual, biennial, triennial (and two hybrid scenarios) -starting age: 40 or 50 years -stopping age: 69 or 74 years.	-O7: from age 40 years until death or age 99. -IC: born in 1960 of average-risk.	40–74	-BC deaths averted. -pp results.

Model D: Dana-Farber Cancer Institute Boston Masachusetts; Model E: Erasmus Misco-Fada; University Medical Center Rotterdam, the Netherlands; Model GE: Georgetown University Medical Center Washington, DC; and Albert Einstein College of Medicine, Bronx, New York; Model H: MD Anderson Cancer Center, Houston, Texas; Model W: University of Wisconsin, CSNET, Cancer Intervention and Surveillance Modeling Network BCSC, Breast Cancer Surveillance Consortium; EB: oestrogen receptor; HER2: human epidermal growth factor receptor 2; NP: Nottingham prognostic index; QALY: quality adjusted life years; SA: sensitivity analysis; (C): women born cohort; (F): time of follow-up; horizon time; (M): mammography accuracy; (ME): mammography effectiveness; (O): overdiagnosis; (P): tailored treatment; (R): radiation dose; BC: breast cancer; FP: false positive. Unpublished data were provided by the authors.

\*A previous study by the same authors and using the same model and population was excluded (Mittmann 2015) as the updated study provided a more detailed description of the outcomes.



Table 3. Summary of the desirable and undesirable effects from RCTs and observational studies for different screening intervals and age groups.\*

Age group	Annual vs. Biennial	Triennial vs. Biennial	Annual vs. Triennial	Certainty of evidence
	N° of studies, countries	N° of studies, countries	N° of studies, countries	
	Relative effect (95% CI)	Relative effect (95% CI)	Relative effect (95% CI)	
	Absolute reduction (95% CI)	Absolute reduction (95% CI)	Absolute reduction (95% CI)	
<b>Breast cancer (BC) mortality</b>				
40–49	–	–	RR 1.14 (0.59 to 2.19)**	3 more (7 fewer to 21 more)
50–69	1 (Canada) [31]	–	RR 0.98 (0.76 to 1.12)	42 fewer (144 fewer to 72 more)
<b>BC stage (IB–IV)</b>				
40–49	1 (US) [37]	1 (US) [38]	–	Very low for all comparisons
50–69	1 (US) [37]	1 (US) [38]	–	Very low for all comparisons
70–74	1 (US) [37]	–	–	Very low
<b>Interval cancer</b>				
40–49	1 (US) [35]	–	–	Very low
50–69	1 (US) [37]	1 (US) [7]	1 (US) [7]	Very low for all comparisons
70–74	1 (US) [37]	–	–	Very low
<b>False positive results—10 year cumulative probability per woman</b>				
40–49	1 (US) [32]	1 (US) [32]	1 (US) [38]	Very low for all comparisons
50–69	1 (US) [32]	1 (US) [38]	1 (US) [38]	Very low for all comparisons
70–74	1 (US) [30]	–	–	Very low
<b>False positive biopsy recommendation—10 year cumulative probability per woman</b>				
40–49	1 (US) [32]	1 (US) [32]	1 (US) [38]	Very low for all comparisons
50–69	1 (US) [32]	1 (US) [38]	1 (US) [38]	Very low for all comparisons
70–74	1 (US) [30]	–	–	Very low

To review the reference for each study and the reasons for downgrading the certainty of the evidence see Supplementary file Table S4–S12.

\*Only the study with the longest time of observation was included when there were several publications with overlapping time periods. When studies provided results stratified by women's characteristics, we extracted data from subgroups more similar to European context (i.e. white women instead of other ethnic groups).

\*\*We calculated the confidence interval from the raw data reported in the publication as the original interval was not consistent with the main effect and lower interval bound.

\*\*\*Randomized clinical trial study.

Table 4. Summary of desirable and undesirable effects (number of events) from modelling studies for different screening intervals and age groups (per 100,000 screened women)\*\*

Age group	Annual vs. Biennial	Triennial vs. Biennial	Annual vs. Triennial	Certainty of evidence across comparisons
	N° of studies, countries	N° of studies, countries	N° of studies, countries	
	N° of events per arm	N° of events per arm	N° of events per arm	
	Absolute reduction	Absolute reduction	Absolute reduction	
Breast cancer deaths averted (BC)				
45-49*	2 (US) [7, 21]**	1 (Spain) [27]	1 (Spain) [27]*	Very low for all comparisons
	A: 70-90 B: 39-40	T: 47 B: 52	5 fewer	14 fewer
50-69	3 (Canada, Japan, Spain) [23, 26, 27]	2 (Canada, Spain) [26, 27]	2 (Canada, Spain) [26, 27]	Very low for all comparisons
	A: 631-870 B: 426-705	T: 397-400 B: 426-520	120 fewer to 29 fewer	234 more to 340 more
70-74*	2 (Canada, Spain) [26, 27]	2 (Canada, Spain) [26, 27]	2 (Canada, Spain) [26, 27]	Very low for all comparisons
	A: 100-142 B: 90-146	T: 80-136 B: 90-146	10 fewer to 9 fewer	6 more to 20 more
Oesophageal				
45-49*	2 (Spain, US) [7, 27]	1 (Spain) [27]	1 (Spain) [27]	Very low for all comparisons
	A: 143-200 B: 0-119	T: 88 B: 119	31 fewer	55 more
50-69	1 (Spain) [27]	1 (Spain) [27]	1 (Spain) [27]	Very low for all comparisons
	A: 904 B: 609	T: 500 B: 609	109 fewer	404 more
70-74*	1 (Spain) [27]	1 (Spain) [27]	1 (Spain) [27]	Very low for all comparisons
	A: 269 B: 236	T: 193 B: 236	43 fewer	76 more
QALY				
45-49*	2 (Spain, US) [7, 27]	1 (Spain) [27]	1 (Spain) [27]	Very low for all comparisons
	A: 727-1540 B: 665-1060	T: 653 B: 665	12 fewer	74 more
50-69	3 (Canada, Germany, Spain) [27-29]	3 (Canada, Germany, Spain) [27-29]	3 (Canada, Germany, Spain) [27-29]	Very low for all comparisons
	A: 4400-7100 B: 3900-5000	T: 3100-4336 B: 3900-5000	1200-328 fewer	1100 to 3100 more
70-74*	2 (Canada, Spain) [26, 27]	2 (Canada, Spain) [26, 27]	2 (Canada, Spain) [26, 27]	Very low for all comparisons
	A: 336-600 B: 427-500	T: 300-308 B: 427-500	200 fewer to 29 fewer	63 fewer to 300 more
Fake positive results				
45-49*	2 (Spain, US) [7, 27]	1 (Spain) [27]	1 (Spain) [27]	Very low for all comparisons
	A: 9150-56,700 B: 86301-26,700	T: 4831 B: 6301	1470 fewer	4,319 more
50-69	3 (Canada, Japan, Spain) [23, 27, 29]	2 (Canada, Spain) [26, 27]	2 (Canada, Spain) [26, 27]	Very low for all comparisons
	A: 42,606-152,600 B: 29,039-80,500	T: 24,253-49,900 B: 29,039-80,500	19,600 fewer to 4787 fewer	18,354 to 82,900 more
70-74*	2 (Canada, Spain) [26, 27]	2 (Canada, Spain) [26, 27]	2 (Canada, Spain) [26, 27]	Very low for all comparisons
	A: 5766-24,100 B: 3459-17,400	T: 2295-12,700 B: 3459-17,400	4700 fewer to 1104 fewer	3,471 more to 11,800 more
Benign biopsy recommendations				
45-49*	2 (Spain, US) [7, 27]	1 (Spain) [27]	1 (Spain) [27]	Very low for all comparisons
	A: 400-5000 B: 208-3000	T: 108 B: 208	100 fewer	301 more
50-69	3 (Canada, Germany, Spain) [26, 27, 29]	3 (Canada, Germany, Spain) [26, 27, 29]	3 (Canada, Germany, Spain) [26, 27, 29]	Very low for all comparisons
	A: 904-16,300 B: 609-14,400	T: 2166-14,100 B: 2487-14,400	1300 fewer to 300 fewer	1,289 fewer to 5,000 fewer
70-74*	2 (Canada, Spain) [26, 27]	2 (Canada, Spain) [26, 27]	2 (Canada, Spain) [26, 27]	Very low for all comparisons
	A: 428-3200 B: 287-3500	T: 171-3200 B: 287-3500	300 fewer to 116 fewer	0 more to 237 more
Radiation induced BC				
45-49*	1 (US) [21]	-	-	Very low for all comparisons
	A: 32 B: 18	-	-	-



Table 4 continued

Age group	Annual vs. Biennial	Triennial vs. Biennial	Annual vs. Triennial	certainty of evidence across comparisons
	N° of studies, countries	N° of studies, countries	N° of studies, countries	
50–69	2 (Canada, US) [21, 23] <sup>†</sup>	13 more to 22 more	-	Very low for all comparisons
Death by radiation induced BC				
45–49 <sup>‡</sup>	1 (US) [21] A: 6 B: 4	2 more	-	Very low for all comparisons
50–69	2 (Canada, US) [21, 23] <sup>†</sup>	1 more to 3 more	-	Very low for all comparisons

To review the reference for each study and the reasons the certainty of the evidence was downgraded see Supplementary file Table S4 to S12. When more than one study informing an outcome, the number represents the range of point estimates reported across studies.

<sup>†</sup>Number of events was not directly reported for this age group. We made an ad hoc calculation subtracting the events from overlapping age groups (e.g. number of QALYs in women 45 to 69 years minus the estimates from 50 to 69 years).

<sup>‡</sup>The certainty of evidence departed from low as the input parameters that inform the modelling studies were of low to very low certainty.

<sup>§</sup>Only one study providing unpublished data informed this comparison. The result was in a different direction than the other bodies of evidence and thus cautious interpretation is recommended.

<sup>¶</sup>Unpublished data from one study (Mapiyo 2014) reported 19 fewer BC deaths averted with annual compared to biennial screening. This result was inconsistent with the other studies and, therefore, is not included in the table.

95%CI 0.80–1.21) and 60–69 (adjusted RR 0.99; 95%CI 0.79–1.24) with annual versus biennial screening [37]. Another study in the US found uncertain evidence that triennial screening compared to biennial screening in white women might be associated to lower odds of stage IIB–IV (OR 0.83; 95%CI 0.65–1.07) but it was not consistent with the observed difference in large tumour size (>20 mm) (OR 1.15; 95%CI 0.93–1.41), or presence of lymph nodes (OR 0.98; 95%CI 0.80–1.21) at BC diagnosis [38].

From a US study using mammography and tumour registries, the 10-year probability of a false positive result was 55.2% (95%CI 54.8–55.7%) with annual screening, 35.4% (95%CI 35.0–35.7%) with biennial screening and 24.8% (95%CI 24.5–25.2%) with triennial screening [38]. The cumulative 10-year probability of having a false positive biopsy recommendation was 9.7% (95%CI 9.3–10.1%) with annual screening, 5.4% (95%CI 5.2–5.6%) with biennial screening and 3.7% (95%CI 3.6–3.9%) with triennial screening [38]. These findings were consistent with the risk of false positive results observed in a retrospective cohort of a screening programme of New York [42].

**Modelling studies.** In a Canadian modelling study, the number of BC deaths averted per 100,000 women aged 50–69 screened annually, biennially or triennially compared to no screening was 740, 520 and 400, respectively [26]. In another study, including three models tailored to the US population, the number of BC deaths averted per 100,000 screened women aged 50–74, with scattered fibroglandular breast density, was 690, 520 and 400 for annual, biennial and triennial screening [22] and the number of QALYs gained was 6000, 4700 and 3600, respectively [22]. A microsimulation model for the German population found a median of 4400, 3900 and 3330 additional QALYs with annual, biennial and triennial screening [29].

The estimated overdiagnosis was greater with more frequent screening intervals. In women with scattered fibroglandular density aged 50–74, a microsimulation model study estimated 2900, 2000 and 1600 for annual, biennial and triennial screening compared to no screening per 100,000 women [22]. A similar trend was reported in a study using non-individual models for a Spanish cohort of women aged 50–69 [27].

A microsimulation model estimated the risk of radiation induced adverse events in 100,000 women aged 50–74 to be of 27 induced BC cases with biennial screening and 49 with annual screening [21]. The attributed number of radiation related deaths simulated was 4 with biennial screening and 7 with annual screening for the same age group [21]. A similar difference between biennial and annual screening intervals was observed from an excess absolute risk model of radiation induced BC [25].

#### Benefits and harms in women aged 70–74 (Tables 3/4)

**Observational studies.** Three studies provided estimations of advanced BC stage (IIB–IV) in older women, using population registries but for different age ranges (i.e. 66–89 [30], 70–85 [37] and 70–89 years [40]). In the age group of 70–85, the proportion of tumours at stage IIB–IV were no different among newly diagnosed BC with a history of biennial or annual screening (OR 0.98 95%CI 0.76–1.27) [37].

One study estimated that the 10-year cumulative probability of false positive results for women between the ages of 75 and 89 may be higher with annual screening (47%, 95%CI 44.9–49.5%) compared to biennial screening (26.6%, 95%CI 25.7–27.5%) [30]. The cumulative probability of false positive biopsy recommendations may also be higher for annual screening (9%, 95%CI 8–11%) compared to biennial screening (4%, 95%CI 4–5%) [30].

**Modelling studies.** The estimated difference for BC deaths between the different intervals might be small. A microsimulation model estimated the number of BC deaths averted for annual, biennial and triennial screening to be 100, 90 and 80, respectively, compared to no screening per 100,000 screened Canadian women

[26]. This result was consistent with the one reported in a non-individual model for a Spanish cohort which showed almost similar benefits for the three screening intervals (unpublished data) [27], and a small number of QALYs gained since life expectancy is lower in this age group.

Only one non-individual based model estimated overdiagnosis for this age group and it showed a small increasing trend with shorter screening intervals from 193 for triennial screening to 269 for annual screening [52].

#### Risk of bias and certainty of the evidence

Overall, the certainty of the evidence was very low, and therefore the differences observed between the possible combinations of screening intervals and age groups are uncertain. The exemption was the evidence from the only RCT included in this systematic review which was downgraded to moderate certainty due to imprecision [19].

The evidence from observational studies was limited among other factors by indirectness as for the age group of 45–49 we only identified studies including a broader age range from 40 to 49 years of age at the time of invitation to screening, and from some studies we had to extract results from specific subgroups of women (e.g. normal weight or white women). All secondary analysis from surveillance registries were also subject to misclassification bias of the interventions as the periodicity of screening was assigned based on different time ranges that elapsed between the two latest mammographies prior to diagnosis. Additionally, US studies used opportunistic screening, thus women might have anticipated or delayed the mammography due to preferences or indications given by radiologists.

We decided that for modelling studies, our GRADE assessment departed from low certainty after considering methodological limitations of key input evidence (i.e. mammography sensitivity estimated from BCSC registries including women from wider age groups than our clinical question and with a clinical follow-up restricted to only one year [53], or no formal assessment of risk of bias in the individual-patient-data meta-analysis used to inform treatment effectiveness [54]) and that credibility assessment of model development was limited due to suboptimal reporting. There was also limited reporting of formal sensitivity analysis to assess the impact of input data assumptions on the simulated events [21, 24, 25]. We had concerns about indirectness given that most models used observational data from the US to inform their input parameters (i.e. radiation induced BC), and because in one modelling study data was only available by different levels of breast density (i.e. scattered fibroglandular density) [22]. Finally, one study providing unpublished data (Vilapriyo 2014) [27] reported fewer BC deaths averted with annual compared to biennial or triennial screening in the age group of 45–49 years. This result was not internally consistent (i.e. annual screening had the largest number of BC deaths averted from 45 to 69) and differed from other studies or bodies of evidence; thus we included this result cautiously only if other studies were not available (Table 4).

The detailed risk of bias assessment per study is available under request. The evidence profiles for all age groups and intervals comparisons describing the reasons for downgrading the certainty of evidence are available from Supplementary Tables S5–S13. In the evidence profiles we prioritised the reporting of evidence from observational/randomised studies over modelling studies (i.e. false positive results).

#### DISCUSSION

##### Main findings

Our systematic review shows that in women of average breast cancer risk, screening intervals may have different trade-offs between benefits and harms for each age group. However, the

available evidence was mostly of very low certainty and precludes us from reaching firm conclusions. In women 50–69 years old, annual compared to biennial screening may have small additional benefits but an important increase in false positive results. Triennial compared to biennial screening suggests the latter provides more benefits but also some additional harms. In younger women (45–49), the more frequent screening intervals (going from biennial to annual screening) provides smaller incremental benefits (i.e. number of BC deaths averted), nearly similar incremental estimates of overdiagnosis and slightly more incremental harms (i.e. false positive results and false positive biopsies recommendations from observational studies) than in women 50–69 years of age. Thus the overall balance between benefits and harms is more favourable in the latter age group. Finally, among women aged 70–74, the smaller incremental harms and similar benefits with shorter screening intervals suggests that longer intervals probably have a more favourable overall balance, but the difference may be small.

We observed sparse data, especially in older women and for critical outcomes, such as BC mortality or disease stage at diagnosis. The only included RCT showed that annual screening, compared to triennial screening, probably reduces BC mortality in women 50–62 years of age. Observational evidence consisted of population registries from different time periods with high uncertainty. We considered modelling evidence when empirical evidence was not available. However, its certainty was very low due to indirectness, since data for input parameters mostly come from opportunistic screening settings. Model studies suggested that in women aged 50–69 the benefits with annual screening may be a bit larger but may also be associated to relevant harms, including the possibility of a small increase of new BC lesions induced by radiation exposure; thus, biennial screening may provide a more favourable balance, while in other age groups the potential benefits gains with more frequent screening intervals may be smaller.

##### Our results in the context of previous research

Our results are broadly consistent but more comprehensive than previous reviews. The USPSTF based their assessment on one modelling study (included in our review), concluding that when moving from biennial to annual mammography, regardless of the starting age, there is a small increase in averted deaths but with a large increase of harms [7]. A systematic review conducted by the American Cancer Society included an indirect comparison between RCTs and a model study from the CISNET collaboration, concluded that beginning screening with more frequent intervals likely results in a greater mortality reduction but the magnitude is uncertain [55].

The modelling estimates of harms due to overdiagnosis remains a matter of debate as there is no consensus on the methods to quantify this outcome [56], and many assumptions are made, including the clinical impact of DCIS and the probability of some cancers to spontaneously regress [50]. It is worth noting that there is also considerable uncertainty in the evidence coming from RCTs. For example, a review including only studies that did not invite women of the control group to screening at the end of the trial period, reported a relevant proportion of overdiagnosis [57]. However, the UK age trial showed that the cumulative incidence of invasive cancers was similar, if not higher, in women who underwent only one mammogram after the age of 50 compared to women who underwent annual mammography from 40 to 49, and then entered a triennial screening programme [58].

The cost-effectiveness of implementing different screening intervals has been studied in few microsimulation models. One study assessed the impact of extending the Dutch screening programme in women under 50, showing that biennial strategies were cost-effective while other alternatives, such as annual



screening starting at 45, resulted in less favourable incremental cost-effectiveness ratios (ICERs) [59]. However, the study used an 80% adherence to screening [59], which might have influenced the relative trade-offs between different screening intervals, as previously described [22]. In women from the US between 50 and 74 years of age, with different breast densities and individual risk level of developing BC, triennial strategies were considered cost-effective (at a threshold of \$100 000 per QALY) for subgroups with average risk and low breast density, while biennial strategies were cost-effective for other breast density subgroups at an average or intermediate risk [22].

#### Limitations and strengths

Although we included only English language articles, the risk of selection bias is probably small as we also screened previous systematic reviews and consulted the GDG experts, not identifying additional studies. Some results are not directly transferable to the European context; for example the cumulative 10-year false positive rates from US studies are higher than those reported in organised European screening programmes. However, we assumed that the difference between intervals would be more comparable across different settings. The scarce available empirical evidence to evaluate the trade-offs between benefits and harms limited our conclusions. We therefore included modelling evidence to complement the gaps in the evidence, an approach that is recommended for interventions such as population screening [60].

#### Implications for practice and research

Our findings may have different implications for practice depending on the age group, the balance between benefits and harms, available resources for public health services, and how women value the different outcomes. In the case of women invited to an opportunistic screening programme (or considering screening) a shared decision-making process to carefully explain the pros and cons of each decision is warranted. Similarly, given the low certainty of evidence and the variability and uncertainty of how women value outcomes at stake, guideline panels are likely to formulate conditional recommendations, as opposed to strong ones. The scope of this review is determined by the European Breast Guidelines screening recommendations; [10] thus, policy makers should note that we did not include modelling estimates for women between the ages of 40 and 44 as screening is not suggested in this age group [10]. Also, readers should be careful when interpreting the effects of screening intervals across the different age groups, as comparisons are limited by the small number of screening rounds in the 45 to 49 and 70 to 74 age groups, compared to the 50–69 age group.

Recommendations about mammography screening intervals will also depend on the magnitude and relative importance of potential harms. Narayan, et al. assessed to what extent harms should decrease in order to make a screening interval with an unfavourable balance of benefits and harms acceptable [61]. They found that for annual screening a reduction of 31% false positive results would be required to support a recommendation in favour of starting at 50, although this was in the context of false positive rates prevailing in the US [61]. Policy makers should probably consider implementing interventions to improve mammography performance, mitigating concerns about potential harms. For example previous studies suggest that comparing mammograms with prior exams can significantly reduce the recall rate while maintaining the same detection rate [62, 63].

Several research priorities were identified during this review, with feedback from the GDG experts, such as need for: (i) empirical research on the effectiveness of the different screening intervals due to the current very low certainty of evidence; (ii) cost-effectiveness studies using unitary costs from different settings, and in particular for women aged 45 to 49; (iii) assessment of alternative imaging modalities; (iv) tailored screening according to risk vs population

screening. For example, previous research has highlighted that breast density influences both mammography accuracy and risk of developing breast cancer [64, 65]. For further information on the complete recommendations formulated in the European Guidelines on Breast Cancer Screening and Diagnosis, please visit the ECIBC website (<https://healthcare-quality.jc.ec.europa.eu/european-breast-cancer-guidelines/screening-ages-and-frequencies>).

#### DISCLAIMER

All views expressed in this article are strictly those of the authors.

#### DATA AVAILABILITY

All data sources used during this study are described in this published article and its additional information files. The datasets analysed are available from the corresponding author on reasonable request.

#### REFERENCES

- Forlay J, Brvik M, Lam F, Colombet M, Mery L, Piñeros M et al. Global Cancer Observatory: Cancer Today. Lyon, France: International Agency for Research on Cancer. Available from: <http://gco.iarc.fr/today>, accessed [25-07-2019]. 2018.
- ECIS. European Cancer Information System From: <https://ecis.jc.ec.europa.eu>, accessed on 20-July-2019. 2018.
- Duffy SW, Tabar L, Yen AM, Dean PB, Smith RA, Jonsson H, et al. Mammography screening reduces rates of advanced and fatal breast cancers: Results in 549,091 women. *Cancer*. 2020;126:2971–9.
- Bonelli C, Cohen S, Duncan A, Given-Wilson R, Jenkins J, Kearns O, et al. NHS Breast Screening Programme. Clinical guidance for breast cancer screening assessment. Fourth edition Nov 2016.
- Siu AL, Force U, S. P. S. T. Screening for breast cancer: U.S. Preventive services task force recommendation statement. *Ann Intern Med*. 2016;164:279–96.
- Oeffinger KC, Farnham ET, Eklert R, Hestig A, Michaelson JS, Shih YC, et al. Breast cancer screening for women at average risk: 2015 guideline update from the American Cancer Society. *JAMA*. 2015;314:1599–614.
- Mandelblatt JS, Stout NK, Schechter CB, van den Broek JJ, Miglioretti DL, Krapcho M, et al. Collaborative modeling of the benefits and harms associated with different U.S. Breast Cancer Screening Strategies. *Ann Intern Med*. 2016;164:215–25.
- Krlikowski K, Zhu W, Hubbard RA, Geller B, Dittus K, Baithwaite D, et al. Outcomes of screening mammography by frequency, breast density, and postmenopausal hormone therapy. *JAMA Intern Med*. 2013;173:807–16.
- Schunemann HJ, Landa D, Dimitrova N, Alonso-Codillo P, Grawinkel A, Quinn C, et al. Methods for development of the European Commission Initiative on Breast Cancer Guidelines: recommendations in the era of guideline transparency. *Ann Intern Med*. 2019;171:273–80.
- Schunemann HJ, Landa D, Quinn C, Follmann M, Alonso-Codillo P, Rios PG, et al. Breast cancer screening and diagnosis: a synopsis of the European Breast Guidelines. *Ann Intern Med*. 2019;172:46–56.
- 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.
- Higgins JP, Altman DG, Gotzsche PC, Juni P, Moher D, Oxman AD, et al. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *BMJ*. 2011;343:d5928.
- Sterne JA, Hernan MA, Reeves BC, Savovic J, Berkman ND, Viswanathan M, et al. ROBINS-2: a tool for assessing risk of bias in non-randomised studies of interventions. *BMJ*. 2019;365:g4019.
- Jaime Caro J, Eddy DM, Kan H, Kaitz C, Patel B, Eldessouki R, et al. Questionnaire to assess relevance and credibility of modelling studies for informing health care decision making: an ISPOR-AMCP-NPC Good Practice Task Force report. *Value Health*. 2014;17:174–82.
- Guyatt GH, Oxman AD, Schunemann HJ, Tugwell P, Knottnerus A. GRADE guidelines: a new series of articles in the Journal of Clinical Epidemiology. *J Clin Epidemiol*. 2011;64:380–2.
- Guyatt GH, Oxman AD, Vist GE, Kunz R, Rakic-Ytter Y, Alonso-Codillo P, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ*. 2008;336:924–6.
- Hultcrantz M, Rind D, Akl EA, Triawek S, Mustafa RA, Iorio A, et al. The GRADE Working Group clarifies the construct of certainty of evidence. *J Clin Epidemiol*. 2017;87:4–13.
- Duffy SW, B. R. Long-term mortality results from the UK screening frequency trial. *EJC Supplements*. 2008;648.



19. Breast Screening Frequency Trial. G. The frequency of breast cancer screening: results from the UKCCCR Randomised Trial. United Kingdom Co-ordinating Committee on Cancer Research. *Eur J Cancer*. 2002;38:1458–64.
20. Gunsoy NB, Garcia-Closas M, Moss SM. Estimating breast cancer mortality reduction and overdiagnosis due to screening for different strategies in the United Kingdom. *Br J Cancer*. 2014;110:2412–9.
21. Miglioretti DL, Lange J, van den Broek JJ, Lee CI, van Ravesteyn NT, Rittley D, et al. Radiation-induced breast cancer incidence and mortality from digital mammography screening: a modeling study. *Ann Intern Med*. 2016;164:205–14.
22. Trentham-Dietz A, Kefauver K, Stout NK, Miglioretti DL, Schechter CB, Eggen MA, et al. Tailoring Breast Cancer Screening Intervals by Breast Density and Risk for Women Aged 50 Years or Older: Collaborative Modeling of Screening Outcomes. *Ann Intern Med*. 2016;165:700–12.
23. Tsunematsu M, Kakihashi M. An analysis of mass screening strategies using a mathematical model: comparison of breast cancer screening in Japan and the United States. *J Epidemiol*. 2015;25:162–71.
24. van Ravesteyn NT, Miglioretti DL, Stout NK, Lee SJ, Schechter CB, Bult DS, et al. Tipping the balance of benefits and harms to favor screening mammography starting at age 40 years: a comparative modeling study of risk. *Ann Intern Med*. 2012;156:609–17.
25. Yaffe MJ, Minkin JC. Risk of radiation-induced breast cancer from mammographic screening. *Radiology*. 2011;258:96–105.
26. Yaffe MJ, Mittmann N, Lee P, Tosteson AN, Trentham-Dietz A, Alagoz O, et al. Clinical outcomes of modeling mammography screening strategies. *Health Rep*. 2015;26:9–15.
27. Vilapinyo E, Forns C, Carles M, Sala M, Pla R, Castells X, et al. Cost-effectiveness and harm-benefit analysis of risk-based screening strategies for breast cancer. *PLoS One*. 2014;9:e8658.
28. Mittmann N, Stout NK, Tosteson AN, Trentham-Dietz A, Alagoz O, Yaffe MJ. Cost-effectiveness of mammography from a publicly funded health care system perspective. *CMAJ Open*. 2018;6:E77–E86.
29. Amdt M, Pfäfer K, Quante AS. Is risk-stratified breast cancer screening economically efficient in Germany? *PLoS ONE*. 2019;14:e0217213.
30. Braithwaite D, Zhu W, Hubbard RA, O'Meara ES, Miglioretti DL, Geller B, et al. Screening outcomes in older US women undergoing multiple mammograms in community practice does interval, age, or comorbidity score affect tumor characteristics or false positive rates? *J Natl Cancer Inst*. 2013;105:334–41.
31. Coldman AJ, Phillips N, Olivotto IA, Gordon P, Warren L, Kan L. Impact of changing from annual to biennial mammographic screening on breast cancer outcomes in women aged 50–79 in British Columbia. *J Med Screen*. 2008;15:182–7.
32. Dittus K, Geller B, Weaver DL, Kefauver K, Zhu W, Hubbard R, et al. Impact of mammography screening interval on breast cancer diagnosis by menopausal status and BMI. *J Gen Intern Med*. 2013;28:1454–62.
33. Goel A, Litenberg B, Barad RC. The association between the pre-diagnosis mammography screening interval and advanced breast cancer. *Breast Cancer Res Treat*. 2007;102:339–45.
34. Hubbard RA, Kefauver K, Flowers C, Yankaskas BC, Zhu W, Miglioretti DL. Cumulative probability of false-positive recall or biopsy recommendation after 10 years of screening mammography: a cohort study. *Ann Intern Med*. 2011;155:481–92.
35. Hunt KA, Rosen EL, Sickles EA. Outcome analysis for women undergoing annual versus biennial screening mammography: a review of 24,211 examinations. *AJR Am J Roentgenol*. 1999;173:285–9.
36. Klemi PJ, Toikkanen S, Räsänen O, Parvinen I, Jousiainen H. Mammography screening interval and the frequency of interval cancers in a population-based screening. *Br J Cancer*. 1997;75:762–6.
37. Miglioretti DL, Zhu W, Kefauver K, Sprague BL, Ortega T, Bult DS, et al. Breast tumor prognostic characteristics and biennial vs annual mammography, age, and menopausal status. *JAMA Oncol*. 2015;1:1069–77.
38. O'Meara ES, Zhu W, Hubbard RA, Braithwaite D, Kefauver K, Dittus K, et al. Mammographic screening interval in relation to tumor characteristics and false-positive risk by race/ethnicity and age. *Cancer*. 2013;119:3959–67.
39. Parvinen I, Chiu S, Rytönen L, Klemi P, Immonen-Raiha P, Kauhava I, et al. Effects of annual vs triennial mammography interval on breast cancer incidence and mortality in ages 40–49 in Finland. *Br J Cancer*. 2011;105:1388–91.
40. White E, Miglioretti DL, Yankaskas BC, Geller BM, Rosenberg RD, Kefauver K, et al. Biennial versus annual mammography and the risk of late-stage breast cancer. *J Natl Cancer Inst*. 2004;96:1832–9.
41. Sanderson M, Levine RS, Fadden MK, Kiboume B, Rieu M, Cain V, et al. Mammography screening among the elderly: a research challenge. *Am J Med*. 2015;128:1362 e1367–1374.
42. McGuinness JE, Unger W, Trivedi MS, Yi HS, David R, Vanegas A, et al. Factors associated with false positive results on screening mammography in a population of predominantly Hispanic women. *Cancer Epidemiol Biomarkers Prev*. 2018;27:446–53.
43. Hubbard RA, Miglioretti DL, Smith RA. Modeling the cumulative risk of a false-positive screening test. *Stat Methods Med Res*. 2010;19:429–40.
44. Lee SJ, Li X, Huang H, Zelen M. The Dana-Farber CISNET model for breast cancer screening strategies: an update. *Med Decis Making*. 2018;38:445–535.
45. van den Broek JJ, van Ravesteyn NT, Huisdijk EA, de Koning HJ. Simulating the impact of risk-based screening and treatment on breast cancer outcomes with MISCAN-Fadia. *Med Decis Making*. 2018;38:445–535.
46. Schechter CB, Niaz AM, Jayakumar J, Chandler Y, Mandelblatt JS. Structure, function, and applications of the Georgetown-Breast (GB) Breast Cancer Simulation Model. *Med Decis Making*. 2018;38:665–775.
47. Huang X, Li Y, Song J, Berry DA. A Bayesian Simulation Model for Breast Cancer Screening, Incidence, Treatment, and Mortality. *Med Decis Making*. 2018 Apr;38(1\_suppl):785–885.
48. Plevritis SK, Sigal BM, Salzman P, Rosenberg J, Gynn P. A stochastic simulation model of U.S. breast cancer mortality trends from 1975 to 2000. *J Natl Cancer Inst Monogr*. 2006;86:95. <https://doi.org/10.1093/jncimonographs/kg102>.
49. Alagoz O, Eggen MA, Cevik M, Sprague BL, Fryback DG, Gengnon RE, et al. The University of Wisconsin breast cancer epidemiology simulation Model: an update. *Med Decis Making*. 2018;38:695–1115.
50. Mandelblatt JS, Niaz AM, Miglioretti DL, Munoz D, Sprague BL, Trentham-Dietz A, et al. Common model inputs used in CISNET collaborative breast cancer modeling. *Med Decis Making*. 2018;38:95–235.
51. van den Broek JJ, van Ravesteyn NT, Mandelblatt JS, Cevik M, Schechter CB, Lee SJ, et al. Comparing CISNET breast cancer models using the maximum clinical incidence reduction methodology. *Med Decis Making*. 2018;38:125–125S.
52. Carles M, Vilapinyo E, Cots F, Gregori A, Pla R, Roman R, et al. Cost-effectiveness of early detection of breast cancer in Catalonia (Spain). *BMC Cancer*. 2011;11:192.
53. Lehman CD, Arao RF, Sprague BL, Lee JM, Bult DS, Kefauver K, et al. National performance benchmarks for modern screening digital mammography: update from the breast cancer surveillance consortium. *Radiology*. 2017;283:40–58.
54. Early Breast Cancer Trialists' Collaborative G, Peto R, Davies C, Godwin J, Gray R, Pan HC, et al. Comparisons between different polychemotherapy regimens for early breast cancer: meta-analyses of long-term outcome among 100,000 women in 123 randomised trials. *Lancet*. 2012;379:432–44.
55. Myers ER, Moorman P, Geirisch JM, Hurlbush LJ, Grimm LJ, Chate S, et al. Benefits and harms of breast cancer screening: a systematic review. *JAMA*. 2015;314:1615–34.
56. Ripping TM, Ten Haaf K, Verbeek ALM, van Ravesteyn NT & Broeders MJM. Quantifying overdiagnosis in cancer screening: a systematic review to evaluate the methodology. *J Natl Cancer Inst*. 2017;109:dx060.
57. Canelo-Aybar C, Ferreira DS, Ballesteros M, Posso M, Montero N, Sola I et al. Benefits and harms of breast cancer mammography screening for women at average risk of breast cancer: a systematic review for the European Commission Initiative on Breast Cancer. *J Med Screen*. 2021; <https://doi.org/10.1177/0969141321998666>.
58. Moss SM, Wake C, Smith R, Evans A, Cudde H, Duffy SW. Effect of mammographic screening from age 40 years on breast cancer mortality in the UK Age trial at 17 years' follow-up: a randomised controlled trial. *Lancet Oncol*. 2015;16:123–32.
59. Sankaling VD, Huisdijk EA, van Lijst PA, van Ravesteyn NT, Racheboud J, de Koning HJ. Cost-effectiveness of digital mammography screening before the age of 50 in The Netherlands. *Int J Cancer*. 2015;137:1990–9.
60. Habbema JD, Wilt TJ, Etzioni R, Nelson HD, Schechter CB, Lawrence WF, et al. Models in the development of clinical practice guidelines. *Ann Intern Med*. 2014;161:812–8.
61. Narayan AK, Elkin EB, Lehman CD, Morris EA. Quantifying performance thresholds for recommending screening mammography: a revealed preference analysis of USPSTF guidelines. *Breast Cancer Res Treatment*. 2018;172:463–8.
62. Hayward JH, Ray KM, Wisner DJ, Kornik J, Lin W, Joe BN, et al. Improving screening mammography outcomes through comparison with multiple prior mammograms. *AJR Am J Roentgenol*. 2016;207:918–24.
63. Rooklofs AA, Kassemaier N, Wedelind N, Beck C, van Woudenberg S, Snoeren PR, et al. Importance of comparison of current and prior mammograms in breast cancer screening. *Radiology*. 2007;242:70–77.
64. McCormack VA, dos Santos Silva I. Breast density and parenchymal patterns as markers of breast cancer risk: a meta-analysis. *Cancer Epidemiol Biomarkers Prev*. 2006;15:159–69.
65. Boyd NF, Guo H, Martin LJ, Sun L, Stone J, Riddell E, et al. Mammographic density and the risk and detection of breast cancer. *N Engl J Med*. 2007;356:227–36.

#### ACKNOWLEDGEMENTS

It is with genuine gratitude and warm regard that we would like to dedicate this publication to our dearest colleague Sue Warner who passed away on 21–06–2021. The authors would like to sincerely thank all members of the Guidelines Development Group of the European Commission Initiative on Breast Cancer for their participation in the

discussions generated by this systematic review which led to the different recommendations they developed in the European Guidelines on Breast Cancer Screening and diagnosis (<https://healthcare-quality.jrc.europa.eu/european-breast-cancer-guidelines>).

#### AUTHOR CONTRIBUTIONS

Carlos Canelo-Aybar, Margarita Posso, Nadia Montoro, Nan Sol and Pablo Alonso-Coello were responsible for conducting the systematic review. Carlos Canelo-Aybar and Margarita Posso conducted the search, data extraction, and analysis. Zulika Saz-Parkinson, Paolo Giorgi Rossi, Stephen W. Duffy, Markus Follmann, Stephen W. Duffy, Markus Follmann, Axel Grüwlinghoff and Paolo Giorgi Rossi contributed to the definition of the research protocol, and provided comments to the preliminary results of the systematic review. Carlos Canelo-Aybar drafted the first version of the article. All authors contributed to the interpretation and reporting of the results and provided comments on subsequent versions of the article. All authors read and approved the final manuscript prior submission.

#### FUNDING INFORMATION

The systematic review was carried out by Iberoamerican Cochrane Collaboration under the Framework contract 443094 for procurement of services between the European Commission's Joint Research Centre and Asociación Colaboración Cochrane Iberoamericana.

#### ETHICS APPROVAL AND CONSENT TO PARTICIPATE

Not applicable.

#### CONSENT TO PUBLISH

Not applicable.

#### COMPETING INTERESTS

Zulika Saz Parkinson was an employee of the Joint Research Centre, European Commission when this systematic review was carried out. Carlos Canelo-Aybar,

Margarita Posso, Ivan Solís and Pablo Alonso-Coello are employees of the Iberoamerican Cochrane Collaboration. Stephen W. Duffy, Markus Follmann, Axel Grüwlinghoff and Paolo Giorgi Rossi are GDG members of the EOCB guidelines.

#### ADDITIONAL INFORMATION

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41416-021-01521-8>.

**Correspondence** and requests for materials should be addressed to Carlos Canelo-Aybar or Zulika Saz Parkinson.

**Reprints and permission information** is available at <http://www.nature.com/reprints>.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© European Union 2021

## **6.2 Second study: Recommendations from the European Commission Initiative on Breast Cancer for multigene testing to guide the use of adjuvant chemotherapy in patients with early breast cancer, hormone receptor positive, HER-2 negative<sup>96</sup>**

### **Systematic review of effectiveness**

We included two RCTs,<sup>97, 98</sup> two secondary analyses from former clinical trials<sup>99, 100</sup> and one pooled analysis of observational studies.<sup>101</sup>

### **21 gene recurrence score**

-Treatment interaction design: One RCTs compared adding chemotherapy to endocrine therapy vs endocrine therapy alone showing a different effect across 21-RS recurrence groups with a hazard ratio (HR) of 1.31 (95% CI 0.46–3.78), 0.61 (95% CI 0.24–1.59) and 0.26 (95% CI 0.13–0.53) in low, intermediate and high-risk groups respectively.<sup>100</sup> Another RCT included stored tumour specimens, reporting adjusted by number of positive nodes no benefit for chemotherapy on disease free survival (DFS) in the low genomic risk group (HR = 1.02; 95% CI 0.54–1.93) and a potential advantage in the high genomic risk (HR = 0.59, 95% CI 0.35–1.01).<sup>99</sup>

-Marker-based strategy: One RCT allocated women with intermediate genomic risk (11 to 26 risk score) to either endocrine therapy alone or chemotherapy plus endocrine therapy. They result suggested little to no difference in the risk of recurrence with chemotherapy plus endocrine therapy for invasive DFS (HR 1.14; 95% CI 0.99–1.31).<sup>98</sup>

### **70-GS**

-Treatment interaction design: One pooled database analysis suggested patients with a low and high genomic risk who received chemotherapy may have a different risk of recurrence compared to endocrine therapy alone (HR 0.26; 95% CI 0.03–2.02 and HR 0.35; 95% CI 0.17–0.71, respectively).<sup>101</sup>

-Marker-based strategy One RCT allocated patients with clinical/genomic discordant-risk groups to receive either chemotherapy in addition to endocrine therapy or endocrine therapy alone. Women with high clinical risk and low genomic risk may have an increase of DFS (HR 0.64; 95% CI 0.43–0.95), and on distant metastases free survival (HR 0.65; 95% CI 0.38–1.10) and of overall survival (OS) (HR 0.63; 95% CI 0.29–1.37). The group of low clinical risk and high genomic risk showed more imprecise effects and lower uncertainty for distant metastases free survival and OS.<sup>97</sup>

### **Decision tree model**

Depending on the different treatment scenarios (all women or only women with high clinical risk are treated with chemotherapy) and genetic testing strategies (all women or only women with high clinical risk are tested), the number of chemotherapies avoided by using the 21-RS would change from more than 600 to about 200 per 1000 women tested. Survival outcomes did not change substantially but may prevent 37 distant metastases compared to a scenario in which only women with high clinical risk were treated with chemotherapy depending on the assumption and inputs used. For the 70-GS, the only scenario considered was one in which only high-risk women would receive chemotherapy leading result to avoidance of about 230 chemotherapies per 1000 women but with small increase of recurrences.

### **Systematic review of economic evidence**

We did not identify economic evaluations models applicable to the clinical question of interest. Therefore, we considered the benefits and harms estimated using our ad-hoc model described above, and unitary costs reported by the studies included in the literature review.

### **Certainty of evidence**

The overall certainty of the evidence of effects was rated as low to very low due to indirectness and risk of bias. Economic evidence was not formally assessed as did not comply with the relevant assumption made by the guideline panel.





## CONSENSUS STATEMENT

# Recommendations from the European Commission Initiative on Breast Cancer for multigene testing to guide the use of adjuvant chemotherapy in patients with early breast cancer, hormone receptor positive, HER-2 negative

Paolo Giorgi Rossi<sup>1</sup>, Annette Lebeau<sup>2</sup>, Carlos Canelo-Aybar<sup>3,4</sup>, Zuleika Saz-Parkinson<sup>5,6</sup>, Cecily Quinn<sup>7</sup>, Miranda Langendam<sup>8</sup>, Helen McGarrigle<sup>9</sup>, Sue Warman<sup>10</sup>, David Rigau<sup>3</sup>, Pablo Alonso-Coello<sup>3</sup>, Mireille Broeders<sup>11,12</sup>, Axel Graewinkel<sup>13</sup>, Margarita Posso<sup>14,15</sup>, Stephen Duffy<sup>16</sup>, Holger J. Schünemann<sup>17</sup> and the ECIBC Contributor Group

**BACKGROUND:** Predicting the risk of recurrence and response to chemotherapy in women with early breast cancer is crucial to optimise adjuvant treatment. Despite the common practice of using multigene tests to predict recurrence, existing recommendations are inconsistent. Our aim was to formulate healthcare recommendations for the question “Should multigene tests be used in women who have early invasive breast cancer, hormone receptor-positive, HER2-negative, to guide the use of adjuvant chemotherapy?”

**METHODS:** The European Commission Initiative on Breast Cancer (ECIBC) Guidelines Development Group (GDG), a multidisciplinary guideline panel including experts and three patients, developed recommendations informed by systematic reviews of the evidence. Grading of Recommendations Assessment, Development and Evaluation (GRADE) Evidence to Decision frameworks were used. Four multigene tests were evaluated: the 21-gene recurrence score (21-RS), the 70-gene signature (70-GS), the PAM50 risk of recurrence score (PAM50-RORS), and the 12-gene molecular score (12-MS).

**RESULTS:** Five studies (2 marker-based design RCTs, two treatment interaction design RCTs and 1 pooled individual data analysis from observational studies) were included; no eligible studies on PAM50-RORS or 12-MS were identified and the GDG did not formulate recommendations for these tests.

**CONCLUSIONS:** The ECIBC GDG suggests the use of the 21-RS for lymph node-negative women (conditional recommendation, very low certainty of evidence), recognising that benefits are probably larger in women at high risk of recurrence based on clinical characteristics. The ECIBC GDG suggests the use of the 70-GS for women at high clinical risk (conditional recommendation, low certainty of evidence), and recommends not using 70-GS in women at low clinical risk (strong recommendation, low certainty of evidence).

*British Journal of Cancer* (2021) 124:1503–1512; <https://doi.org/10.1038/s41416-020-01247-z>

## BACKGROUND

Breast cancer is the most frequently diagnosed cancer among women.<sup>1</sup> In the European Union, including UK, 404,920 women were diagnosed with breast cancer and 98,755 died because of this disease in 2018.<sup>2</sup> Hormone receptor (HoR)-positive (i.e. oestrogen receptor (ER)- and/or progesterone receptor

(PR)-positive), human epidermal growth factor receptor 2 (HER2)-negative breast cancer represents about 70% of breast cancer diagnosed in western countries.<sup>3</sup> At the time of diagnosis, around 60% of this type of cancer has not spread to lymph nodes,<sup>4</sup> and approximately 15% of these women will develop a recurrence within 10 years if treated with adjuvant endocrine

<sup>1</sup>Azienda Sanitaria Locale—IRCCS di Reggio Emilia, Reggio Emilia, Italy; <sup>2</sup>Department of Pathology, University Medical Center Hamburg-Eppendorf, Hamburg, Germany; <sup>3</sup>Iberoamerican Cochrane Center, Biomedical Research Institute (IB-Sant Pau-CIBERS), Barcelona, Spain; <sup>4</sup>Department of Paediatrics, Obstetrics and Gynaecology, Preventive Medicine, and Public Health, PhD Programme in Methodology of Biomedical Research and Public Health, Universitat Autònoma de Barcelona, Bellaterra, Spain; <sup>5</sup>European Commission, Joint Research Centre (JRC), Ispra, Italy; <sup>6</sup>Instituto de Salud Carlos III, Health Technology Assessment Agency, Avenida Monforte de Lemos 5, Madrid, Spain; <sup>7</sup>St. Vincent's University Hospital, Dublin, Ireland; <sup>8</sup>Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Amsterdam UMC, University of Amsterdam, Amsterdam Public Health Institute, Amsterdam, The Netherlands; <sup>9</sup>Cardiff and Vale UHB - General Surgery, Cardiff, UK; <sup>10</sup>Havart Lodge, Havart Road, Langford, North Somerset, UK; <sup>11</sup>Department for Health Evidence, Radboud University Medical Center, Nijmegen, the Netherlands; <sup>12</sup>Dutch Expert Centre for Screening, Nijmegen, the Netherlands; <sup>13</sup>Radiologie am Theater, Paderborn, NRW, Germany; <sup>14</sup>Department of Epidemiology and Evaluation, IMIM (Hospital del Mar Medical Research Institute), Barcelona, Spain; <sup>15</sup>Research Network on Health Services in Chronic Diseases (REDISSEC), Barcelona, Spain; <sup>16</sup>Centre for Cancer Prevention, Queen Mary University of London, Charterhouse Square, London, UK and <sup>17</sup>Michael G. DeGroote Cochrane Canada and McGRADE Centres, Department of Health Research Methods, Evidence and Impact, McMaster University Health Sciences Centre, Hamilton, Ontario, Canada

Correspondence: Zuleika Saz-Parkinson ([zuleika.saz-parkinson@ec.europa.eu](mailto:zuleika.saz-parkinson@ec.europa.eu))

Members of the ECIBC Contributor Group are listed above Acknowledgements.

These authors contributed equally: Paolo Giorgi Rossi, Annette Lebeau

Received: 1 October 2020 Revised: 10 December 2020 Accepted: 17 December 2020

Published online: 18 February 2021



Table 1. Structured clinical questions.

Population	Intervention	Comparison	Outcomes
Patients with hormone receptor-positive, HER2-negative, lymph node-negative or up to 3 lymph nodes-positive invasive breast cancer.	70-gene signature to decide chemotherapy: women with low clinical risk do not receive chemotherapy independently from 70-gene signature; women with high clinical risk will receive chemotherapy only if at high genomic risk.  21-gene recurrence score to decide chemotherapy, in two alternative scenarios: • Women with low or intermediate genomic risk will not receive chemotherapy; women with high genomic risk will; • Only women with high clinical risk will be tested for genomic risk and receive chemotherapy if genomic risk is high; women with low clinical risk will not receive chemotherapy.	• According to the considered clinical trial designs the comparison would be either: • All patients receive chemotherapy • All patients with low clinical risk do not receive chemotherapy and those with high clinical risk subgroup receive chemotherapy. Direct comparisons between the different multigene tests were not performed.	• Overall survival • Disease-free survival • Adverse effects of drugs • Quality of life.

Initially PICO covered four multigene tests, but two of them, the PAM50 risk of recurrence score (RORS) and 12-gene molecular score (12-MS), were excluded because of study selection criteria. These selection criteria allowed no direct comparison between the different multigene tests.

therapy alone.<sup>4,5</sup> The risk of recurrence could be reduced by the addition of chemotherapy.<sup>7</sup> However, given the relatively low risk of recurrence and the partial effectiveness of chemotherapy in these women, most would be over-treated if all received chemotherapy. The same rationale would apply to women with HoR-positive, HER2-negative invasive breast cancer with 1–3 positive lymph nodes.<sup>6</sup> Several prognostic factors, including clinical-pathological features such as age, tumour size, percentage of ER- and PR-positive cells as well as Ki67-index,<sup>7–9</sup> predict the risk of recurrence and can help identify women who would benefit the most from chemotherapy. Although these factors have been shown to discriminate different prognostic groups, they showed no or minimal predictive value on the response to chemotherapy.<sup>10</sup>

In the last 15 years, different tests have been developed to stratify patients with early breast cancer into different risk of recurrence groups by analysing the activity of various genes. Although these multigene tests use diverse techniques (RT-PCR, microarray, and others) and diverse target gene combinations, they all focus on genes involved in cell proliferation. The tests provide recurrence risk profiles categorised in different ways. Some tests have been explicitly proposed to provide additional information to clinical-pathological features, as the 12-MS and the PAM50-RORS. Based on the results of the MINDACT trial,<sup>6</sup> the application of the 70-GS also takes into account clinical prognostic characteristics, while the 21-RS has been proposed to substitute clinical risk-based treatment decisions.

The European Commission Initiative on Breast Cancer (ECIBC) aims to provide evidence-based recommendations for screening and diagnosis of breast cancer.<sup>11</sup> The Guidelines Development Group (GDG) of the ECIBC prioritised a clinical question on the use of multigene tests to guide the use of adjuvant chemotherapy in HoR-positive, HER2-negative and lymph node-negative or up to 3 lymph nodes-positive invasive breast cancer. Four multigene tests, used to stratify women with breast cancer into different groups according to recurrence risk,<sup>12–17</sup> are included in the clinical question (Supplementary Table 1): 21-RS (Oncotype DX, Genomic Health Inc), 12-MS (EndoPredict, Myriad Genetics Inc), PAM50-RORS (Prosigna test, NanoString Technologies Inc), and 70-GS (MammaPrint, Agendia Inc). Direct comparison between different tests is beyond the scope of these recommendations.

## METHODS

### Structured question and outcome prioritisation

The clinical question “Should multigene tests be used in patients who have HoR-positive, HER2-negative, lymph node-negative or up to 3 lymph nodes-positive invasive breast cancer to guide the use of

adjuvant chemotherapy” was structured following the Population, Intervention, Comparison and Outcomes (PICO) format (Table 1). The outcomes were also prioritised by the GDG using a nine-point scale (7 to 9 critical; 4 to 6 important; 1 to 3 of limited importance), as suggested by the Grading of Recommendations Assessment, Development and Evaluation (GRADE) approach.<sup>11,18</sup> The GDG decided not to attempt any head-to-head comparisons between the different tests.

### Systematic review

**Data sources and searches.** MEDLINE (May 2018), EMBASE (May 2018) and CENTRAL (May 2018) databases were searched, using pre-defined algorithms, for both systematic reviews and individual studies (Supplementary Table 2); this original search was continuously run up to October 2018. Lists of references of the included studies were reviewed and members of the GDG were requested to provide additional studies.

**Study selection.** Randomised controlled trials (RCT) and cohort studies (including pooled analyses of studies), either from prospective or retrospective analysis, of stored specimen samples were included as long as they applied any of the four tests as predictive markers for guiding the use of adjuvant chemotherapy (Supplementary Fig. 1).

A predictive marker identifies the differential benefit of a treatment based on the marker status. Thus, we included the following assessment approaches: (a) *Marker-based strategy design*: patients are assigned to a treatment arm depending on whether they received treatment (i.e. endocrine therapy or endocrine plus chemotherapy) according to the test results or according to usual clinical practice. The predictive value is assessed by comparing the outcomes from the testing-based arm versus the non-testing arm; (b) *Treatment interaction design*: patients are divided into groups based on the marker status (i.e. high and low marker status). Then they are allocated to receive endocrine therapy or endocrine treatment plus chemotherapy. The predictive value is assessed by observing the relative efficacy of treatment differences between marker status and treatment assignments.

Studies that only reported prognosis data based on marker status (without considering differential treatment effect), individual observational studies, abstracts or conference communications not published as full text articles, and articles published in a language other than English were excluded.

With respect to economic evidence, cost-utility, cost-benefit, and cost-consequences, analyses were included if conducted within clinical trials, as well as observational and modelling studies, published in English during the last decade (Supplementary Table 3).



After a calibration process, each reviewer (CCA and KP) assessed titles and abstracts for eligibility. Subsequently, two reviewers (CCA and KP), independently, reviewed the full text of all the pre-selected references. Discrepancies were solved either by consensus or with the help of a third reviewer (DR) (Supplementary Fig. 2 and Supplementary Table 4).

**Data extraction and risk of bias assessment.** Two reviewers (CCA and KP) independently assessed risk of bias and extracted the following information from each study: first author, year of publication, country, study design, inclusion and exclusion criteria, number of patients, age, participants' characteristics and prioritised outcomes.

The risk of bias of the included RCTs was assessed using the Cochrane Risk of Bias tool for randomised trials.<sup>19</sup> Cohort studies were assessed with the 'Risk Of Bias In Non-randomised Studies - of Interventions' (ROBINS-I) tool.<sup>20</sup> For economic evaluations, one reviewer (MP) screened the search results and used the NICE methodology checklist to assess applicability and methodological limitations.<sup>21</sup> Studies with poor applicability and/or high risk of bias were excluded (Supplementary Fig. 2).

**Data analysis.** Descriptive statistics were used to summarise the characteristics of the included patients across studies. The effect measures for prioritised outcomes and their corresponding 95% confidence intervals (CIs) were reported as presented in individual studies.

**Certainty of the evidence.** The certainty of evidence per outcome and overall certainty was rated using the GRADE approach. For each recommendation, the GDG received a Summary of Findings (SoF) table and a first draft of an evidence to decision framework (EtD).<sup>22</sup>

#### Comparison scenarios and modelling

A simple deterministic decision tree model without discounting was built by PGR with input from the rest of the GDG to estimate the downstream consequences of testing patients with the multigene tests versus different scenarios of usual care (Supplementary Fig. 3). For the 21-RS, the general model assumptions were: the population of eligible women was divided into the three risk groups, as reported in the TAILORx trial at recruitment until 2008 (14% low risk of recurrence, 68% intermediate, 18% high).<sup>23</sup> Rate of events, observed in the RCTs, were applied to the simulated usual care arms. Clinical risk of recurrence was classified as low and high according to the modified AdjuvantOnlineScore.<sup>24,25</sup> Results are based on a fixed observation time of 10 years.

Two strategies for implementing the multigene test to guide the use of adjuvant chemotherapy were considered as interventions (Supplementary Fig. 3a):

1. All women would undergo multigene testing and adjuvant chemotherapy would be given accordingly (only to those classified in the high genomic risk group, i.e. with a score  $\geq 26$ ).
2. Only women with high clinical risk would undergo multigene testing, and only those with high genomic risk would receive chemotherapy. Women with low clinical risk would not receive it.

Two scenarios were considered as usual care comparators ('C' of the PICO framework) (Supplementary Fig. 3b):

1. All women would be referred to adjuvant chemotherapy (assuming 18.4% would not comply, i.e. the proportion of women not receiving chemotherapy among those assigned to the treatment arm in TAILORx).<sup>26</sup>
2. Women would receive adjuvant chemotherapy only if the clinical risk is high. The model assumes that women with

low clinical and high genomic risk, as well as those with low clinical and low genomic risk, do not benefit from adjuvant chemotherapy. For sensitivity analyses, in women of low clinical and high genomic risk, two different assumptions were used to estimate the benefits: (a) The advantage of receiving adjuvant chemotherapy is equivalent to that observed in the MINDACT trial<sup>6</sup> at five years, and the effect is maintained at 10 years; (b) the advantage from adjuvant chemotherapy is equivalent to that observed by Paik and colleagues<sup>26</sup> for all women at high genomic risk, independent of their clinical risk. Distributions of the clinical risk within the multigene risk strata are those reported in the TAILORx trial.<sup>23</sup>

For the 70-GS, we focused only on comparing strategy 2 with scenario 2 in which women at high clinical risk would be tested and/or treated, because the evidence from the MINDACT trial indicates a very small benefit, if any, from adjuvant chemotherapy in women with low clinical risk, independent of their genomic risk.<sup>6</sup>

#### Evidence to decision and recommendation formulation

The process the ECIBC GDG used to formulate recommendations has been described in a dedicated article published elsewhere.<sup>11</sup> In brief, a subgroup of GDG members including experts on the topic and an informed patient (the so-called PICO responsible unit), took primary responsibility for the review and completion of the first draft of the SoF tables and the EtD frameworks, conducted initially by the systematic review team. The frameworks were used in the meetings to help the complete GDG formulate the recommendations. Subsequently they were reviewed by a technical team from the Joint Research Centre, the PICO responsible unit and the systematic review team. Finally, the recommendations and frameworks were approved by the GDG.

## RESULTS

### Included studies

We included five studies (Supplementary Fig. 1): two RCTs,<sup>6,23</sup> two secondary analyses of stored tissue blocks collected from former parent clinical trials<sup>26,27</sup> and one pooled analysis of observational studies<sup>12</sup> from four previously reported validation studies, including unpublished data (Supplementary Table 5).<sup>28-30</sup>

### 21 gene recurrence score

**Treatment interaction design studies.** Paik and colleagues<sup>26</sup> provided estimates for distant recurrence free survival in patients with lymph node-negative breast cancer stratified into three levels of the 21-RS risk groups. Adding chemotherapy to endocrine therapy, compared to endocrine therapy alone, may have a different effect on recurrence across groups, i.e. a larger effect in women with higher 21-RS, but the evidence is very uncertain: hazard ratio (HR) of 1.31 (95% CI 0.46-3.78), 0.61 (95% CI 0.24-1.59) and 0.26 (95% CI 0.13-0.53) in low, intermediate and high risk groups, respectively (Supplementary Table 6).

Albain and colleagues<sup>27</sup> included stored tumour specimens for genomic testing of postmenopausal women with HbR-positive, node-positive breast cancer. They performed an analysis adjusted by the number of positive nodes that suggests no benefit for chemotherapy on disease free survival (DFS) in the low genomic risk group (HR = 1.02; 95% CI 0.54-1.93) and a potential advantage in the high genomic risk (HR = 0.59, 95% CI 0.35-1.01) (Supplementary Table 6). The authors refer similar results for overall survival (OS).

**Marker-based strategy.** Sparano and colleagues<sup>23</sup> provided results for several disease-free survival (DFS)-related outcomes among women with an intermediate genomic risk group (11 to 26

risk score) allocated to either endocrine therapy alone or chemotherapy plus endocrine therapy. The as-treated results suggest little to no difference in the risk of recurrence with chemotherapy plus endocrine therapy for invasive DFS (HR 1.14; 95% CI 0.99–1.31). For distant metastases, local recurrence and OS, similar results were observed (Supplementary Table 6).

#### 70-GS

**Treatment interaction design studies.** Knauer and colleagues<sup>12</sup> described results from a pooled database analysis with a median follow-up time of 7.1 years. Patients with a low and high genomic risk who received chemotherapy may have a lower risk of recurrence than those with endocrine therapy alone, but the evidence is very uncertain (HR 0.26; 95% CI 0.03–2.02 and HR 0.35; 95% CI 0.17–0.71, respectively). The results for mortality were consistent with the observed pattern of the risk of recurrence but the evidence was also very uncertain (HR 0.58; 95% CI 0.07–4.98; HR 0.21; 95% CI 0.07–0.59, respectively) (Supplementary Table 7).

**Marker-based strategy.** Cardoso and colleagues<sup>6</sup> reported DFS and OS among patients in the clinical/genomic discordant-risk groups which were allocated to receive either chemotherapy in addition to endocrine therapy or endocrine therapy alone.

Women with high clinical risk and low genomic risk may have an increase of DFS (HR 0.64; 95% CI 0.43–0.95), of distant metastases free survival (HR 0.65; 95% CI 0.38–1.10) and of OS (HR 0.63; 95% CI 0.29–1.37) (Supplementary Table 7). The group of low clinical risk and high genomic risk showed imprecise effects and uncertain evidence for DFS (HR 0.74; 95% CI 0.40–1.39), distant metastases free survival (HR 0.90; 95% CI 0.40–2.01) and OS (HR 0.72; 95% CI 0.23–2.24).<sup>6</sup>

**Modelling for predicting impact of testing on patient's outcomes** Depending on the different treatment scenarios (all women are referred to chemotherapy or only women with high clinical risk are treated with chemotherapy) and genetic testing strategies (genetic testing carried out in all women or testing only those with high clinical risk), the number of women who avoid chemotherapy by using the 21-RS would change from more than 600 to about 200. Survival outcomes did not change substantially (Table 2) for the 21-RS based on the benefits from adding chemotherapy in the MINDACT trial.<sup>6</sup> However, on the assumption that all women with high genomic score would obtain the same benefits from adding chemotherapy as observed by Paik and colleagues,<sup>26</sup> independently from their clinical risk, the intervention could potentially prevent 37 distant metastases compared to a scenario in which only women with high clinical risk would be treated with chemotherapy.

On the other hand, we considered for the 70-GS a two-step strategy according to the results of the MINDACT trial, testing only women with high clinical risk.<sup>6</sup> Consequently, the only scenario considered was one in which only high-risk women would receive chemotherapy (Table 3). The use of the 70-GS would result in an avoidance of chemotherapy in about 230 women out of 1000 associated with small increase of recurrences.

#### Results from the systematic review of economic evidence

From the primary literature search and from the two identified systematic reviews,<sup>31,32</sup> 12 cost-effectiveness evaluations were identified (Supplementary Fig. 2 and Supplementary Table 4).<sup>31–44</sup> The GDG agreed that these economic evaluations used models that were not directly applicable to the clinical question of interest. Therefore, cost-effectiveness was evaluated considering the benefits and harms estimated using the GDG's ad-hoc model described above (Supplementary Fig. 3) and the costs reported by the studies included in the literature review. Eight studies reported costs for the use of the 21-RS in women with negative lymph nodes,<sup>31–41</sup> whereas three reported costs for women with up to 3 positive lymph nodes.<sup>42–44</sup> The reported costs of the 21-RS were

**Table 2.** Anticipated outcomes for different comparisons between the 21-gene recurrence score testing strategies (interventions) and comparator scenarios (no testing) per 1000 women with hormone receptor-positive, HER2-negative, lymph node-negative invasive breast cancer.

Outcome	Intervention strategy 1 <sup>a</sup>	Comparator scenario 1 <sup>b</sup>	Absolute difference	Intervention strategy 1 <sup>a</sup>	Comparator scenario 2 <sup>b</sup>	Absolute difference	Intervention strategy 2 <sup>a</sup>	Comparator scenario 2 <sup>b</sup>	Absolute difference
Number of	180	800	-620	180	314	-134	103	314	-211
Chemotherapy									
Invasive disease	80	78	2	80	79	-1	80	79	-1
recurrence									
Distant recurrence	27	26	1	28	27	0	27	27	0
Local and distant	39	38	1	39	39	0	39	39	0
recurrence									
Deaths	24	23	1	24	24	0	24	24	0

<sup>a</sup>According to Supplementary Fig. 3a.

<sup>b</sup>According to Supplementary Fig. 3b.



EUR 3180 per patient in five out of the 11 studies included. These costs did not show significant differences between countries or over time. For the 70-GS, two studies reported costs of EUR 3153 per patient for the use of this assay in women with negative lymph nodes.<sup>45,46</sup>

#### Certainty of evidence

The overall certainty of the evidence was rated as low to very low. The main concerns across studies were risk of bias, indirectness of trial populations and imprecision (Supplementary Tables 6 and 7). The evidence was also downgraded for indirectness due to the assumptions used for the model that implied the use of evidence from one population in another population and from different duration of follow-up across studies.

#### Evidence to decision frameworks

**21-RS.** The GDG judged the anticipated desirable effects (i.e. the avoided chemotherapy treatments) of using the test to guide chemotherapy to be large and the undesirable effects (i.e. increase in recurrence) trivial, with very low certainty of the evidence. The costs were considered large, though no cost-effectiveness study

was included. A negative impact on equity was considered a potential concern (Table 4).

For women with HoR-positive, HER2-negative, node-negative invasive breast cancer, the ECIBC GDG suggests the use of the 21-RS to guide the use of chemotherapy (conditional recommendation, very low certainty of the evidence, Table 5). The recommendation is conditional because the certainty of evidence was very low and the downstream consequences of avoiding chemotherapy were not quantified, thus making the balance of benefits and harms difficult to determine, together with the large resource (costs) requirements (Table 4, see also [https://healthcare-quality.jrc.ec.europa.eu/sites/default/files/Guidelines/ETDs/Updated/ECIBC\\_GLS\\_ETD\\_21\\_gene\\_recurrence\\_score.pdf](https://healthcare-quality.jrc.ec.europa.eu/sites/default/files/Guidelines/ETDs/Updated/ECIBC_GLS_ETD_21_gene_recurrence_score.pdf)).

The GDG did not consider women with node-positive invasive breast cancer in this recommendation, because they were not included in the TAILORx trial,<sup>23</sup> the main source for model parameters. The GDG also stated that sub-populations with high clinical risk (defined according to AdjuvantOnline!<sup>24,25</sup>) may experience larger net desirable consequences and provide a more favourable cost-effectiveness profile (Fig. 1). On the other hand, women with low clinical risk may experience smaller or no net desirable consequences. Indirect evidence from the MINDACT trial using the 70-GS supports that conclusion. In fact, in this trial there are very small, if any, benefits from chemotherapy in low clinical risk women, independently of the genomic risk.<sup>6</sup>

New relevant results have been published on the 21-RS since the systematic review used for this recommendation was conducted.<sup>45,46</sup> Recent data from the TAILORx trial stratified by age and clinical risk have been published.<sup>45</sup> The authors suggest that women below 50 with an intermediate genomic risk score could have a benefit from adding chemotherapy to endocrine therapy if their clinical risk is high but the absence of a dose response and very imprecise effect estimates suggest that chance could play a major role. Furthermore, the reported analysis suggests that data for evaluating the potentially most efficient two-step testing strategy is missing.<sup>47,48</sup> Mariotto and colleagues<sup>49</sup> showed that application of the 21-RS risk to decide whether or not to provide chemotherapy would produce savings in the actual US real clinical practice. Another analysis using the same data

**Table 3.** Anticipated outcomes for the comparison between the 70-gene signature assay testing strategy (intervention) and comparator scenario (no testing) per 1000 women with hormone receptor-positive, HER2-negative, lymph node-negative or up to 3 lymph nodes-positive invasive breast cancer.

Outcome	Intervention strategy 2 <sup>a</sup>	Comparator scenario 2 <sup>b</sup>	Absolute difference
Treated women	270	501.4	-232
Distant disease	53	48	4.5
Disease free	100	93	7
Deaths	30	26	3.5

<sup>a</sup>According to supplementary Fig. 3a.

<sup>b</sup>According to supplementary Fig. 3b.

**Table 4.** Judgements by the Guideline Development Group (GDG) in Evidence to decision framework for the question: Should multigene tests be used in patients who have hormone receptor-positive, HER-2 negative, lymph node-negative or up to 3 lymph nodes-positive invasive breast cancer to guide the use of adjuvant chemotherapy?

	21-gene recurrence score (limited to women with negative lymph nodes)	70-gene signature assay	
		Low clinical risk	High clinical risk
Problem	Yes	Yes	Yes
Desirable effects	Large	Trivial	Large
Undesirable effects	Trivial	Trivial	Small
Certainty of evidence	Very low	Low	Low
Values	Probably no important uncertainty or variability	Probably no important uncertainty or variability	Probably no important uncertainty or variability
Balance of effects	Probably favours the intervention	Favours the comparison	Probably favours the intervention
Resources required	Large costs	Large costs	Large savings
Certainty of evidence of required resources	Very low	Very low	Very low
Cost effectiveness	No included studies	No included studies	No included studies
Equity	Probably reduced	Probably reduced	Probably reduced
Acceptability	Varies	Varies	Varies
Feasibility	Varies	Varies	Varies
Final recommendation	Conditional in favour of the intervention	Strong against intervention	Conditional in favour of the intervention

**Table 5.** European Commission Initiative on Breast Cancer guidelines development group recommendations on the use of multigene tests to guide the use of adjuvant chemotherapy in patients with early breast cancer, hormone receptor-positive, HER-2 negative.

	21-gene recurrence score	70-gene signature	
		Women with low clinical risk <sup>a</sup>	Women with high clinical risk <sup>a</sup>
Recommendation	For women with hormone receptor-positive, HER2-negative, lymph node-negative invasive breast cancer, the ECIBC's Guidelines Development Group (GDG) suggests using the 21-gene recurrence score to guide the use of chemotherapy.	For women with hormone receptor-positive, HER2-negative, lymph node-negative or up to 3 lymph nodes-positive invasive breast cancer at low clinical risk, the ECIBC's Guidelines Development Group (GDG) recommends not using the 70-gene signature test to guide the use of chemotherapy.	For women with hormone receptor-positive, HER2-negative, lymph node-negative or up to 3 lymph nodes-positive invasive breast cancer at high clinical risk, the ECIBC's Guidelines Development Group (GDG) suggests using the 70-gene signature test to guide the use of chemotherapy.
Strength	Conditional recommendation for the intervention Very low certainty of the evidence	Strong recommendation against the intervention Low certainty of the evidence	Conditional recommendation for the intervention Low certainty of the evidence
Sub-group considerations	The GDG did not consider women with node-positive invasive breast cancer to be included in this recommendation. Women with high clinical risk <sup>a</sup> and low genomic risk (larger tumour diameter and higher grade) may experience larger net desirable consequences and provide a better cost-benefit profile. Women with low clinical risk <sup>a</sup> and high genomic risk may experience smaller or no net desirable consequences. Indirect evidence from other gene based testing (e.g. 70-gene signature) supports that conclusion.	The proportion of women with 2 or 3 node-positive breast cancer was small, so the results may be less clear in this subgroup.	The proportion of women with 2 or 3 node-positive breast cancer was small, so the results may be less clear in this subgroup.

<sup>a</sup>For definitions of low and high clinical risk, see Supplementary Table 8.

indicates that savings would be much larger if testing would be performed according to a two-step strategy.<sup>50</sup> The GDG, for the moment, judged that the new evidence is consistent with the recommendation.

**70-GS.** In light of the results from the MINDACT trial,<sup>6</sup> the GDG decided to split the recommendation according to clinical risk of the population under study (at low and high clinical risk, Table 5 and Fig. 1). In the low clinical risk group the GDG recommends against using the 70-GS testing to guide the use of chemotherapy (strong recommendation, very low certainty of the evidence) as there are no apparent benefits and there are very large costs (EUR 3153 per patient) ([https://healthcare-quality.jrc.ec.europa.eu/sites/default/files/Guidelines/ETDs/Updated/ECIBC\\_GLS\\_ETD\\_70\\_gene\\_testing\\_low\\_risk.pdf](https://healthcare-quality.jrc.ec.europa.eu/sites/default/files/Guidelines/ETDs/Updated/ECIBC_GLS_ETD_70_gene_testing_low_risk.pdf)).

For women with hormone receptor-positive, HER2-negative, node-negative or up to 3 lymph nodes-positive invasive breast cancer at high clinical risk, the ECIBC GDG suggests using the 70-GS test to guide the use of chemotherapy. The judgments favoured the intervention in the high clinical risk population due to the moderate desirable effect, a balance that probably favours the use of 70-GS testing, and the large savings (Table 4). The recommendation is conditional mainly because of the low certainty of the evidence about the effects. The GDG also stated that the proportion of women with 2 or 3 positive lymph nodes was small, therefore making the results less clear in this subgroup ([https://healthcare-quality.jrc.ec.europa.eu/sites/default/files/Guidelines/ETDs/Updated/ECIBC\\_GLS\\_ETD\\_70\\_gene\\_testing\\_high\\_risk.pdf](https://healthcare-quality.jrc.ec.europa.eu/sites/default/files/Guidelines/ETDs/Updated/ECIBC_GLS_ETD_70_gene_testing_high_risk.pdf)).

## DISCUSSION

### Statement of principal findings

The ECIBC GDG suggests the use of 21-RS in lymph node-negative women, recognising that benefits are probably larger in women at

high clinical risk and suggests the use of the 70-GS only for women at high clinical risk.

### Strengths and weaknesses of the study

The strength of the recommendations includes the ECIBC's adherence to the requirements for trustworthy development of guidelines.<sup>51-53</sup> Previously, we described some limitations of our guidelines.<sup>11,54</sup> The weakness of the deterministic decision tree model used is that it is, to some extent, a simplistic approach and some assumptions are questionable (i.e. negligible effects in low clinical risk, same effects in studies with different duration of follow-up). Furthermore, we did not actually quantify the side effects of chemotherapy, considering that avoiding any unnecessary chemotherapy was a desirable effect.

### Relation to other guidelines

NICE recently published guidance on multigene testing.<sup>55</sup> The panel decided to evaluate the evidence for the four commercially available multigene tests included in the ECIBC question, and for the IHC4 + C test.<sup>56,57</sup> The NICE guidelines did not follow the GRADE methodology and also had a different goal, i.e. deciding which tests should be funded by the UK National Health Service. The NICE 21-RS recommendation is similar to the ECIBC recommendation, a conditional recommendation limited to those patients in which the risk of distant recurrence is intermediate, using a validated tool such as PREDICT<sup>58</sup> or the Nottingham Prognostic Index (Table 6). This approach is based on the assumption that in some patients the clinical and pathological prognostic parameters are consistent (low or high risk), and that multigene testing does not provide additional information. In particular, the NICE panel recommended against the use of the 70-GS, based on cost-effectiveness considerations (Table 6). Unfortunately, the NICE guideline does not allow comparison of our estimates of desirable and undesirable health effects. Some



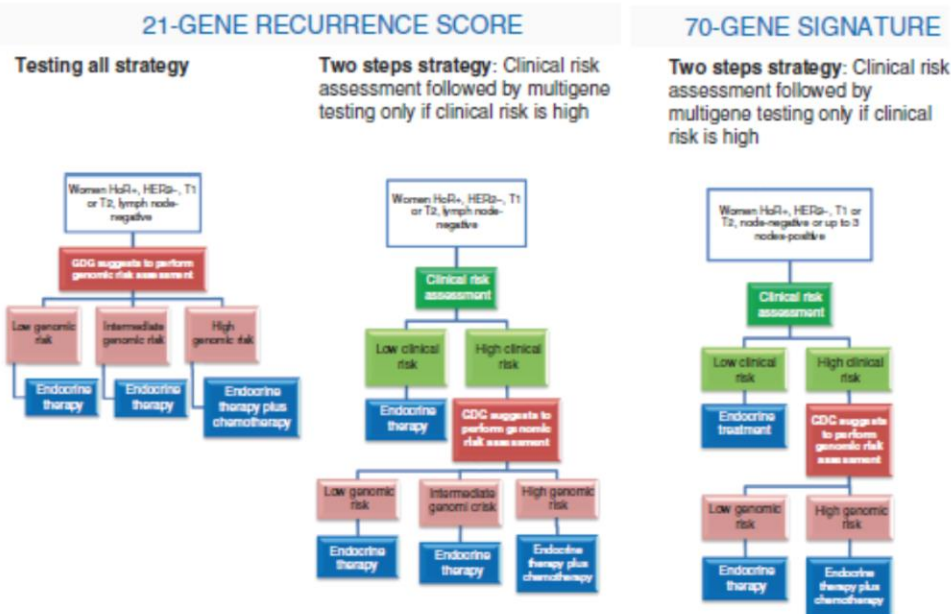


Fig. 1 Flow charts reporting the possible uses of 21-gene recurrence score and for the 70-gene signature test to guide the use of adjuvant chemotherapy in patients with early invasive breast cancer, hormone receptor-positive, HER-2 negative. Two strategies are proposed for 21-gene recurrence score, the first in which all women are tested for genomic risk assessment and treated accordingly, the second in which only women with high clinical risk are tested for genomic assessment, while those at low clinical risk are referred to endocrine therapy alone without genomic risk assessment. According to sub-group considerations reported by the GDG, the latter strategy is probably more cost effective and women might experience larger net desirable consequences. For the 70-gene signature only a two-step strategy is proposed where only women at high clinical risk are tested for genomic risk; testing women at low clinical risk is not recommended.

**Table 6.** Synopsis of American Society of Clinical Oncology (ASCO)<sup>59-61</sup> and the UK National Institute for Health and Care Excellence (NICE)<sup>62</sup> recommendations on 21-gene recurrence score and 70-gene signature assay in hormone receptor-positive, HER2-negative, lymph node-negative or up to 3 nodes-positive invasive breast cancer.

	ASCO <sup>59,60</sup>	NICE <sup>64</sup>
<b>21-gene recurrence score</b>		
Low clinical risk	Strong recommendation	
High clinical risk	Strong recommendation	Recommended if: clinical risk is "intermediate" according to the PREDICT tool <sup>63</sup> or the Nottingham Prognostic Index, i.e. the additional benefit of chemotherapy is between 3 and 5% increase in survival, the decision on the therapy will depend on the test result, the test is provided at reduced price.
Negative lymph nodes	Strong recommendation	Recommendation applies to both lymph node-negative patients and lymph node-positive patients, restricted to micro-metastases
1 to 3 positive lymph nodes	Not recommended	
	ASCO <sup>60</sup>	NICE <sup>65</sup>
<b>70-gene signature assay</b>		
Low clinical risk	Strong against	Recommendation against because not cost effective
High clinical risk	Strong in favour	
Negative lymph nodes	Strong in favour	
1 to 3 positive lymph nodes	Moderate in favour	

methodological differences might explain the divergent results: since real practice varies across Europe, we preferred to use theoretical scenarios as comparators and interventions not accounting for non-compliance, while the NICE model used real practice as comparator (the prevalence of chemotherapy used in this group of women in the UK), and as intervention (a change in the probability of receiving chemotherapy given the test result). Furthermore, the model used by NICE assumed that the tests are only prognostically relevant but are not predictive of the response to treatment, i.e. they calculated a constant HR of 0.77 for chemotherapy in addition to endocrine therapy compared to endocrine therapy alone in all risk groups. In contrast, the ECIBC GDG judged the benefits to be trivial or small in the low clinical risk group. Finally, for the 21-RS a commercial-in-confidence discounted test cost was used to model cost-effectiveness, while for the 70-GS the regular market price was used. It is worth noting that despite NICE stating that the major benefit of the genetic testing strategies would be a reduction of chemotherapy, the cost models predict health benefits only if chemotherapy is increased.

The American Society of Clinical Oncology (ASCO) also provided recommendations on the use of multigene tests.<sup>19-21</sup> Despite a different methodological approach, the direction of the recommendations is the same, but the strength is not (Table 6). There are differences in the grading of certainty of the evidence, considered as high by the ASCO panelists, while the GDG valued the evidence as very low for the 21-RS and low for the 70-GS. Unfortunately, we were unable to deduce the details of the ASCO evidence rating approach and also the criteria and judgments that were used to determine the strength of the recommendations. Unlike us, ASCO and NICE made recommendations on the 12-M5 and the PAM50-RS to guide treatment decisions,<sup>22,23</sup> mainly because they did not exclude studies based on prognostic results only.

#### Meaning of the study

The implications of our recommendations are context dependent. The criteria used for making decisions on the provision or not of adjuvant treatment differ between countries. Therefore, the cost-benefit profile of introducing one of the multigene tests might also vary across countries. Decreasing costs for the tests would support a more widespread use. For these considerations, the GDG decided not to establish a threshold of recurrence risk to recommend genomic risk assessment, or a threshold for adding adjuvant chemotherapy, since these thresholds are context specific.

In conclusion, the ECIBC GDG recommendations for or against the use of 21-RS and 70-RS are justified based on the judgments made. The transparency of our approach allows understanding the rationale for making different recommendations for the two tests and risk groups.

#### Unanswered questions and future research

Data protection issues may be of relevance for the 21-RS because processing of samples is centralised in one US lab and requires shipping samples abroad. Furthermore, transparent information on test results is not available and reproducibility has been questioned.<sup>6,3</sup>

The GDG recommends research on exploring in what subgroups the use of 21-RS would have larger anticipated benefits as well as carrying out longer follow-up studies for 70-GS. The recommendations will be updated according to the ECIBC monitoring strategy in place (<https://healthcare-quality.jrc.ec.europa.eu/discover-ecibc/methodologies/guidelines-updating>).

Furthermore, the GDG is exploring the possibility of evaluating biomarkers that may assist decision making regarding the administration of adjuvant chemotherapy on the basis of their ability to identify women with a sufficiently low risk of relapse that would allow them to be spared from chemotherapy. In contrast to

the presented evidence evaluation, this would enable evidence-based and transparent recommendations based on prognostic cohort studies, randomised or not, that predict the recurrence risk of different subgroups. We are currently working on a healthcare question on the significance of Ki67 using this strategy. In a next step, such an approach might also make it possible to evaluate multigene tests in the assessment for which we found no usable evidence in the predictive search strategy used here, but for which data on the prognostic value are available.<sup>13-17,24</sup>

#### ECIBC CONTRIBUTOR GROUP

Mariangela Autelitano<sup>18</sup>, Bettina Borch<sup>19</sup>, Xavier Castells<sup>20</sup>, Edward Colanzi<sup>21</sup>, Jan Daniel<sup>22</sup>, Patricia Fitzpatrick<sup>23</sup>, Ulva Giordano<sup>24</sup>, Solveig Holmvid<sup>25</sup>, Lydia Ioannidou-Mouzaki<sup>26</sup>, Susan Knox<sup>27</sup>, Lennarth Nyström<sup>28</sup>, Bina Parmell<sup>29</sup>, Elsa Perez<sup>30</sup>, Alberto Torresin<sup>31</sup>, Ruben Van Engen<sup>32</sup>, Cary Van Landuyt-Vorhagen<sup>33</sup>, Ken Young<sup>34</sup>

<sup>18</sup>Cancer Registry of Milan, Milan, Italy; <sup>19</sup>Institute of Global Health, University of Geneva, Geneva, Switzerland; <sup>20</sup>MIM, Barcelona, Spain; <sup>21</sup>European Centre for Disease Control and Prevention (ECDC), Solna, Sweden; <sup>22</sup>Charles University in Prague, Prague, Czech Republic; <sup>23</sup>National Screening Service, Dublin, Ireland; <sup>24</sup>CPO-Piedmont - AOU Città della Salute e della Scienza, Torino, Italy; <sup>25</sup>Cancer Registry of Norway, Oslo, Norway; <sup>26</sup>University of Athens Medical School, Athens, Greece; <sup>27</sup>Europa Donna, Milan, Italy; <sup>28</sup>Umeå University, Umeå, Sweden; <sup>29</sup>European Commission, Joint Research Centre, Ispra, Italy; <sup>30</sup>University Hospital Dr. Josep Trueta, Girona, Spain; <sup>31</sup>Ospedale Niguarda Ca' Granda, Milan, Italy; <sup>32</sup>Dutch Reference Centre for Screening, Nijmegen, The Netherlands and <sup>33</sup>National Coordinating Centre for the Physics of Mammography, Guildford, UK

#### ACKNOWLEDGEMENTS

To Kevin Pacheco (KP) for his support in the systematic review of effectiveness of the multigene tests described in this publication.

#### AUTHOR CONTRIBUTIONS

All authors contributed to the definition of the research protocol. C.C.A., D.R., P.A.C. and M.P. were responsible for conducting the systematic review. P.G.R. and A.L. prepared the first draft of the article. All authors contributed to the interpretation and reporting of the results. All authors revised the manuscript and provided comments on subsequent versions of the article. All authors read and approved the final manuscript prior to submission and are accountable for all aspects of the work.

#### ADDITIONAL INFORMATION

**Ethics approval and consent to participate** Not applicable.

**Consent to publish** Not applicable.

**Data availability** All data sources used during this study are described in this published article and its additional information files. The datasets analysed are available from the corresponding author on reasonable request.

**Competing interests** Members of the Guideline Development Group (GDG) do not receive financial compensation for their work but are reimbursed by the E.C. for travel-related expenses for the meetings organised by the JRC. Dr. Giorgi Rossi as former PI of an independent study on HPV-based cervical cancer screening, funded by the Italian Ministry of Health, data owner, conducted negotiations with Roche diagnostics, Hologic-Gemprobe, Becton-Dickinson to obtain reagents at reduced price or for free; the reagents obtained were not used in his institution. Dr. Lebeau reports grants and reimbursement for travel-related expenses related to consultancy from Roche Pharma AG, reimbursement for travel-related expenses related to consultancy from Novartis Oncology, and grants from BioNTech Diagnostics GmbH outside the submitted work. Dr. Sz-Parkinson was employed by the European Commission, coordinating the ECIBC Guidelines Development Group. Dr. Quinn is Chair of the European Working Group for Breast Screening Pathology (EWGESP). Various companies have provided some sponsorship to the EWGESP for group meetings. Dr. Grdwinholt is the responsible radiologist for screening unit Paderborn, Germany, consultant radiologist for screening programs in Switzerland, and consultant radiologist for Hellenic School of Senology. Dr. Canelo-Aybar, Dr. Rigau, Dr. Posso Rivera, and Dr. Alonso-Coello reports that his institution received payments from the European Commission to develop the systematic reviews informing the



recommendations. Authors not named here have disclosed no conflicts of interest. Disclosures can also be viewed at [www.aoponline.org/authors/fcmj/ConflictOfInterestForm.do?msNum=M18-3445](http://www.aoponline.org/authors/fcmj/ConflictOfInterestForm.do?msNum=M18-3445).

**Funding information** This work was supported by the European Commission. The EC did not have any role in the study design, collection, analysis and interpretation of the data. The researchers were independent of the funders and all authors, external and internal, had full access to all of the data (including statistical reports and tables) in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. Dissemination declaration: Results will be disseminated through the EBC website (<http://healthcare-quality.ec.europa.eu/european-breast-cancer-guidelines>) and specific women versions of these recommendations will also be made available to facilitate access to lay people.

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41416-020-01247-z>.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## REFERENCES

1. Forlay, J. E. M., Lam, F., Colombet, M., Mery, L., Pfisterer, M., Znaor, A. et al. *Global Cancer Observatory: Cancer Today* (International Agency for Research on Cancer, Lyon, France, 2018).
2. EBC. European Cancer Information System From <https://ecis.jrc.ec.europa.eu> (2018).
3. Howlader, N., Altekruse, S. F., Li, C. I., Chen, V. W., Clarke, C. A., Ries, L. A. G. et al. US incidence of breast cancer subtypes defined by joint hormone receptor and HER2 status. *J. Natl Cancer Inst.* **28**, 106 (2014).
4. Early Breast Cancer Trialists' Collaborative Group. Effects of chemotherapy and hormonal therapy for early breast cancer on recurrence and 15-year survival: an overview of the randomised trials. *Lancet* **365**, 1687–1717 (2005).
5. Early Breast Cancer Trialists' Collaborative Group. Relevance of breast cancer hormone receptors and other factors to the efficacy of adjuvant tamoxifen: patient-level meta-analysis of randomised trials. *Lancet* **378**, 771–784 (2011).
6. Cardoso, F., van't Veer, L. J., Bogaerts, J., Slaats, L., Viale, G., Delalogu, S. et al. 70-gene signature as an aid to treatment decisions in early-stage breast cancer. *N. Engl. J. Med.* **375**, 717–729 (2016).
7. Goldhirsch, A., Wood, W. C., Coates, A. S., Gelber, R. D., Therasse, B., Senn, H. J. et al. Strategies for subtypes—dealing with the diversity of breast cancer: highlights of the St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2011. *Ann. Oncol.* **22**, 1736–1747 (2011).
8. Coates, A. S., Winer, E. P., Goldhirsch, A., Gelber, R. D., Grant, M., Rocca-Gelbhart, M. et al. Tailoring therapies—improving the management of early breast cancer: St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2015. *Ann. Oncol.* **26**, 1533–1546 (2015).
9. Petrelli, F., Viale, G., Cabiddu, M. & Barni, S. Prognostic value of different cut-off levels of Ki-67 in breast cancer: a systematic review and meta-analysis of 64,196 patients. *Breast Cancer Res. Treat.* **153**, 477–491 (2015).
10. Early Breast Cancer Trialists' Collaborative Group. Comparisons between different polychemotherapy regimens for early breast cancer: meta-analyses of long-term outcome among 100,000 women in 123 randomised trials. *Lancet* **379**, 432–444 (2012).
11. Schumacher, H. J., Lenda, D., Dimitrova, N., Alonso-Codillo, P., Grønhøj, A., Quinn, C. et al. Methods for development of the European Commission Initiative on Breast Cancer Guidelines: Recommendations in the Bo of Guideline Transparency. *Ann. Intern. Med.* <https://doi.org/10.7326/M18-3445> (2018).
12. Knauer, M., Mook, S., Rutgers, E. J., Bender, R. A., Hauptmann, M., van de Vijver, M. J. et al. The predictive value of the 70-gene signature for adjuvant chemotherapy in early breast cancer. *Breast Cancer Res. Treat.* **120**, 655–661 (2010).
13. Dowsett, M., Sestak, I., Lopez-Knowles, E., Sidhu, K., Durbior, A. K., Cowens, J. W. et al. Comparison of PAM50 risk of recurrence score with oncoprint DX and IHC4 for predicting risk of distant recurrence after endocrine therapy. *J. Clin. Oncol.* **31**, 2783–90 (2013).
14. Grant, M., Sestak, I., Filips, M., Dowsett, M., Balic, M., Lopez-Knowles, E. et al. Identifying clinically relevant prognostic subgroups of postmenopausal women with node-positive hormone receptor-positive early-stage breast cancer treated with endocrine therapy: a combined analysis of ABCSG8 and ATAC using the PAM50 risk of recurrence score and intrinsic subtype. *Ann. Oncol.* **26**, 1685–1691 (2015).
15. Sestak, I., Cuzick, J., Dowsett, M., Lopez-Knowles, E., Filips, M., Dubsley, P. et al. Prediction of late distant recurrence after 5 years of endocrine treatment: a combined analysis of patients from the Austrian breast and colorectal cancer study group 8 and arimidex, tamoxifen alone or in combination randomized trials using the PAM50 risk of recurrence score. *J. Clin. Oncol.* **33**, 916–922 (2015).
16. Buis, R., Sestak, I., Kronenwett, R., Denkert, C., Dubsley, P., Krappmann, K. et al. Comparison of EndoPredict and EPclin with oncoprint DX recurrence score for prediction of risk of distant recurrence after endocrine therapy. *J. Natl Cancer Inst.* **108**, djw140 (2016).
17. Sestak, I., Buis, R., Cuzick, J., Dubsley, P., Kronenwett, R., Denkert, C. et al. Comparison of the performance of 6 prognostic signatures for estrogen receptor-positive breast cancer: a secondary analysis of a randomized clinical trial. *JAMA Oncol.* **4**, 546–553 (2018).
18. Guyatt, G. H., Oxman, A. D., Kunz, R., Adkins, D., Brook, J., Vist, G. et al. GRADE guidelines: 2. Framing the question and deciding on important outcomes. *J. Clin. Epidemiol.* **64**, 395–400 (2011).
19. Higgins, J. P., Altman, D. G., Gøtzsche, P. C., Jüni, P., Moher, D., Oxman, A. D. et al. The cochrane collaboration's tool for assessing risk of bias in randomised trials. *BMJ* **343**, d5928 (2011).
20. Sterne, J. A., Hernán, M. A., Reeves, B. C., Savovic, J., Barkman, N. D., Viswanathan, M. et al. ROBINS-4: a tool for assessing risk of bias in non-randomised studies of interventions. *BMJ* **355**, i4019 (2016).
21. NICE (National Institute for Health and Care Excellence). The guidelines manual: appendix G. NICE methodology checklist for economic evaluations. <https://www.nice.org.uk/process/pmg9/resources/the-guidelines-manual-appendix-bi-2549703709/chapter/appendix-g-methodology-checklist-economic-evaluations> (2012).
22. Alonso-Codillo, P., Oxman, A. D., Moher, J., Brignardello-Peterson, R., Alt, E. A., Davoli, M. et al. GRADE evidence to decision (EtD) frameworks: a systematic and transparent approach to making well informed healthcare choices. 2: clinical practice guidelines. *BMJ* **353**, i2089 (2016).
23. Sparano, J. A., Gray, R. J., Makower, D. F., Pritchard, K. I., Albain, K. S., Hayes, D. F. et al. Adjuvant chemotherapy guided by a 21-gene expression assay in breast cancer. *N. Engl. J. Med.* **379**, 111–121 (2018).
24. Ravdin, P. M., Siminoff, L. A., Davis, G. J., Mercer, M. B., Hawick, J., Garzon, N. et al. Computer program to assist in making decisions about adjuvant therapy for women with early breast cancer. *J. Clin. Oncol.* **19**, 980–991 (2001).
25. Olivotto, I. A., Bajdik, C. D., Ravdin, P. M., Sparano, J. A., Goldstein, A. J., Nomi, B. D. et al. Population-based validation of the prognostic model ADJUVANT for early breast cancer. *J. Clin. Oncol.* **23**, 2716–2725 (2005).
26. Park, S., Tang, G., Shak, S., Kim, C., Baker, J., Kim, W. et al. Gene expression and benefit of chemotherapy in women with node-negative, estrogen receptor-positive breast cancer. *J. Clin. Oncol.* **24**, 3726–3734 (2006).
27. Albain, K. S., Barlow, W. E., Shak, S., Hortobagyi, G. N., Livingston, R. B., Yeh, I. T. et al. Breast cancer intergroup of north america: prognostic and predictive value of the 21-gene recurrence score assay in a randomized trial of chemotherapy for postmenopausal, node-positive, estrogen receptor-positive breast cancer. *Lancet Oncol.* **11**, 55–65 (2010).
28. van de Vijver, M. J., He, Y. D., van't Veer, L. J., Dai, H., Hart, A. A. M., Voskuil, D. W. et al. A gene expression signature as a predictor of survival in breast cancer. *N. Engl. J. Med.* **347**, 1999–2009 (2002).
29. Bueno-de-Mesquita, J. M., Linn, S. C., Krijger, R., Wesseling, J., Nuyten, D. S. A., van Krimpen, C. et al. Validation of 70-gene prognosis signature in node-negative breast cancer. *Breast Cancer Res. Treat.* **117**, 483–495 (2009).
30. Mook, S., Schmidt, M. K., Viale, G., Pruneri, G., Eklund, L., Floore, A. et al. The 70-gene prognosis signature predicts disease outcome in breast cancer patients with 1–3 positive lymph nodes in an independent validation study. *Breast Cancer Res. Treat.* **116**, 295–302 (2009).
31. Blok, E. J., Bastiaant, E., van den Hou, W. B., Liefers, G. J., Smit, V. T. H. B. M., Koope, J. R. et al. Systematic review of the clinical and economic value of gene expression profiles for invasive early breast cancer available in Europe. *Cancer Treat. Rev.* **62**, 74–90 (2018).
32. Wang, S. Y., Dang, W., Richman, I., Mougall, S. S., Evans, S. B. & Gross, C. P. Cost-effectiveness analyses of the 21-gene assay in breast cancer: systematic review and critical appraisal. *J. Clin. Oncol.* **36**, 1619–1627 (2018).
33. Davidson, J. A., Cromwell, L., Bland, S. L., Lohrich, C., Galmon, K. A., Shenker, T. et al. A prospective clinical utility and pharmacoeconomic study of the impact of the 21-gene Recurrence Score(R) assay in estrogen receptor positive node negative breast cancer. *Eur. J. Cancer* **49**, 2469–2475 (2013).
34. Jahn, B., Rochau, U., Kurzthaler, C., Hubalek, M., Mikszad, R., Soczynski, G. et al. Personalized treatment of women with early breast cancer: a risk-group specific cost-effectiveness analysis of adjuvant chemotherapy accounting for companion prognostic tests OncotypeDX and AdjuvantOnline. *BMC Cancer* **17**, 685 (2017).
35. Katz, G., Romano, O., Foa, C., Vataine, A. L., Chantalat, J. V., Hervé, R. et al. Economic impact of gene expression profiling in patients with early-stage breast cancer in France. *PLoS ONE* **10**, e0128880 (2015).
36. Lyman, G. H., Cook, L. E., Naderi, N. M. & Homburger, J. Impact of a 21-gene RT-PCR assay on treatment decisions in early-stage breast cancer: An economic



- analysis based on prognostic and predictive validation studies. *Cancer* 109, 1011–1018 (2007).
37. Ontario Health Technology Advisory Committee, Medical Advisory Secretariat. Gene expression profiling for guiding adjuvant chemotherapy decisions in women with early breast cancer: an evidence-based and economic analysis. *Ont. Health Technol. Assess. Ser.* 10, 1–57 (2010).
  38. Paudyal, M., Faneli, J., Pham, B., Bedard, P. L., Trudeau, M. & Kohn, M. Cost-effectiveness of the 21-gene assay for guiding adjuvant chemotherapy decisions in early breast cancer. *Value Health* 16, 729–739 (2013).
  39. Vataire, A. L., Laas, E., Aballea, S., Gligorov, J., Rouzier, R. & Chéreau, E. Cost-effectiveness of a chemotherapy predictive test. *Bull. Cancer* 99, 907–914 (2012).
  40. Kondo, M., Hoshi, S. I., Ishiguro, H. & Toi, M. Economic evaluation of the 70-gene prognosis-signature (MammaPrint) in hormone receptor-positive, lymph node-negative, human epidermal growth factor receptor type 2-negative early stage breast cancer in Japan. *Breast Cancer Res. Treat.* 133, 759–68 (2012).
  41. Ward, S., Scope, A., Rafia, R., Pandor, A., Heman, S., Bains, P. et al. Gene expression profiling and expanded immunohistochemistry tests to guide the use of adjuvant chemotherapy in breast cancer management: a systematic review and cost-effectiveness analysis. *Health Technol. Assess.* 17, 1–302 (2013).
  42. Bohmer, J. U., Reza, M., Kummel, S., Köhn, T., Wam, M., Friedrich, K. et al. Using the 21-gene assay to guide adjuvant chemotherapy decision making in early-stage breast cancer: a cost-effectiveness evaluation in the German setting. *J. Med. Econ.* 16, 30–40 (2013).
  43. Neirich, V., Curtil, E., Bazan, F., Montcaquet, P., Villanueva, C., Chaigneau, L. et al. Economic assessment of the routine use of Oncotype DX assay for early breast cancer in the Centre-Val de Loire region. *Bull. Cancer* 101, 681–689 (2014).
  44. Vandelaar, B. F., Broder, M. S., Chang, E. Y., Oatz, R. & Bentley, T. G. K. Cost-effectiveness of 21-gene assay in node-positive, early-stage breast cancer. *Am. J. Manag. Care* 17, 455–464 (2011).
  45. Sparano, J. A., Gray, R. J., Makower, D. F., Albain, K. S., Saphner, T. J., Badve, S. S. et al. Clinical outcomes in early breast cancer with a high 21-gene recurrence score of 26 to 100 assigned to adjuvant chemotherapy plus endocrine therapy: a secondary analysis of the TAILORx randomized clinical trial. *JAMA Oncol.* <https://doi.org/10.1001/jamaoncol.2019.4794> (in press).
  46. Sparano, J. A., Gray, R. J., Ravdin, P. M., Makower, D. F., Pritchard, K. I., Albain, K. S. et al. Clinical and genomic risk to guide the use of adjuvant therapy for breast cancer. *N. Engl. J. Med.* 380, 2395–2405 (2019).
  47. Giorgi Rossi, P., Lebeau, A. & Schönmacher, H. J. Clinical and genomic risk in adjuvant therapy for breast cancer. *N. Engl. J. Med.* 381, 1289–1290 (2019).
  48. Dowsett, M. & Turner, N. Estimating risk of recurrence for early breast cancer: integrating clinical and genomic risk. *J. Clin. Oncol.* 37, 689–692 (2019).
  49. Marotto, A., Jayasekera, J., Petkov, V., Schichler, C. B., Enewold, L., Hultsinger, K. J. et al. Expected monetary impact of oncoprint DX score-concordant systemic breast cancer therapy based on the TAILORx trial. *J. Natl. Cancer Inst.* 112, 154–160 (2020).
  50. Giorgi Rossi, P. & Paci, E. Re: expected monetary impact of oncoprint DX score-concordant systemic breast cancer therapy based on the TAILORx trial. *J. Natl. Cancer Inst.* <https://doi.org/10.1093/jnci/kjz125>, in press (2021).
  51. Oemraw, A. D., Frothingham, J. & Schönmacher, H. J. SURE: Improving the use of research evidence in guideline development introduction. *Health Res. Policy Syst.* 4, 12 (2006).
  52. Schönmacher, H. J., Wierdich, W., Branda, I., Falavigna, M., Santos, N., Mustafa, R. et al. Guidelines 2.0: systematic development of a comprehensive checklist for a successful guideline enterprise. *CMAJ* 186, E123–E142 (2014).
  53. Wodt, S., Schönmacher, H. J., Eccles, M. P., Grimshaw, J. M. & Shekelle, P. Developing clinical practice guidelines: types of evidence and outcomes, values and economics, synthesis, grading, and presentation and deriving recommendations. *Implement. Sci.* 7, 61 (2012).
  54. Schönmacher, H. J., Lenda, D., Quinn, C., Follmann, M., Alonso-Coello, P., Giorgi Rossi, P. et al. Breast cancer screening and diagnosis: a synopsis of the European breast guidelines. *Ann. Intern. Med.* <https://doi.org/10.7326/M19-2125>, in press (2021).
  55. NICE National Institute for Health Care and Excellence (2018). Tumour profiling tests to guide adjuvant chemotherapy decisions in early breast cancer. Diagnostic Guidance DG34. <https://www.nice.org.uk/guidance/dg34/resources/tumour-profiling-tests-to-guide-adjuvant-chemotherapy-decisions-in-early-breast-cancer-pdf-1053750722345> (2020).
  56. Cuzick, J., Dowsett, M., Pineda, S., Wallis, C., Sillar, J., Quinn, E. et al. Prognostic value of a combined estrogen receptor, progesterone receptor, Ki-67, and human epidermal growth factor receptor 2 immunohistochemical score and comparison with the Genomic Health recurrence score in early breast cancer. *J. Clin. Oncol.* 29, 4273–4278 (2011).
  57. Barton, S., Zabaglio, L., A'Hern, R., Turner, N., Fergusson, T., O'Neill, S. et al. Assessment of the contribution of the IHC4 + C score to decision making in clinical practice in early breast cancer. *Br. J. Cancer* 106, 1760–1765 (2012).
  58. Candido Dos Reis, F. J., Wishart, G. C., Dicks, E. M., Greenberg, D., Rashbass, J., Schmidt, M. K. et al. An updated PREDICT breast cancer prognostication and treatment benefit prediction model with independent validation. *Breast Cancer Res.* 19, 58 (2017).
  59. Hark, L. N., Ismail, N., McShane, L. M., Andre, F., Collier, D. E., Gonzalez-Angulo, A. M. et al. Use of biomarkers to guide decisions on adjuvant systemic therapy for women with early-stage invasive breast cancer: American Society of Clinical Oncology Clinical Practice Guideline. *J. Clin. Oncol.* 34, 1134–1150 (2016).
  60. Krop, I., Ismail, N., Andre, F., Bast, R. C., Barlow, W., Collier, D. E. et al. Use of biomarkers to guide decisions on adjuvant systemic therapy for women with early-stage invasive breast cancer: American Society of Clinical Oncology Clinical Practice Guideline Focused Update. *J. Clin. Oncol.* 35, 2638–2647 (2017).
  61. Andre, F., Ismail, N., Henry, N. L., Somerfeld, M. R., Bast, R. C., Barlow, W. et al. Use of biomarkers to guide decisions on adjuvant systemic therapy for women with early-stage invasive breast cancer: ASCO Clinical Practice Guideline Update: Integration of Results From TAILORx. *J. Clin. Oncol.* 37, 1956–1964 (2019).
  62. Haman, S., Tappenden, P., Cooper, K., Stevens, J., Besary, A., Rafia, R. et al. Tumour profiling tests to guide adjuvant chemotherapy decisions in early breast cancer: a systematic review and economic analysis. *Health Technol. Assess.* 23, 1–328 (2019).
  63. Schildgen, V., Wam, M., Brodermann, M. & Schildgen, O. Oncotype DX breast cancer recurrence score results: inter-assay reproducibility with RT2-profiler multiplex RT-PCR. *Sci. Rep.* 9, 20266 (2019).
  64. Sestak, I., Martin, M., Dubsky, P., Konecny, R., Rojo, F., Cuzick, J. et al. Prediction of chemotherapy benefit by EndoPredict in patients with breast cancer who received adjuvant endocrine therapy plus chemotherapy or endocrine therapy alone. *Breast Cancer Res. Treat.* 176, 377–386 (2019).



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021

### 6.3 Third study: GRADE Guidelines 30: the GRADE approach to assessing the certainty of modeled evidence—An overview in the context of health decision-making<sup>21</sup>

Researchers should start by conceptualizing the problem and the ideal target model that would best represent the actual phenomenon or decision problem they are considering.<sup>23</sup>

This conceptualization would either guide the development of a new model or serve as a reference against which existing models could be compared. The ideal target model should reflect the following: 1) the relevant population, 2) the exposures or health interventions being considered, 3) the outcomes of interest in that context, and 4) their relationships.<sup>23</sup>

Conceptualizing the model will also reduce the risk of intentional or unintentional development of data-driven models, in which inputs and structure would be determined only by what is feasible to develop given the available data at hand.

#### Outline of an approach to using model outputs for decision-making

Workshop participant identified three options in which users may incorporate model outputs in health decision-making:

1. Develop a model de novo designed specifically to answer the very question at hand.  
Workshop participants agreed that in an ideal situation, such an approach would almost always be the most appropriate. Following this approach, however, requires suitable skills, ample resources, and time being available.
2. Search for an existing model describing the same or a very similar problem and use it “*off-the-shelf*” or adapt it appropriately to answer the current question. In practice, many researchers initially use this approach because of the aforementioned limitations of developing a new model. However, it is often not possible to find an existing model that would be directly relevant to the problem at hand and/or it is not feasible to adapt an existing model when found. Any adaptation of a model requires availability of input data relevant for current problem, appropriate expertise and resources, and access to the original model. The latter is often not available or the structure of the original model is not being transparent enough to allow adaptation.
3. Use the results from multiple existing models found in the literature. This approach may be useful when a limited knowledge about the phenomenon being modeled makes it impossible to decide which of the available models are more relevant, or when many alternative models are relevant but use different input parameters. In such situations, one may be compelled to rely on the results of several models because selection of the single,

seemingly “best” model may provide incorrect estimates of outputs and lead to incorrect decisions.

If a systematic search revealed one or more models meeting the eligibility criteria, then researchers would assess the certainty of outputs from each model. Depending on this assessment, researchers may be able to use the results of a single most direct and lowest risk of bias model “off-the-shelf” or proceed to adapt that model. If researchers failed to find an existing model that would be sufficiently direct and low risk of bias, then they would ideally develop their own model *de novo*.

### **Assessing the certainty of outputs**

When researchers develop their own model or when they identify a single model that is considered sufficiently direct to the problem at hand, they should assess the certainty of its outputs (i.e., evidence generated from that model). Note that if a model estimates multiple outputs, researchers need to assess the certainty of each output separately. Workshop participants agreed that all GRADE domains are applicable to assess the certainty of model outputs, but further work is needed to identify examples and develop specific criteria to be assessed.

#### *Risk of bias in a single model*

The risk of bias of model outputs is determined by the credibility of a model itself and the certainty of evidence for each of model inputs.

The credibility of a model, also referred to as the quality of a model, is influenced by its conceptualization, structure, calibration, validation, and other factors. There are some discipline-specific guidelines or checklists developed for the assessment of credibility of a model and other factor affecting the certainty of model outputs. Workshop participants agreed that there is a need for comprehensive tools developed specifically to assess credibility of various types of models in different modelling disciplines.

The certainty of evidence in each of the model inputs is another critical determinant of the risk of bias in a model. A model has several types of input data, when researchers develop a model *de novo*, to minimize the risk of bias, they need to specify those input parameters to which the model outputs are the most sensitive. Model inputs should reflect the entire body of relevant evidence satisfying clear prespecified criteria rather than an arbitrarily selected evidence that is based on convenience (“any available evidence”) or picked in any other non-systematic way.

The appropriate approach will depend on the type of data and may require performing a systematic review of evidence on each important or crucial input variable.<sup>102, 103</sup> Some inputs may have very narrow inclusion criteria, and therefore, evidence from single epidemiological survey or population surveillance may provide all relevant data for the population of interest (e.g., baseline population incidence or prevalence).

The certainty of evidence for each input needs to be assessed following the established GRADE approach specific to that type of evidence (e.g., estimates of intervention effects or baseline risk of outcomes).<sup>7, 10, 104</sup>

#### *Indirectness in a single model*

By directness or relevance, we mean the extent to which model outputs directly represent the phenomenon being modeled. To evaluate the relevance of a model, one needs to compare it against the conceptual ideal target model. Determining the directness of model outputs includes assessing to what extent the modeled population, the assumed interventions and comparators, the time horizon, the analytic perspective, as well as the outcomes being modeled reflect those that are current interest.

Assessing indirectness in a single model also requires evaluating two separate sources of indirectness:

- Indirectness of input data with respect to the ideal target model's inputs.
- Indirectness of model outputs with respect to the decision problem at hand.

This conceptual distinction is important because, one needs to address each type of indirectness separately. Even if the outputs might be direct to the problem of interest, the final assessment should consider if the inputs used were also direct for the target model.

#### *Inconsistency in a single model*

A single model may yield inconsistent outputs owing to unexplained variability in the results of individual studies informing the pooled estimates of input variables. For instance, when developing a health economic model, a systematic review may yield several credible, but discrepant, utility estimates in the population of interest. If there is no plausible explanation for that difference in utility estimates, outputs of a model based on those inputs may also be qualitatively inconsistent. Again, sensitivity analysis may help to make a judgment to what extent such inconsistency of model inputs would translate into a meaningful inconsistency in model outputs with respect to the decision problem at hand.

#### *Imprecision in a single model*

Sensitivity analysis characterizes the response of model outputs to parameter variation and helps to determine the robustness of model's qualitative conclusions.<sup>105</sup> The overall certainty of model outputs may also be lower when estimated imprecisely. For quantitative outputs, one should examine not only the point estimate (e.g., average predicted event) but also the variability of that estimate (e.g., results of the probabilistic sensitivity analysis based in the distribution of the input parameters). It is essential that a report from a modelling study always includes information about output variability. Further guidance on how to assess imprecision in model outputs will need to take into account if the conclusions change in accordance with that specific parameter.

*Risk of publication bias in the context of a single model*

Risk of publication bias may not be relevant when assessing the certainty of outputs of a single model constructed de novo. However, when one intends to reuse an existing model but is aware or strongly suspects that similar models had been developed but are not available, then one may be inclined to think that their outputs might have systematically differed from the model that is available. In such a case, one may have lower confidence in the outputs of the identified model if there is no reasonable explanation for the inability to obtain those other models.

*Domains that increase the certainty of outputs from a single model*

Workshop participants agreed that presence of a dose—response gradient in model outputs may be applicable in some modelling disciplines (e.g., environmental health). Similarly, whether or not a large magnitude of an effect in model outputs increases the certainty of the evidence may depend on the modelling discipline. The effect of an opposite direction of plausible residual confounding seems theoretically also applicable in assessing the certainty of model outputs (i.e., a conservative model not incorporating input data parameter in favor of an intervention but still finding favorable outputs), but an actual example of this phenomenon in modelling studies is still under discussion.



# HHS Public Access

Author manuscript

*J Clin Epidemiol*. Author manuscript; available in PMC 2022 January 01.

Published in final edited form as:

*J Clin Epidemiol*. 2021 January ; 129: 138–150. doi:10.1016/j.jclinepi.2020.09.018.

## GRADE Guidelines 30: the GRADE approach to assessing the certainty of modeled evidence—An overview in the context of health decision-making

Jan L. Brozek<sup>a,b,c,\*</sup>, Carlos Canelo-Aybar<sup>d,e,1</sup>, Elie A. Akl<sup>f</sup>, James M. Bowen<sup>g</sup>, John Bucher<sup>h</sup>, Weihsueh A. Chiu<sup>i</sup>, Mark Cronin<sup>j</sup>, Benjamin Djulbegovic<sup>k</sup>, Maicon Falavigna<sup>l</sup>, Gordon H. Guyatt<sup>a,b,c</sup>, Ami A. Gordon<sup>m</sup>, Michele Hilton Boon<sup>n</sup>, Raymond C.W. Hutubessy<sup>o</sup>, Manuela A. Joore<sup>p</sup>, Vittal Katikireddi<sup>q</sup>, Judy LaKind<sup>r,s</sup>, Miranda Langendam<sup>t</sup>, Veena Manja<sup>u,v</sup>, Kristen Magnuson<sup>w</sup>, Alexander G. Mathioudakis<sup>x</sup>, Joerg Meerpohl<sup>y,z</sup>, Dominik Mertz<sup>a</sup>, Roman Mezencev<sup>y</sup>, Rebecca Morgan<sup>a</sup>, Gian Paolo Morgano<sup>a,c</sup>, Reem Mustafa<sup>a,z</sup>, Martin O'Flaherty<sup>aa</sup>, Grace Patlewicz<sup>ab</sup>, John J. Riva<sup>c,ac</sup>, Margarita Posso<sup>e</sup>, Andrew Rooney<sup>h</sup>, Paul M. Schlosser<sup>y</sup>, Lisa Schwartz<sup>a</sup>, Ian Shemilt<sup>ad</sup>, Jean-Eric Tarride<sup>ae</sup>, Kristina A. Thayer<sup>u</sup>, Katya Tsaion<sup>af</sup>, Luke Vale<sup>ag</sup>, John Wambough<sup>ab</sup>, Jessica Wignall<sup>m</sup>, Ashley Williams<sup>m</sup>, Feng Xie<sup>a</sup>, Yuan Zhang<sup>a,ah</sup>, Holger J. Schünemann<sup>a,b,c</sup>, GRADE Working Group

<sup>a</sup>Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, Ontario, Canada

<sup>b</sup>Department of Medicine, McMaster University, Hamilton, Ontario, Canada

<sup>c</sup>McMaster GRADE Centre & Michael DeGroote Cochrane Canada Centre, McMaster University, Hamilton, Ontario, Canada

<sup>d</sup>Department of Paediatrics, Obstetrics and Gynaecology, Preventive Medicine, and Public Health. PhD Programme in Methodology of Biomedical Research and Public Health. Universitat Autònoma de Barcelona, Bellaterra, Spain

<sup>e</sup>Iberoamerican Cochrane Center, Biomedical Research Institute (IIB Sant Pau-CIBERESP), Barcelona, Spain

<sup>f</sup>Department of Internal Medicine, American University of Beirut, Beirut, Lebanon

<sup>g</sup>Toronto Health Economics and Technology Assessment (THETA) Collaborative, Toronto, Ontario, Canada

<sup>h</sup>National Toxicology Program, National Institute of Environmental Health Sciences, Durham, NC, USA

\*Corresponding author: Jan Brozek, McMaster University, Health Sciences Centre, Area 2C, 1280 Main Street West, Hamilton, Ontario L8S 4K1, Canada.

<sup>1</sup>Co-first author.

Authors' contribution

All authors analyzed and interpreted the data. J.B. and C.C.-A. wrote the first version of the article. All authors of this article have read and approved the final version submitted.

Appendix A Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.jclinepi.2020.09.018>.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript



<sup>1</sup>Department of Veterinary Integrative Biosciences, Texas A&M University, College Station, TX, USA

<sup>2</sup>School of Pharmacy and Chemistry, Liverpool John Moores University, Liverpool, UK

<sup>3</sup>Center for Evidence-Based Medicine and Health Outcome Research, Morsani College of Medicine, University of South Florida, Tampa, Florida, USA

<sup>4</sup>Institute for Education and Research, Hospital Moinhos de Vento, Porto Alegre, Rio Grande do Sul, Brazil

<sup>5</sup>ICF International, Durham, NC, USA

<sup>6</sup>Institute of Health & Wellbeing, University of Glasgow, Glasgow, UK

<sup>7</sup>Department of Immunization, Vaccines and Biologicals, World Health Organization, Geneva, Switzerland

<sup>8</sup>Clinical Epidemiology and Medical Technology Assessment, Maastricht University Medical Centre+, Maastricht, the Netherlands

<sup>9</sup>LaKind Associates, LLC, Catonsville, MD, USA

<sup>10</sup>Department of Epidemiology and Public Health, University of Maryland School of Medicine, Baltimore, MD, USA

<sup>11</sup>Department of Clinical Epidemiology, Biostatistics, and Bioinformatics, Academic Medical Center, University of Amsterdam, Amsterdam, the Netherlands

<sup>12</sup>Department of Surgery, University of California Davis, Sacramento, CA, USA

<sup>13</sup>Department of Medicine, Department of Veterans Affairs, Northern California Health Care System, Mather, CA, USA

<sup>14</sup>Division of Infection, Immunity and Respiratory Medicine, University Hospital of South Manchester, University of Manchester, Manchester, UK

<sup>15</sup>Institute for Evidence in Medicine, Medical Center, University of Freiburg, Freiburg-am-Breisgau, Germany

<sup>16</sup>Cochrane Germany, Freiburg-am-Breisgau, Germany

<sup>17</sup>National Center for Environmental Assessment, U.S. Environmental Protection Agency, Washington, DC, USA

<sup>18</sup>Department of Medicine, University of Kansas Medical Center, Kansas City, KS, USA

<sup>19</sup>Institute of Population Health Sciences, University of Liverpool, Liverpool, UK

<sup>20</sup>National Center for Computational Toxicology, U.S. Environmental Protection Agency, Durham, NC, USA

<sup>21</sup>Department of Family Medicine, McMaster University, Hamilton, Ontario, Canada

<sup>22</sup>EPPI-Centre, Institute of Education, University College London, London, UK

<sup>23</sup>Programs for Assessment of Technology in Health, McMaster University, Hamilton, Ontario, Canada

<sup>af</sup> Evidence-Based Toxicology Collaboration, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA

<sup>ag</sup> Health Economics Group, Institute of Health and Society, Newcastle University, Newcastle upon Tyne, UK

<sup>ah</sup> Health Quality Ontario, Toronto, Ontario, Canada

## Abstract

**Objectives:** The objective of the study is to present the Grading of Recommendations Assessment, Development, and Evaluation (GRADE) conceptual approach to the assessment of certainty of evidence from modeling studies (i.e., certainty associated with model outputs).

**Study Design and Setting:** Expert consultations and an international multidisciplinary workshop informed development of a conceptual approach to assessing the certainty of evidence from models within the context of systematic reviews, health technology assessments, and health care decisions. The discussions also clarified selected concepts and terminology used in the GRADE approach and by the modeling community. Feedback from experts in a broad range of modeling and health care disciplines addressed the content validity of the approach.

**Results:** Workshop participants agreed that the domains determining the certainty of evidence previously identified in the GRADE approach (risk of bias, indirectness, inconsistency, imprecision, reporting bias, magnitude of an effect, dose—response relation, and the direction of residual confounding) also apply when assessing the certainty of evidence from models. The assessment depends on the nature of model inputs and the model itself and on whether one is evaluating evidence from a single model or multiple models. We propose a framework for selecting the best available evidence from models: 1) developing de novo, a model specific to the situation of interest, 2) identifying an existing model, the outputs of which provide the highest certainty evidence for the situation of interest, either “off-the-shelf” or after adaptation, and 3) using outputs from multiple models. We also present a summary of preferred terminology to facilitate communication among modeling and health care disciplines.

**Conclusion:** This conceptual GRADE approach provides a framework for using evidence from models in health decision-making and the assessment of certainty of evidence from a model or models. The GRADE Working Group and the modeling community are currently developing the detailed methods and related guidance for assessing specific domains determining the certainty of evidence from models across health care—related disciplines (e.g., therapeutic decision-making, toxicology, environmental health, and health economics). © 2020 Published by Elsevier Inc.

## Keywords

GRADE; Certainty of evidence; Mathematical models; Modelling studies; Health care Decision making; Guidelines

## 1. Introduction

When direct evidence to inform health decisions is not available or not feasible to measure (e.g., long-term effects of interventions or when studies in certain populations are perceived



as unethical), modeling studies may be used to predict that “evidence” and inform decision-making [1,2]. Health decision makers arguably face many more questions than can be reasonably answered with studies that directly measure the outcomes. Modeling studies, therefore, are increasingly used to predict disease dynamics and burden, the likelihood that an exposure represents a health hazard, the impact of interventions on health benefits and harms, or the economic efficiency of health interventions, among others [1]. Irrespective of the modeling discipline, decision makers need to know the best estimates of the modeled outcomes and how much confidence they may have in each estimate [3]. Knowing to what extent one can trust the outputs of a model is necessary when using them to support health decisions [4].

Although a number of guidance documents on how to assess the trustworthiness of estimates obtained from models in several health fields have been previously published [5–16], they are limited by failing to distinguish methodological rigor from completeness of reporting and by failing to clearly distinguish among various components affecting the trustworthiness of model outputs. In particular, they lack clarity regarding sources of uncertainty that may arise from model inputs and from the uncertainty about a model itself. Modelers and those using results from models should assess the credibility of both [4].

Authors have attempted to develop tools to assess model credibility, but many addressed only selected aspects, such as statistical reproducibility of data, the quality of reporting [17], or a combination of reporting with aspects of good modeling practices [7,18–21]. Many tools also do not provide sufficiently detailed guidance on how to apply individual domains or criteria. There is therefore a need for further development and validation of such tools in specific disciplines. Sufficiently detailed guidance for making and reporting these assessments is also necessary.

Models predict outcomes based on model inputs—previous observations, knowledge, and assumptions about the situation being modeled. Thus, when developing new models or assessing whether an existing model has been optimally developed, one should specify a priori the most appropriate and relevant data sources to inform different parameters required for the model. These may be either (seldom) a single study that provides the most direct information for the situation being modeled or (more commonly) a systematic review of multiple studies that identify all relevant sources of data. The risk of bias, directness and consistency of input data, precision of these estimates, and other domains specified in the Grading of Recommendations Assessment, Development, and Evaluation (GRADE) approach determine the certainty of each of the model inputs [22–28].

When assessing the evidence generated, various disciplines in health care and related areas that use modeling face similar challenges and may benefit from shared solutions. Table 1 presents examples of selected models used in health-related disciplines. Building on the existing GRADE approach, we refined and expand guidance regarding assessment of the certainty of model outputs. We formed a GRADE project group which comprised individuals with expertise in developing models and using model results in health-related disciplines, to create a unified framework for assessing the certainty of model outputs in the context of systematic reviews [29], health technology assessments, health care guidelines,

and other health decision-making. In this article, we outline the proposed conceptual approach and clarify key terminology (Table 2). The target audience for this article includes researchers who develop models and those who use models to inform health care—related decisions.

### 1.1. What we mean by a model

Authors have used the term *model* to describe a variety of different concepts [2] and suggested several broader or narrower definitions [6,30], so even modelers in the relatively narrow context of health sciences can differ in their views regarding what constitutes a model. Models vary in their structure and degree of complexity. A very simple model might be an equation estimating a variable not directly measured, such as the absolute effect of an intervention estimated as the product of the intervention's relative effect and the assumed baseline risk in a defined population (risk difference equals relative risk reduction multiplied by an assumed baseline risk). On the other end of the spectrum, elaborate mathematical models, such as system dynamics models (e.g., infectious disease transmission) may contain dozens of sophisticated equations that require considerable computing power to solve.

By their nature, such models only *resemble* the phenomena being modeled—that is, specific parts of the world that are interesting in the context of a particular decision—with necessary approximations and simplifications and to the extent that one actually knows and understands the underlying mechanisms [1]. Given the complexity of the world, decision makers often rely on some sort of a model to answer health-related questions.

In this article, we focus on quantitative mathematical models defined as “mathematical framework representing variables and their interrelationships to describe observed phenomena or predict future events” [30] used in health-related disciplines for decision-making (Table 1). These may be models of systems representing causal mechanisms (aka mechanistic models), models predicting outcomes from input data (aka empirical models), and models combining mechanistic with empirical approaches (aka hybrid models). We do not consider here statistical models used to estimate the associations between measured variables (e.g., proportional hazards models or models used for meta-analysis).

### 1.2. The GRADE approach

The GRADE Working Group was established in the year 2000 and continues as a community of people striving to create systematic and transparent frameworks for assessing and communicating the certainty of the available evidence used in making decisions in health care—related and health-related disciplines [31]. The GRADE Working Group now includes over 600 active members from 40 countries and serves as a think tank for advancing evidence-based decision-making in multiple health-related disciplines ([www.gradeworkinggroup.org](http://www.gradeworkinggroup.org)). GRADE is widely used internationally by over 110 organizations to address topics related to clinical medicine, public health, coverage decisions, health policy, and environmental health.

The GRADE framework uses concepts familiar to health scientists, grouping specific items to evaluate the certainty of evidence in conceptually coherent domains. Specific approaches to the concepts may differ depending on the nature of the body of evidence (Table 2).

GRADE domains include concepts such as risk of bias [28], directness of information [24], precision of an estimate [23], consistency of estimates across studies [25], risk of bias related to selective reporting [26], strength of the association, presence of a dose—response gradient, and the presence of plausible residual confounding that can increase confidence in estimated effects [27].

The general GRADE approach is applicable irrespective of health discipline. It has been applied to rating the certainty of evidence for management interventions, health care—related tests and strategies [32,33], prognostic information [34], evidence from animal studies [35], use of resources and cost-effectiveness evaluations [36], and values and preferences [37,38]. Although the GRADE Working Group has begun to address certainty of modeled evidence in the context of test—treatment strategies [39], health care resource use and costs [36], and environmental health [40], more detailed guidance is needed for complex models such as those used in infectious diseases, health economics, public health, and decision analysis.

## 2. Methods

On May 15 and 16, 2017, health scientists participated in a GRADE modeling project group workshop in Hamilton, Ontario, Canada, to initiate a collaboration in developing common principles for the application of the GRADE assessment of certainty of evidence to modeled outputs. The National Toxicology Program of the Department of Health and Human Services in the United States of America and the MacGRADE Center in the Department of Health Research Methods, Evidence, and Impact at McMaster University sponsored the workshop which was co-organized by MacGRADE Center and ICF International.

Workshop participants were selected to ensure a broad representation of all modeling related fields (Appendix). Participants had expertise in modeling in the context of clinical practice guidelines, public health, environmental health, dose—response modeling, physiologically based pharmacokinetic (PBPK) modeling, environmental chemistry, physical/chemical property prediction, evidence integration, infectious disease, computational toxicology, exposure modeling, prognostic modeling, diagnostic modeling, cost-effectiveness modeling, biostatistics, and health ethics.

Leading up to the workshop, we held three webinars to introduce participants to the GRADE approach. Several workshop participants (VM, KT, JB, AR, JW, JLB, HJS) collected and summarized findings from literature and the survey of experts as background material that provided a starting point for discussion. The materials included collected terminology representing common concepts across multiple disciplines that relate to evaluating modeled evidence and a draft framework for evaluating modeled evidence. Participants addressed specific tasks in small groups and large group discussion sessions and agreed on key principles both during the workshop and through written documents.



### 3. Results

#### 3.1. Terminology

Workshop participants agreed on the importance of clarifying terminology to facilitate communication among modelers, researchers, and users of model outputs from different disciplines. Modeling approaches evolved somewhat independently, resulting in different terms being used to describe the same or very similar concepts or the same term being used to describe different concepts. For instance, the concept of extrapolating from the available data to the context of interest has been referred to as directness, applicability, generalizability, relevance, or external validity. The lack of standardized terminology leads to confusion and hinders effective communication and collaboration among modelers and users of models.

Overcoming these obstacles would require clarifying the definitions of concepts and agreeing on terminology across disciplines. Realizing that this involves changing established customary use of terms in several disciplines, workshop participants suggested accepting the use of alternative terminology while always being clear about the preferred terms to be used and the underlying concept to which it refers (Table 2). Experts attending a World Health Organization's consultation have very recently suggested a more extensive set of terms [41]. To facilitate future communication, participants of this workshop will further collaborate to build a comprehensive glossary of terminology related to modeling.

#### 3.2. Outline of an approach to using model outputs for decision-making

Workshop participants suggested an approach to incorporate model outputs in health-related decision-making (Figure 1). In this article, we describe only the general outline of the suggested approach; in subsequent articles, we will discuss the details of the approach and provide more specific guidance on its application to different disciplines and contexts.

Researchers should start by conceptualizing the problem and the ideal target model that would best represent the actual phenomenon or decision problem they are considering [13]. This conceptualization would either guide the development of a new model or serve as a reference against which existing models could be compared. The ideal target model should reflect the following: 1) the relevant population (e.g., patients receiving some diagnostic procedure or exposed to some hazardous substance), 2) the exposures or health interventions being considered, 3) the outcomes of interest in that context, and 4) their relationships [42]. Conceptualizing the model will also reduce the risk of intentional or unintentional development of data-driven models, in which inputs and structure would be determined only by what is feasible to develop given the available data at hand.

Participants identified three options in which users may incorporate model outputs in health decision-making (Figure 1):

1. Develop a model *de novo* designed specifically to answer the very question at hand. Workshop participants agreed that in an ideal situation, such an approach would almost always be the most appropriate. Following this approach, however, requires suitable skills, ample resources, and time being available. It also

requires enough knowledge about the phenomenon being modeled to be able to tell whether or not the new model would have any advantage over already existing models.

2. Search for an existing model describing the same or a very similar problem and use it “*off-the-shelf*” or adapt it appropriately to answer the current question. In practice, many researchers initially use this approach because of the aforementioned limitations of developing a new model. However, it is often not possible to find an existing model that would be directly relevant to the problem at hand and/or it is not feasible to adapt an existing model when found. Any adaptation of a model requires availability of input data relevant for current problem, appropriate expertise and resources, and access to the original model. The latter is often not available (e.g., proprietary model or no longer maintained) or the structure of the original model is not being transparent enough to allow adaptation (“black-box”).
3. Use the results from multiple existing models found in the literature [43]. This approach may be useful when a limited knowledge about the phenomenon being modeled makes it impossible to decide which of the available models are more relevant, or when many alternative models are relevant but use different input parameters. In such situations, one may be compelled to rely on the results of several models because selection of the single, seemingly “best” model may provide incorrect estimates of outputs and lead to incorrect decisions.

Identifying existing models that are similar to the ideal target model often requires performing a scoping of the literature or a complete systematic review of potentially relevant models—a structured process following a standardized set of methods with a goal to identify and assess all available models that are accessible, transparently reported, and fulfill the prespecified eligibility criteria based on the conceptual ideal target model. Some prefer the term systematic survey that differs from a systematic review in the initial intention to use the results: in systematic reviews, the initial intention is to combine the results across studies either statistically through a meta-analysis or narratively summarizing their results when appropriate, whereas in a systematic survey, the initial intention is to examine the various ways that an intervention or exposure has been modeled, to review the input evidence that has been used, and ultimately to identify a single model that fits the conceptual ideal target model the best or requires the least adaptation; only when one cannot identify a single such model will it be necessary to use the results of multiple existing models.

If a systematic search revealed one or more models meeting the eligibility criteria, then researchers would assess the certainty of outputs from each model. Depending on this assessment, researchers may be able to use the results of a single most direct and lowest risk of bias model “*off-the-shelf*” or proceed to adapt that model. If researchers failed to find an existing model that would be sufficiently direct and low risk of bias, then they would ideally develop their own model *de novo*.

### 3.3. Assessing the certainty of outputs from a single model

When researchers develop their own model or when they identify a single model that is considered sufficiently direct to the problem at hand, they should assess the certainty of its outputs (i.e., evidence generated from that model). Note that if a model estimates multiple outputs, researchers need to assess the certainty of each output separately [23–28]. Workshop participants agreed that all GRADE domains are applicable to assess the certainty of model outputs, but further work is needed to identify examples and develop specific criteria to be assessed, which may differ depending on the model being used and/or situation being modeled.

### 3.4. Risk of bias in a single model

The risk of bias of model outputs (i.e., model outputs being systematically overestimated or underestimated) is determined by the credibility of a model itself and the certainty of evidence for each of model inputs.

The credibility of a model, also referred to as the quality of a model (Table 2), is influenced by its conceptualization, structure, calibration, validation, and other factors. Determinants of model credibility are likely to be specific to a modeling discipline (e.g., health economic models have different determinants of their credibility than PBPK models). There are some discipline-specific guidelines or checklists developed for the assessment of credibility of a model and other factors affecting the certainty of model outputs such as the framework to assess adherence to good practice guidelines in decision-analytic modeling [18], the questionnaire to assess relevance and credibility of modeling studies [18,44,45], good research practices for modeling in health technology assessment [5,6,8,9,12–14], the approaches to assessing uncertainty in read-across [46], and the quantitative structure–activity relationships [47] in predictive toxicology. Workshop participants agreed that there is a need for comprehensive tools developed specifically to assess credibility of various types of models in different modeling disciplines.

The certainty of evidence in each of the model inputs is another critical determinant of the risk of bias in a model. A model has several types of input data—bodies of evidence used to populate a model (Table 2). When researchers develop their model *de novo*, to minimize the risk of bias, they need to specify those input parameters to which the model outputs are the most sensitive. For instance, in economic models, these key parameters may include health effects, resource use, utility values, and baseline risks of outcomes. Model inputs should reflect the entire body of relevant evidence satisfying clear prespecified criteria rather than an arbitrarily selected evidence that is based on convenience (“any available evidence”) or picked in any other nonsystematic way (e.g., “first evidence found”—single studies that researchers happen to know about or are the first hits in a database search).

The appropriate approach will depend on the type of data and may require performing a systematic review of evidence on each important or crucial input variable [48–50]. Some inputs may have very narrow inclusion criteria, and therefore, evidence from single epidemiological survey or population surveillance may provide all relevant data for the population of interest (e.g., baseline population incidence or prevalence).



The certainty of evidence for each input needs to be assessed following the established GRADE approach specific to that type of evidence (e.g., estimates of intervention effects or baseline risk of outcomes) [22,32,34,37]. Following the logic of the GRADE approach that the overall certainty of evidence cannot be higher than the lowest certainty for any body of evidence that is critical for a decision [51], the overall rating of certainty of evidence across model inputs should be limited by the lowest certainty rating for any body of evidence (in this case, input data) to which the model output(s) was proved sensitive.

Application of this approach requires a priori consideration of likely critical and/or important inputs when specifying the *conceptual ideal target model* and the examination of the results of *back-end* sensitivity analyses. It further requires deciding how to judge whether results are or are not sensitive to alternative input parameters. Authors have described several methods to identify the most influential parameters including global sensitivity analysis to obtain “parameter importance measures” (i.e., information-based measures) [52] or alternatively by varying one parameter at a time and assessing their influence in “base case” outputs [52]. For example, in a model-based economic evaluation, one might be looking for the influence of sensitivity analysis on cost-effectiveness ratios at a specified willingness-to-pay threshold.

### 3.5. Indirectness in a single model

By directness or relevance, we mean the extent to which model outputs directly represent the phenomenon being modeled. To evaluate the relevance of a model, one needs to compare it against the conceptual ideal target model. When there are concerns about the directness of the model or there is limited understanding of the system being modeled making it difficult to assess directness, then one may have lower confidence in model outputs.

Determining the directness of model outputs includes assessing to what extent the modeled population, the assumed interventions and comparators, the time horizon, the analytic perspective, as well as the outcomes being modeled reflect those that are current interest. For instance, if the question is about the risk of birth defects in children of mothers chronically exposed to a certain substance, there may be concerns about the directness of the evidence if the model assumed short-term exposure, the route of exposure was different, or the effects of exposure to a similar but not the same substance were measured.

Assessing indirectness in a single model also requires evaluating two separate sources of indirectness:

1. Indirectness of input data with respect to the ideal target model's inputs.
2. Indirectness of model outputs with respect to the decision problem at hand.

This conceptual distinction is important because, although they are interrelated, one needs to address each type of indirectness separately. Even if the outputs might be direct to the problem of interest, the final assessment should consider if the inputs used were also direct for the target model.

Using an existing model has potential limitations: its inputs might have been direct for the decision problem addressed by its developers but are not direct with respect to the problem



currently at hand. In this context, sensitivity analysis can help to assess to what extent model outputs are robust to the changes in input data or assumptions used in model development.

### 3.6. Inconsistency in a single model

A single model may yield inconsistent outputs owing to unexplained variability in the results of individual studies informing the pooled estimates of input variables. For instance, when developing a health economic model, a systematic review may yield several credible, but discrepant, utility estimates in the population of interest. If there is no plausible explanation for that difference in utility estimates, outputs of a model based on those inputs may also be qualitatively inconsistent. Again, sensitivity analysis may help to make a judgment to what extent such inconsistency of model inputs would translate into a meaningful inconsistency in model outputs with respect to the decision problem at hand.

### 3.7. Imprecision in a single model

Sensitivity analysis characterizes the response of model outputs to parameter variation and helps to determine the robustness of model's qualitative conclusions [52,53]. The overall certainty of model outputs may also be lower when the outputs are estimated imprecisely. For quantitative outputs, one should examine not only the point estimate (e.g., average predicted event) but also the variability of that estimate (e.g., results of the probabilistic sensitivity analysis based in the distribution of the input parameters). It is essential that a report from a modeling study always includes information about output variability. Further guidance on how to assess imprecision in model outputs will need to take into account if the conclusions change in accordance with that specific parameter. In some disciplines, for instance in environmental health, model inputs are frequently qualitative. Users of such models may assess "adequacy" of the data, that is, the degree of "richness" and quantity of data supporting particular outputs of a model.

### 3.8. Risk of publication bias in the context of a single model

The risk of publication bias, also known as "reporting bias", "non-reporting bias", or "bias owing to missing results", as it is currently called in the Cochrane Handbook [54], is the likelihood that relevant models have been constructed but were not published or otherwise made publicly available. Risk of publication bias may not be relevant when assessing the certainty of outputs of a single model constructed de novo. However, when one intends to reuse an existing model but is aware or strongly suspects that similar models had been developed but are not available, then one may be inclined to think that their outputs might have systematically differed from the model that is available. In such a case, one may have lower confidence in the outputs of the identified model if there is no reasonable explanation for the inability to obtain those other models.

### 3.9. Domains that increase the certainty of outputs from a single model

The GRADE approach to rating the certainty of evidence recognized three situations when the certainty of evidence can increase: large magnitude of an estimated effect, presence of a dose—response gradient in an estimated effect, and an opposite direction of plausible residual confounding [27]. Workshop participants agreed that presence of a dose

—response gradient in model outputs may be applicable in some modeling disciplines (e.g., environmental health). Similarly, whether or not a large magnitude of an effect in model outputs increases the certainty of the evidence may depend on the modeling discipline. The effect of an opposite direction of plausible residual confounding seems theoretically also applicable in assessing the certainty of model outputs (i.e., a conservative model not incorporating input data parameter in favor of an intervention but still finding favorable outputs), but an actual example of this phenomenon in modeling studies is still under discussion.

### 3.10. Assessing the certainty of outputs across multiple models

Not infrequently, particularly in disciplines relying on mechanistic models, the current knowledge about the real system being modeled is very limited precluding the ability to determine which of the available existing models generate higher certainty outputs. Therefore, it may be necessary to rely on the results across multiple models. Other examples include using multiple models when no model was developed for the population directly of interest (e.g., the European breast cancer guidelines for screening and diagnosis relied on a systematic review of modeling studies that compared different mammography screening intervals [55]) or when multiple models of the same situation exist but vary in structure, complexity, and parameter choices (e.g., HIV Modelling Consortium compared several different mathematical models simulating the same antiretroviral therapy program and found that all models predicted that the program has the potential to reduce new HIV infections in the population [56]).

When researchers choose or are compelled to include outputs from several existing models, they should assess the certainty of outputs across all included models. This assessment may be more complex than for single models and single bodies of evidence. The feasibility of GRADE's guidance to judge the certainty of evidence lies in the availability of accepted methods for assessing most bodies of evidence from experimental to observational studies. However, the methods for systematic reviews of modeling studies are less well-established; some stages of the process are more complex, the number of highly skilled individuals with experience in such systematic reviews is far lower, and there is larger variability in the results [57]. In addition, researchers must be careful to avoid “double counting” the same model as if it were multiple models. For instance, the same model (i.e., same structure and assumptions) may have been used in several modeling studies, in which investigators relied on different inputs. When facing this scenario, researchers may need to decide which of the inputs are the most direct to their particular question and include in only this model in the review.

### 3.11. Risk of bias across multiple models

The assessment of risk of bias across models involves an assessment of the risk of bias in each individual model (see aforementioned discussion of risk of bias in single model) and subsequently making a judgment about the overall risk of bias across all included models. Specific methods for operationalizing this integration remain to be developed.

### 3.12. Indirectness across multiple models

As for the risk of bias, researchers need to assess indirectness of outputs initially for each of included models and then integrate the judgments across models. Likewise, specific methods for operationalizing this integration still remain to be developed. During this assessment, researchers may find some models too indirect to be informative for their current question and decide to exclude them from further consideration. However, the criteria to determine which models are too indirect should be developed a priori, before the search for the models is performed and their results are known.

### 3.13. Imprecision across multiple models

The overall certainty of model outputs may also be lower when model outputs are not estimated precisely. If researchers attempt a quantitative synthesis of outputs across models, they will report the range of estimates and variability of that estimates. When researchers choose to perform only a qualitative summary of the results across models, it is desirable that they report some estimate of variability in the outputs of individual models and an assessment of how severe the variability is (e.g., range of estimated effects).

### 3.14. Inconsistency of outputs across multiple models

The assessment of inconsistency should focus on unexplained differences across model outputs for a given outcome. If multiple existing models addressing the same issue produce considerably different outputs or reach contrasting conclusions, then careful comparison of the models may lead to a deeper understanding of the factors that drive outputs and conclusions. Ideally, the different modeling groups that developed relevant models would come together to explore the importance of differences in the type and structure of their models and of the data used as model inputs.

Invariably there will be some differences among the estimates from different models. Researchers will need to assess whether or not these differences are important, that is, whether they would lead to different conclusions. If the differences are important but can be explained by model structure, model inputs, the certainty of the evidence of the input parameters, or other relevant reasons, one may present the evidence separately for the relevant subgroups. If differences are important, but cannot be clearly explained, the certainty of model outputs may be lower.

### 3.15. Risk of publication bias across multiple models

The assessment is similar to that of the risk of publication bias in the context of a single model.

### 3.16. Domains that increase the certainty of outputs across multiple models

All considerations are the same to those in the context of a single model.

## 4. Discussion

The goal of the GRADE project group on modeling is to provide concepts and operationalization of how to rate the certainty of evidence in model outputs. This article



provides an overview of the conclusions of the project group. This work is important because there is a growing need and availability of modeled information resulting from a steadily increasing knowledge of the complexity of the structure and interactions in our environment and computational power to construct and run models. Users of evidence obtained from modeling studies need to know how much trust they may have in model outputs. There is a need to improve the methods of constructing models and to develop methods for assessing the certainty in model outputs. In this article, we have attempted to clarify the most important concepts related to developing and using model outputs to inform health-related decision-making. Our preliminary work identified confusion about terminology, lack of clarity of what is a model, and need for methods to assess certainty in model outputs as priorities to be addressed to improve the use of evidence from modeling studies.

In some situations, decision makers might be better off developing a new model specifically designed to answer their current question. However, we suggest that it is not always feasible to develop a new model or that developing a new model might not be any better than using already existing models, when the knowledge of the real life system to be modeled is limited precluding the ability to choose one model that would be better than any other. Thus, sometimes it may be necessary or more appropriate to use one or multiple existing models depending on their availability, credibility, and relevance to the decision-making context. The assessment of the certainty of model outputs will be conceptually similar when a new model is constructed, or one existing model is used. The main difference between the latter two approaches is the availability of information to perform a detailed assessment. That is, information for one's own model may be easily accessible, but information required to assess someone else's model will often be more difficult to obtain. Assessment of the certainty evidence across models can build on existing GRADE domains but requires different operationalization.

Because it builds on an existing, widely used framework that includes a systematic and transparent evaluation process, modeling disciplines' adoption of the GRADE approach and further development of methods to assess the certainty of model outputs may be beneficial for health decision-making. Systematic approaches improve rigor of research, reducing the risk of error and its potential consequences; transparency of the approach increases its trustworthiness. There may be additional benefits related to other aspects of the broader GRADE approach, for instance, a potential to reduce unnecessary complexity and workload in modeling by careful consideration of the most direct evidence as model inputs. This may allow, for instance, optimization of the use of different streams of evidence as model inputs. Frequently, authors introduce unnecessary complexity by considering multiple measures of the same outcome when focus could be on the most direct outcome measure.

The GRADE Working Group will continue developing methods and guidance for using model outputs in health-related decision-making. In subsequent articles, we will provide more detailed guidance about choosing the "best" model when multiple models are found, using multiple models, integrating the certainty of evidence from various bodies of evidence with credibility of the model and arriving at the overall certainty in model outputs, how to assess the credibility of various types of models themselves, and further clarification

of terminology. In the future, we aim to develop and publish the detailed guidance for assessing certainty of evidence from models, the specific guidance for the use of modeling across health care—related disciplines (e.g., toxicology, environmental health, or health economics), validation of the approach, and accompanying training materials and examples.

## Acknowledgments

A.R. was supported by the National Institutes of Health, National Institute of Environmental Health Sciences.

## Appendix. List of workshop participants

Elie Akl (EA)— American University of Beirut, Lebanon  
 Jim Bowen (JMB)— McMaster University, Canada  
 Chris Brinkerhoff (CB)— US Environmental Protection Agency, USA  
 Jan Brozek (JLB)— McMaster University, Canada  
 John Bucher (JB)— US National Toxicology Program, USA  
 Carlos Canelo-Aybar (CCA)— Iberoamerican Cochrane Centre, Spain  
 Marcy Card (MC)— US Environmental Protection Agency, USA  
 Weihsueh A. Chiu (WCh)— Texas A&M University, USA  
 Mark Cronin (MC)— Liverpool John Moores University, UK  
 Tahira Devji (TD)— McMaster University, Canada  
 Ben Djulbegovic (BD)— University of South Florida, USA  
 Ken Eng (KE)— Public Health Agency of Canada  
 Gerald Gartlehner (GG)— Donau-Universität Krems, Austria  
 Gordon Guyatt (GGu)— McMaster University, Canada  
 Raymond Huxhessy (RH)— World Health Organization Initiative for Vaccine Research, Switzerland  
 Manuela Joore (MJ)— Maastricht University, the Netherlands  
 Richard Judson (RJ)— US Environmental Protection Agency, USA  
 S. Vittal Katikireddi (SK)— University of Glasgow, UK  
 Nicole Kleinstreuer (NK)— US National Toxicology Program, USA  
 Judy LaKind (JL)— University of Maryland, USA

Miranda Langendam (ML)– University of Amsterdam, the Netherlands  
 Zbyszek Leś (ZL)– Evidence Prime Inc., Canada  
 Veena Manja (VM)– McMaster University, Canada  
 Joerg Meerpohl (JM)– GRADE Center Freiburg, Cochrane Germany, University Medical Center Freiburg  
 Dominik Mertz (DM)– McMaster University, Canada  
 Roman Mezenцев (RM)– US Environmental Protection Agency, USA  
 Rebecca Morgan (RMO)– McMaster University, Canada  
 Gian Paolo Morgano (GPM)– McMaster University, Canada  
 Reem Mustafa (RMu)– University of Kansas, USA  
 Bhash Naidoo (BN)– National Institute for Health and Clinical Excellence, UK  
 Martin O'Flaherty (MO)– Public Health and Policy, University of Liverpool, UK  
 Grace Padlewicz (GP)– US Environmental Protection Agency, USA  
 John Riva (JR)– McMaster University, Canada  
 Alan Sasso (AS)– US Environmental Protection Agency, USA  
 Paul Schlosser (PS)– US Environmental Protection Agency, USA  
 Holger Schünemann (HS)– McMaster University, Canada  
 Lisa Schwartz (LS)– McMaster University, Canada  
 Ian Shemilt (IS)– University College London, UK  
 Marek Smieja (MS)– McMaster University, Canada  
 Ravi Subramaniam (RS)– US Environmental Protection Agency, USA  
 Jean-Eric Tarride (JT)– McMaster University, Canada  
 Kris Thayer (KAT)– US Environmental Protection Agency, USA  
 Katya Tsaioun (KT)– John Hopkins University, USA  
 Bernhard Uitsch (BU)– Robert Koch Institute, Germany  
 John Wambough (JW)– US Environmental Protection Agency, USA  
 Jessica Wignall (JWi)– ICF, USA



Ashley Williams (AW)– ICF, USA

Feng Xie (FX)– McMaster University, Canada

## References

- [1] Oreskes N. The role of quantitative models in science. In: Canham CD, Cole JJ, Lauenroth WK, editors. *Models in ecosystem science*. Princeton, NJ: Princeton University Press; 2003:13–31.
- [2] Frigg R, Hartmann S. Models in science. In: Zalta EN, editor. *The Stanford Encyclopedia of Philosophy*. Stanford, CA: Spring 2017 Edition; 2017.
- [3] Guyatt GH, Oxman AD, Kunz R, Vist GE, Falck-Ytter Y, Schunemann HJ, et al. What is “quality of evidence” and why is it important to clinicians? *BMJ* 2008;336:995–8. [PubMed: 18456631]
- [4] Oreskes N. Evaluation (not validation) of quantitative models. *Environ Health Perspect* 1998;106(Suppl 6):1453–60. [PubMed: 9860904]
- [5] Briggs AH, Weinstein MC, Fenwick EA, Karon J, Sculpher MJ, Paltiel AD, et al. Model parameter estimation and uncertainty: a report of the ISPOR-SMDM modeling good research practices task force-6. *Value Health* 2012;15:835–42. [PubMed: 22999133]
- [6] Caro JJ, Briggs AH, Siebert U, Kuntz KM, Force I-SMGRPT. Modeling good research practices—overview: a report of the ISPOR-SMDM modeling good research practices task force-1. *Med Decis Making* 2012;32:667–77. [PubMed: 22990082]
- [7] Caro JJ, Eddy DM, Kan H, Kaitz C, Patel B, Eldessouki R, et al. Questionnaire to assess relevance and credibility of modeling studies for informing health care decision making: an ISPOR-AMCP-NPC Good Practice Task Force report. *Value Health* 2014;17:174–82. [PubMed: 24636375]
- [8] Eddy DM, Hollingworth W, Caro JJ, Tsevat J, McDonald KM, Wong JB, et al. Model transparency and validation: a report of the ISPOR-SMDM modeling good research practices task force-7. *Med Decis Making* 2012;32:733–43. [PubMed: 22990088]
- [9] Karon J, Stahl J, Brennan A, Caro JJ, Mar J, Moller J. Modeling using discrete event simulation: a report of the ISPOR-SMDM modeling good research practices task force-4. *Med Decis Making* 2012;32:701–11. [PubMed: 22990085]
- [10] Marshall DA, Burgos-Liz L, MJ LJ, Crown W, Padula WV, Wong PK, et al. Selecting a dynamic simulation modeling method for health care delivery research-part 2: report of the ISPOR Dynamic Simulation Modeling Emerging Good Practices Task Force. *Value Health* 2015;18:147–60. [PubMed: 25773550]
- [11] Marshall DA, Burgos-Liz L, MJ LJ, Osgood ND, Padula WV, Higashi MK, et al. Applying dynamic simulation modeling methods in health care delivery research-the SIMULATE checklist: report of the ISPOR simulation modeling emerging good practices task force. *Value Health* 2015;18:5–16. [PubMed: 25595229]
- [12] Pitman R, Fisman D, Zaric GS, Postma M, Kretzschmar M, Edmunds J, et al. Dynamic transmission modeling: a report of the ISPOR-SMDM modeling good research practices task force working group-5. *Med Decis Making* 2012;32:712–21. [PubMed: 22990086]
- [13] Roberts M, Russell LB, Paltiel AD, Chambers M, McEwan P, Krahn M, et al. Conceptualizing a model: a report of the ISPOR-SMDM modeling good research practices task force-2. *Med Decis Making* 2012;32:678–89. [PubMed: 22990083]
- [14] Siebert U, Alagoz O, Bayoumi AM, Jahn B, Owens DK, Cohen DJ, et al. State-transition modeling: a report of the ISPOR-SMDM modeling good research practices task force-3. *Med Decis Making* 2012;32:690–700. [PubMed: 22990084]
- [15] Vemer P, van Vroom GA, Ramos IC, Krabbe PF, Al MJ, Feenstra TL. Improving model validation in health technology assessment: comments on guidelines of the ISPOR-SMDM modeling good research practices task force. *Value Health* 2013;16:1106–7. [PubMed: 24041364]
- [16] Weinstein MC, O'Brien B, Hornberger J, Jackson J, Johannesson M, McCabe C, et al. Principles of good practice for decision analytic modeling in health-care evaluation: report of the ISPOR Task Force on Good Research Practices—Modeling Studies. *Value Health* 2003;6:9–17. [PubMed: 12535234]



- [17]. Bennett C, Manuel DG. Reporting guidelines for modelling studies. *BMC Med Res Methodol* 2012;12:168. [PubMed: 23134698]
- [18]. naloza Ramos MC, Barton P, Jowett S, Sutton AJ. A systematic review of research guidelines in decision-analytic modeling. *Value Health* 2015;18:512–29. [PubMed: 26091606]
- [19]. Philips Z, Bojke L, Sculpher M, Claxton K, Golder S. Good practice guidelines for decision-analytic modelling in health technology assessment: a review and consolidation of quality assessment. *Pharmacoeconomics* 2006;24:355–71. [PubMed: 16605282]
- [20]. LaKind JS, O'Mahony C, Armstrong T, Tibaldi R, Blount BC, Naiman DQ. ExpoQual: evaluating measured and modeled human exposure data. *Environ Res* 2019;171:302–12. [PubMed: 30708234]
- [21]. Husereau D, Drummond M, Petrou S, Carswell C, Moher D, Greenberg D, et al. Consolidated health economic evaluation reporting standards (CHEERS) statement. *BMJ* 2013;346:f1049. [PubMed: 23529982]
- [22]. Balshem H, Helfand M, Schunemann HJ, Ozman AD, Kunz R, Brozek J, et al. GRADE guidelines: 3. Rating the quality of evidence. *J Clin Epidemiol* 2011;64:401–6. [PubMed: 21208779]
- [23]. Guyatt GH, Ozman 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;64:1283–93. [PubMed: 21839614]
- [24]. Guyatt GH, Ozman AD, Kunz R, Woodcock J, Brozek J, Helfand M, et al. GRADE guidelines: 8. Rating the quality of evidence—indirectness. *J Clin Epidemiol* 2011;64:1303–10. [PubMed: 21802903]
- [25]. Guyatt GH, Ozman AD, Kunz R, Woodcock J, Brozek J, Helfand M, et al. GRADE guidelines: 7. Rating the quality of evidence—inconsistency. *J Clin Epidemiol* 2011;64:1294–302. [PubMed: 21803546]
- [26]. Guyatt GH, Ozman 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;64:1277–82. [PubMed: 21802904]
- [27]. Guyatt GH, Ozman 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;64:1311–6. [PubMed: 21802902]
- [28]. Guyatt GH, Ozman 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;64:407–15. [PubMed: 21247734]
- [29]. Cumpston M, Chandler J, Thomas J, Higgins JPT, Deeks JJ, Clarke MJ. Chapter 1: Introduction. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al., editors. *Cochrane Handbook for Systematic Reviews of Interventions*. 6.1. Cochrane; 2020. <https://training.cochrane.org/handbook/current/chapter-1>. Accessed October 13, 2020.
- [30]. Eyrkhoff P. System identification: parameter and state estimation. London: Wiley-Interscience; 1974.
- [31]. Schunemann HJ, Best D, Vist G, Ozman AD, Group GW. Letters, numbers, symbols and words: how to communicate grades of evidence and recommendations. *CMAJ* 2003;169:677–80. [PubMed: 14517128]
- [32]. Schunemann HJ, Mustafa R, Brozek J, Santesso N, Alonso-Coello P, Guyatt G, et al. GRADE Guidelines: 16. GRADE evidence to decision frameworks for tests in clinical practice and public health. *J Clin Epidemiol* 2016;76:89–98. [PubMed: 26931285]
- [33]. Schunemann HJ, Ozman AD, Brozek J, Glasziou P, Jaeschke R, Vist GE, et al. Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. *BMJ* 2008;336:1106–10. [PubMed: 18483053]
- [34]. Iorio A, Spencer FA, Falavigna M, Alon C, Lang E, Burnand B, et al. Use of GRADE for assessment of evidence about prognosis: rating confidence in estimates of event rates in broad categories of patients. *BMJ* 2015;350:h870. [PubMed: 25775931]

- [35]. Hooijmans CR, de Vries RBM, Ritskes-Hoitinga M, Rovers MM, Leeflang MM, Int'Hout J, et al. Facilitating healthcare decisions by assessing the certainty in the evidence from preclinical animal studies. *PLoS One* 2018;13:e0187271.
- [36]. Brunetti M, Shemilt I, Pregno S, Vale L, Ozman AD, Lord J, et al. GRADE guidelines: 10. Considering resource use and rating the quality of economic evidence. *J Clin Epidemiol* 2013;66:140–50. [PubMed: 22863410]
- [37]. Zhang Y, Alonso-Coello P, Guyatt GH, Yepes-Nunez JJ, Aki EA, Hazlewood G, et al. GRADE Guidelines: 19. Assessing the certainty of evidence in the importance of outcomes or values and preferences-Risk of bias and indirectness. *J Clin Epidemiol* 2018.
- [38]. Zhang Y, Coello PA, Guyatt GH, Yepes-Nunez JJ, Aki EA, Hazlewood G, et al. GRADE guidelines: 20. Assessing the certainty of evidence in the importance of outcomes or values and preferences-inconsistency, imprecision, and other domains. *J Clin Epidemiol* 2018.
- [39]. World Health Organization. WHO guidelines for screening and treatment of precancerous lesions for cervical cancer prevention. Geneva, Switzerland: World Health Organization; 2013.
- [40]. Thayer KA, Schunemann HJ. Using GRADE to respond to health questions with different levels of urgency. *Environ Int* 2016;92–93: 585–9.
- [41]. Porgo TV, Norris SL, Salanti G, Johnson LF, Simpson JA, Low N, et al. The use of mathematical modeling studies for evidence synthesis and guideline development: a glossary. *Res Synth Methods* 2019; 10:125–33. [PubMed: 30508309]
- [42]. National Institute for Health and Care Excellence. Chapter 5: The reference case. Guide to the methods of technology appraisal 2013. NICE; 2013. <https://www.nice.org.uk/process/pmg9/>. Accessed October 13, 2020.
- [43]. Eyles H, Ni Mhurchu C, Nghiem N, Blakely T. Food pricing strategies, population diets, and non-communicable disease: a systematic review of simulation studies. *PLoS Med* 2012;9:e1001353.
- [44]. Jaime Caro J, Eddy DM, Kam H, Kaltz C, Patel B, Eldessouki R, et al. Questionnaire to assess relevance and credibility of modeling studies for informing health care decision making: an ISPOR-AMCP-NPC Good Practice Task Force report. *Value Health* 2014;17:174–82. [PubMed: 24636375]
- [45]. National Institute for Health and Care Excellence. Appendix G: Methodology checklist: economic evaluations. The guidelines manual: Process and methods. NICE; 2012. <https://www.nice.org.uk/process/pmg6/>. Accessed October 13, 2020.
- [46]. Schultz TW, Richarz A-N, Cronin MTD. Assessing uncertainty in read-across: questions to evaluate toxicity predictions based on knowledge gained from case studies. *Comput Toxicol* 2019;9:1–11.
- [47]. Cronin MTD, Richarz AN, Schultz TW. Identification and description of the uncertainty, variability, bias and influence in quantitative structure-activity relationships (QSARs) for toxicity prediction. *Regul Toxicol Pharmacol* 2019;106:90–104. [PubMed: 31026540]
- [48]. 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;22:267–75. [PubMed: 30832964]
- [49]. Kaltenthaler E, Tappenden P, Praisley S, Squires H. NICE DSU Technical Support Document 13: Identifying and Reviewing Evidence to Inform the Conceptualisation and Population of Cost-Effectiveness Models. London, UK: National Institute for Health and Care Excellence; 2011.
- [50]. Praisley S. Identification of evidence for key parameters in decision-analytic models of cost effectiveness: a description of sources and a recommended minimum search requirement. *Pharmacoeconomics* 2016;34:597–608. [PubMed: 26861793]
- [51]. Guyatt G, Ozman AD, Sultan S, Brozek J, Glasziou P, Alonso-Coello P, et al. GRADE guidelines: 11. Making an overall rating of confidence in effect estimates for a single outcome and for all outcomes. *J Clin Epidemiol* 2013;66:151–7. [PubMed: 22542023]
- [52]. Bilcke J, Beutels P, Brisson M, Jit M. Accounting for methodological, structural, and parameter uncertainty in decision-analytic models: a practical guide. *Med Decis Making* 2011;31:675–92. [PubMed: 21653805]

- [53]. Saltelli A, Ratto M, Andres T, Campolongo F, Cariboni J, Gatelli D, et al. Global sensitivity analysis. The primer. Chichester, England: John Wiley & Sons; 2008.
- [54]. Page MJ, Higgins JPT, Sterne JAC. Chapter 13: Assessing risk of bias due to missing results in a synthesis. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al., editors. *Cochrane Handbook for Systematic Reviews of Interventions*. 6.1 Cochrane; 2019. <https://training.cochrane.org/handbook/current/chapter-13>. Accessed October 13, 2020.
- [55]. Schünemann HJ, Lerda D, Quinn C, Follmann M, Alonso-Coello P, Rossi PG, et al. Breast cancer screening and Diagnosis: a synopsis of the European Breast guidelines. *Ann Intern Med* 2020;172:46–56. [PubMed: 31766052]
- [56]. Eaton JW, Johnson LF, Salomon JA, Barnighausen T, Bendavid E, Bershteyn A, et al. HIV treatment as prevention: systematic comparison of mathematical models of the potential impact of antiretroviral therapy on HIV incidence in South Africa. *Plos Med* 2012;9: e1001245.
- [57]. Gomersall JS, Jadotte YT, Xue Y, Lockwood S, Riddle D, Preda A. Conducting systematic reviews of economic evaluations. *Int J Evid Based Healthc* 2015;13:170–8. [PubMed: 26288063]
- [58]. Mandelblatt JS, Stout NK, Schechter CB, van den Broek JJ, Miglioretti DL, Krapcho M, et al. Collaborative modeling of the benefits and harms associated with different U.S. Breast cancer screening strategies. *Ann Intern Med* 2016;164:215–25. [PubMed: 26756606]
- [59]. Davies NG, Kucharski AJ, Eggo RM, Gimma A, Edmunds WJ. Centre for the Mathematical Modelling of Infectious Diseases C-wg. Effects of non-pharmaceutical interventions on COVID-19 cases, deaths, and demand for hospital services in the UK: a modelling study. *Lancet Public Health* 2020;5:e375–85.
- [60]. Tibaldi R, ten Berge W, Drolet D. Dermal absorption of chemicals: estimation by IH SkinPerm. *J Occup Environ Hyg* 2014;11:19–31. [PubMed: 24283333]
- [61]. Young BM, Tulve NS, Egeghy PP, Driver JH, Zartarian VG, Johnston JE, et al. Comparison of four probabilistic models (CARES(R)), Calendex, ConsExpo, and SHEDS) to estimate aggregate residential exposures to pesticides. *J Expo Sci Environ Epidemiol* 2012;22:522–32. [PubMed: 22781436]
- [62]. United States Environmental Protection Agency. Human Exposure Modeling - Overview. <https://www.epa.gov/fera/human-exposure-modeling-overview>. Accessed October 13, 2020.
- [63]. Levin S, Dugas A, Gurses A, Kirsch T, Kelen G, Hinson J, et al. Hopscore: An Electronic Outcomes-Based Emergency Triage System. Agency for Healthcare Research and Quality; 2018. <https://digital.ahrq.gov/sites/default/files/docs/citation/r21hs023641-levin-final-report-2018.pdf>. Accessed October 13, 2020.
- [64]. Smith RD, Keogh-Brown MR, Barnett T, Tait J. The economy-wide impact of pandemic influenza on the UK: a computable general equilibrium modelling experiment. *BMJ* 2009;339:b4571. [PubMed: 19926697]
- [65]. Hultcrantz M, Rind D, Akl EA, Treweek S, Mustafa RA, Iorio A, et al. The GRADE Working Group clarifies the construct of certainty of evidence. *J Clin Epidemiol* 2017;87:4–13. [PubMed: 28529184]



**What is new?****Key findings:**

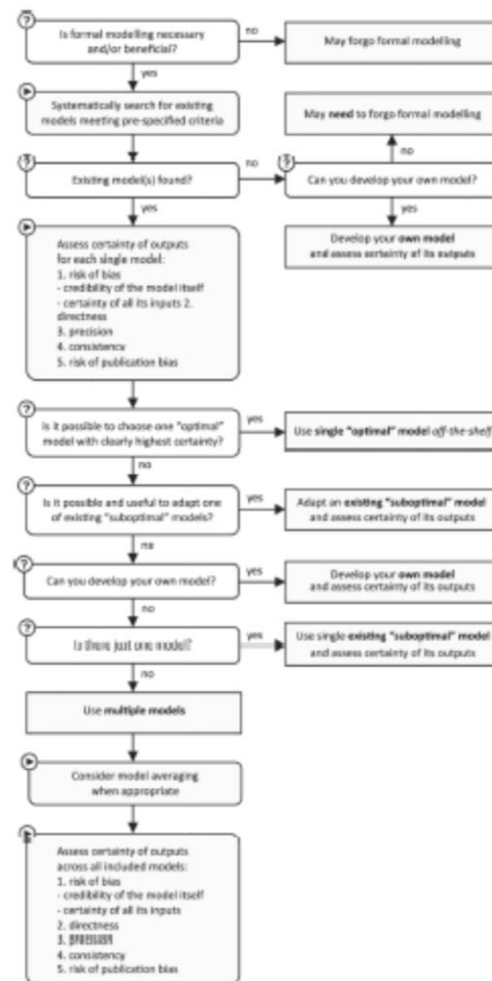
- General concepts determining the certainty of evidence in the GRADE approach (risk of bias, indirectness, inconsistency, imprecision, reporting bias, magnitude of an effect, dose—response relation, and the direction of residual confounding) also apply in the context of assessing the certainty of evidence from models (model outputs).
- Detailed assessment of the certainty of evidence from models differs for the assessment of outputs from a single model compared with the assessment of outputs across multiple models.

**What this adds to what was known?**

- We propose a framework for selecting the best available evidence from models to inform health care decisions: to develop a model de novo, to identify an existing model, the outputs of which provide the highest certainty evidence, or to use outputs from multiple models.

**What is the implication and what should change now?**

- We suggest that the modeling and health care decision-making communities collaborate further to clarify terminology used in the context of modeling and make it consistent across the disciplines to facilitate communication.



**Fig 1.**  
The general approach to using modeled evidence and assessing its certainty in health-related disciplines.

Table 1.

Examples of modeling methods in health-related disciplines (not comprehensive)<sup>a</sup>

Decision analysis models	Structured model representing health care pathways examining effects of an intervention on outcomes of interest.
	<p>Types</p> <ul style="list-style-type: none"> <li>■ Decision tree models</li> <li>■ State transition models               <ul style="list-style-type: none"> <li>○ Markov cohort simulation</li> <li>○ Individual-based microsimulation (first-order Monte Carlo)</li> </ul> </li> <li>■ Discrete event simulation</li> <li>■ Dynamic transmission models</li> <li>■ Agent-based models</li> </ul> <p>Examples</p> <ul style="list-style-type: none"> <li>■ Estimation of long-term benefits and harm outcomes from complex intervention, e.g., minimum unit pricing of alcohol</li> <li>■ Estimation of benefits and harms of population mammography screening based in the microsimulation model, e.g., Wisconsin model from CISNET collaboration [58]</li> <li>■ Susceptible-Infectious-Recovery transmission dynamic model to assess effectiveness of lockdown during the SARS-CoV-2 pandemic [59]</li> </ul>
Pharmacology and toxicology models	<p>Computational models developed to organize, analyze, simulate, visualize, or predict toxicological and ecotoxicological effects of chemicals. In some cases, these models are used to estimate the toxicity of a substance even before it has been synthesized.</p> <p>Types</p> <ul style="list-style-type: none"> <li>■ Structural alerts and rule-based models</li> <li>■ Read-across</li> <li>■ Dose response and time response</li> <li>■ Toxicokinetic (TK) and toxicodynamic(TD)</li> <li>■ Uncertainty factors</li> <li>■ Quantitative structure activity relationship (QSAR)</li> <li>■ Biomarker-based toxicity models</li> </ul> <p>Examples</p> <ul style="list-style-type: none"> <li>● Structural alerts for mutagenicity and skin sensitization</li> <li>● Read-across for complex endpoints such as chronic toxicity</li> <li>● Pharmacokinetic (PK) models to calculate concentrations of substances in organs, after a variety of exposures and QSAR models for carcinogenicity</li> <li>● TGx-DDI biomarker to detect DNA damage-inducing agents</li> </ul>
Environmental models	<p>The EPA defined these models as 'A simplification of reality that is constructed to gain insights into select attributes of a physical, biological, economic, or social system.' It involves the application of multidisciplinary knowledge to explain, explore, and predict the Earth's response to environmental change and the interactions between human activities and natural processes.</p> <p>Classification (based on the CREM guidance document):</p> <ul style="list-style-type: none"> <li>● Human activity models</li> <li>● Natural system process</li> <li>● Emission models</li> <li>● Fate and transport models</li> <li>● Exposure models</li> <li>● Human health effect models</li> </ul>

*J Clin Epidemiol*. Author manuscript; available in PMC 2022 January 01.

Author Manuscript

- Ecological effect models
  - Economic impact models
  - Noneconomic impact models
- Examples
- Land use regression models
  - IH Skin Pains [60]
  - CostExpo [61]
  - Other exposure models [62]

- Other
- HojScore: An Electronic Outcomes-Based Emergency Triage System [63]
  - Computational general equilibrium (CGE) models [64]

<sup>a</sup>Although not described in this classification, simple calculations incorporating two or more pieces of evidence, as for example, the multiplication of an RR by the baseline risk to obtain the absolute risk difference of an intervention are models, although pragmatic, with their respective assumptions.

Author Manuscript

Author Manuscript

Author Manuscript



Table 2.

Selected commonly used and potentially confusing terms used in the context of modeling and the GRADE approach\*

Term	General definition
Sources of evidence (may come from in vitro or in vivo experiment or a mathematical model)	
Streams of evidence	Parallel information about the same outcome that may have been obtained using different methods of estimating that outcome. For instance, evidence of the increased risk for developing lung cancer in humans after an exposure to certain chemical compound may come from several streams of evidence: 1) mechanistic evidence—models of physiological mechanisms, 2) studies in animals—observations and experiments in animals from different phyla, classes, orders, families, genera, and species (e.g., bacteria, nematodes, insects, fish, mice, rats), and 3) studies in humans.
Bodies of evidence	Information about multiple different aspects around a decision about the best course of action. For instance, to decide whether or not a given diagnostic test should be used in some people, one needs to integrate the bodies of evidence about the accuracy of the test, the prevalence of the conditions being suspected, the natural history of these conditions, the effects of potential treatments, values and preferences of affected individuals, cost, feasibility, etc.
Quality (may refer to many concepts, thus alternative terms are preferred to reduce confusion)	
Certainty of model outputs	
Alternative terms:	In the context of health decision-making, the certainty of evidence (term preferred over "quality" to avoid confusion with the risk of bias in an individual study) reflects the extent to which one's confidence in an estimate of an effect is adequate to make a decision or a recommendation. Decisions are influenced not only by the best estimates of the expected desirable and undesirable consequences but also by one's confidence in these estimates. In the context of evidence syntheses of separate bodies of evidence (e.g., systematic reviews), the certainty of evidence reflects the extent of confidence that an estimate of effect is correct. For instance, the attributable national risk of cardiovascular mortality resulting from exposure to air pollution measured in selected cities.
■ certainty of modeled evidence	
■ quality of evidence	
■ quality of model output	
■ strength of evidence	
■ confidence in model outputs	
	The GRADE Working Group published several articles explaining the concept in detail [22–28,65]. Note that the phrase "confidence in an estimate of an effect" does not refer to statistical confidence intervals. Certainty of evidence is always assessed for the whole body of evidence rather than on a single-study level (single studies are assessed for risk of bias and indirectness).
Certainty of model inputs	
Alternative term:	Characteristics of data that are used to develop, train, or run the model, e.g., source of input values, their manipulation before input into a model, quality control, risk of bias in data, etc.
■ quality of model inputs	
Credibility of a model	
Alternative terms:	To avoid confusion and keep with terminology used by modeling community [7], we suggest using the term <i>credibility</i> rather than <i>quality</i> of a model. The concept refers to the characteristics of a model itself—its design or execution—that affect the risk that the results may overestimate or underestimate the true effect. Various factors influence the overall credibility of a model, such as its structure, the analysis, and the validation of the assumptions made during modeling.
■ quality of a model	
■ risk of bias in a model	
■ validity of a model	
Quality of reporting	Refers to how comprehensively and clearly model inputs, a model itself, and model outputs have been documented and described such that they can be critically evaluated and used for decision-making. Quality of reporting and quality of a model are separate concepts: a model with a low quality of reporting is not necessarily a low-quality model and vice versa.
Directness	
Directness of a model	
Alternative terms:	By directness of a model, we mean the extent to which the model represents the real-life situation being modeled which is dependent on how well the input data and the model structure reflect the scenario of interest.
■ relevance	Directness is the term used in the GRADE approach because each of the alternatives has been used usually in a narrower meaning.
■ external validity	

*J Clin Epidemiol.* Author manuscript; available in PMC 2022 January 01.

Term	General definition
■ applicability	
■ generalizability	
■ transferability	
■ translatability	

\* There may be either subtle or fundamental differences among some disciplines in how these terms are being used, for the purposes of this article, these terms are generalized rather than discipline specific. GRADE, Grading of Recommendations Assessment, Development, and Evaluation.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

---

## **DISCUSSION**

---

## 7. DISCUSSION

---

### 7.1 Main findings

In this doctoral thesis, we provide new knowledge in the methods of developing CPGs recommendations when findings from systematic reviews of RCTs or observational studies may not apply directly to the guideline development setting. This may include the assessment of multiple ways to deliver an intervention (i.e. age to start or stop screening), projecting benefits and harms to a lifetime horizon, to estimate the impact of interventions for underrepresented populations (i.e. with comorbidities), or applying results to different health care conditions from where studies have been conducted.

Our first publication integrates modelling evidence to inform the benefits and harms of a screening programme; we adapted standard GRADE evidence profiles to be capable of combining modelled and empirical evidence, across outcomes, and tailored the GRADE domains (i.e. risk of bias or imprecision) to the characteristics of modelling research evidence. In the second publication, we describe the procedures of how a European guideline panel developed a decision tree model, to assist their assessment of the downstream clinical consequences of diagnostic test interventions, and used it for formulating recommendations. Finally, partially informed by the previous studies, our third publication proposes a framework on how to incorporate modelling evidence for development of clinical guidelines, and provides guidance on using the GRADE approach to assess the certainty of this type of evidence.

Breast cancer (BC) is the second most prevalent cancer in the world and the most frequent among women [1]. BC mortality has decreased over the last decades due to improvements in treatment, services quality, and the implementation of population-based screening programmes [3]. However, there is debate on how to best implement with diverse recommendations on mammography screening frequencies. Thus, we conducted a systematic review that informed the recommendations of the European Guidelines for Breast Cancer Screening and Diagnosis. Our approach of including modelling evidence allowed us to inform outcomes (i.e. overdiagnosis, radiation induced breast cancer, life years saved) for which the evidence was sparse or not available. Modelling evidence also made possible to estimate the impact of different screening intervals in younger age groups, such as 45 to 49 years old; given that observational evidence from population registries provides evidence only for a larger age range (40 to 49 years old) the effects may be different to the age group of interest.

To inform the European guideline panel, we adapted the standard GRADE evidence profile, including modelling evidence only when other type of studies was not available and specifying the type of evidence and our judgments over its certainty (Table 11). We later used this evidence profile during the discussion with the EBCI guideline panel meeting to issue the recommendations on mammography intervals. Noteworthy, during the development of this review, we started the work of adapting the GRADE approach to assess the certainty of modelling evidence. As it was described in the methods section our first article, after discussion with other experts in the field, we decided to apply the GRADE rating of certainty, departing from the lowest certainty of the input evidence (this was subsequently modified during the development of the final guidance – see article three -). In most instance, the certainty of studies included in the systematic review on screening intervals was very low due to indirectness, since data for input parameters mostly come from opportunistic screening settings.

In some situations, decision makers may prefer developing a new model specifically designed to answer their question of interest. Although this approach would be the most appropriate, it requires suitable skills, considerable resources, and time. In our case, we first developed a systematic review of multigene tests to decide the provision of adjuvant chemotherapy in women with early BC; however, as the available evidence did not inform all downstream consequences of interest, a decision tree model without discounting was built by the guideline development group, to assist the panel on the assessment of benefits and harms of multigene testing. We presented the model, in conjunction with the evidence from our systematic review, in the Evidence to Decision frameworks to conduct the discussion and recommendation process by the ECIBC guideline (Table 12).

This process was similar to another guideline developed by the WHO for screening of cervical cancer, which projected long term consequences of different testing strategies (ref). Both experiences underline that for some clinical questions (i.e. diagnostic tests), guidelines developers should consider in advance, the need and type of modelling needed (pragmatic or more “sophisticated”) depending of the requirements, complexity of interventions, and resources available.



**Table 11.** Evidence profile presented at the ECIBC guideline meeting (age group: women 45 to 49 years)<sup>46</sup>

Certainty assessment							№ of patients		Effect		Certainty
№ of studies	Study design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations	annual mammography screening	biennial mammography screening	Relative (95% CI)	Absolute (95% CI)	
Breast cancer death averted											
2 <sup>1,2,e</sup>	modelling studies	serious <sup>f,g</sup>	not serious	very serious <sup>h,i,j</sup>	not serious	none	70 to 90	39 to 40	Ratio 1.75 to 2.31	from 30 more to 51 more per 100.000	⊕○○○ VERY LOW
Stage of breast cancer (IIB-IV)											
1 <sup>3</sup>	observational studies	serious <sup>k</sup>	not serious	very serious <sup>c,l</sup>	not serious	none	2052 cases 3573 controls		OR 0.85 (0.75 to 0.96)	-	⊕○○○ VERY LOW
							-	0.0%		--	
QALYs											
2 <sup>1,5,m</sup>	modelling studies	not serious	not serious	very serious <sup>h,i,j</sup>	not serious	none	727 to 1,540	665 to 1,060	Ratio 1.09 to 1.45	62 more to 480 more per 100.000	⊕○○○ VERY LOW
Interval cancer											
1 <sup>4,n</sup>	observational studies	serious <sup>k</sup>	serious	very serious <sup>o</sup>	not serious	none	10/14285 (0.1%)	5/3333 (0.2%)	RR 0.46 (0.16 to 1.36)	81 fewer per 100.000 (from 126 fewer to 54 more)	⊕○○○ VERY LOW
Overdiagnosis											
2 <sup>1,5,m</sup>	modelling studies	not serious	not serious	serious <sup>h,i,j</sup>	not serious	none	143 to 200	0 to 119	Ratio :Not estimable to 1.2	24 more to 200 more per 100.000	⊕○○○ VERY LOW
False positive results -10 year cumulative probability											
1 <sup>6</sup>	observational studies	serious <sup>q</sup>	not serious	very serious <sup>c,l</sup>	not serious	none	Annual screening 67% (95%CI 65% to 68%) Biennial screening 45% (95%CI 44% to 46%) Difference: 22,000 more per 100,000.				⊕○○○ VERY LOW

**False positive biopsy recommendation -10 year cumulative probability**

1 <sup>6</sup>	observational studies	serious <sup>a</sup>	not serious	very serious <sup>c,l</sup>	not serious	none	Annual screening 11% (10% to 13%) Biennial screening 6% (5% to 7%) Difference: 5,000 more per 100,000.				⊕○○○ VERY LOW
----------------	-----------------------	----------------------	-------------	-----------------------------	-------------	------	---	--	--	--	------------------

**Radiation induce breast cancer**

1 <sup>7,s</sup>	modelling studies	serious <sup>f,g</sup>	not serious	very serious <sup>h,i,j,t</sup>	not serious	none	32	18	<b>Ratio: 1.78</b>	14 more per 100.000	⊕○○○ VERY LOW
------------------	-------------------	------------------------	-------------	---------------------------------	-------------	------	----	----	--------------------	---------------------	------------------

**Death by radiation induced breast cancer**

1 <sup>7,s</sup>	modelling studies	serious <sup>f,g</sup>	not serious	very serious <sup>h,i,j,t</sup>	not serious	none	6	4	<b>Ratio:1.5</b>	2 more per 100.000	⊕○○○ VERY LOW
------------------	-------------------	------------------------	-------------	---------------------------------	-------------	------	---	---	------------------	--------------------	------------------

**CI:** Confidence interval; **RR:** Risk ratio; **OR:** Odds ratio. For modelling studies, certainty of evidence starts from low certainty and when there is more than one study informing an outcome, the number represents the range of point estimates reported across studies.

**Explanations**

- a. Rate ratio comparing annual screening relative to biennial screening was estimated by an indirect meta-analysis. Absolute effects were calculated taken as basal risk the proportion of breast cancer mortality in intervention arms of the trials of annual screening.
- b. Comparison was done by performing indirect meta-analysis of RCT (n=3) of annual mammography interval versus no screening against RCT of biennial mammography interval versus no screening.
- c. Estimations based on studies that included women from 40 to 49 years old
- d. Wide confidence interval based in indirect comparison
- e. Modelling studies used different number of women screened for calculations: 1,000 in 2 studies, and 100,000 in 2 studies. One modelling study (Vilaprinco 2017) gave inconsistent results in this year period (less deaths averted for annual interval) and then it was not included in the results of breast cancer deaths averted.
- f. One or more studies did not report information about external validation for the estimated parameters of the models.
- g. One or more studies did not report sensitivity analysis information for the estimated parameters of the models.
- h. The comparison for any interval in the models was a no screening scenario. No direct comparisons were reported.
- i. Modelling studies with data available for the 45 to 49 age period. Results were calculated by subtracting the absolute number of events from overlapping periods of screening i.e. 45 to 74 minus 50 to 74.
- j. Most models were constructed using data of surveillance registries from United States.
- k. Intervals were classified based on the month ranges elapsed between two screening mammograms prior to diagnosis. Potential high risk of misclassification.
- l. Results were extracted from groups of women with selected characteristics (e.g. normal weight, fatty or scattered fibroglandular breast density, or white race).
- m. Modelling study, used 1,000 women screened for calculations.
- n. From the In the Swedish two county trial with an average screening interval of 24 months, the calculated interval cancers for >0 to <12 months was 38%, and for 12 to <24 months was 68% (Tabar 1987).
- o. Estimations based on one study that included women from 40 to 79 years old
- p. Two modelling studies estimated the number of false positive results in annual screening of 9,150 to 56,700 and for biennial of 6,301 to 26,700 per 100,000 screened women from 45 to 49 years old (difference 2,849 to 30,000 more events).
- q. No clear information of how the intervals were estimated for the false positive cohorts or the number of individuals per interval.
- r. Two modelling studies estimated the number of benign biopsy results in annual screening of 409 to 5,600 and for biennial of 208 to 3,000 per 100,000 screened women from 45 to 49 years old (difference 201 to 2,600 more events).
- s. Modelling study, used 100,000 women screened for estimates.
- t. Incremental effects were estimated for a screening program starting at 50 and ending at 74.

**Table 12.** Evidence to Decision presented for discussion at ECIBC panel meeting (only desirable and undesirable effects are shown)

**Sparano (lymph node negative –intermediate risk only)**

Outcomes	№ of participants (studies) Follow up	Certainty of the evidence (GRADE)	Relative effect (95% CI)	Anticipated absolute effects* (95% CI)	
				Risk with endocrine therapy plus chemotherapy	Risk difference with endocrine therapy
Invasive disease-free survival	6712 (1 RCT) <sup>1,a,b</sup>	⊕⊕○○ LOW <sup>c,d,e,f</sup>	HR 1.14 (0.99 to 1.31)	Study population 153 per 1,000	<b>19 more per 1,000</b> (1 fewer to 43 more)
Freedom from recurrence at a distant site	6712 (1 RCT) <sup>1,a,b</sup>	⊕⊕○○ LOW <sup>c,d,e,f</sup>	HR 1.03 (0.80 to 1.33)	Study population 71 per 1,000	<b>2 more per 1,000</b> (14 fewer to 22 more)
Freedom from recurrence at a distant or local-regional site	6712 (1 RCT) <sup>1,a,b</sup>	⊕⊕○○ LOW <sup>c,d,e,f</sup>	HR 1.12 (0.91 to 1.38)	Study population 50 per 1,000	<b>6 more per 1,000</b> (4 fewer to 18 more)
Overall survival	6712 (1 RCT) <sup>1,a,b</sup>	⊕⊕○○ LOW <sup>c,d,e,f</sup>	HR 0.97 (0.78 to 1.21)	Study population 62 per 1,000	<b>2 fewer per 1,000</b> (13 fewer to 12 more)

1. Sparano JA, Gray RJ Makower DF Pritchard KI Albain KS Hayes DF et al. Adjuvant Chemotherapy Guided by a 21-Gene Expression Assay in Breast Cancer.. N Engl J Med.; 2018.

The PRU members developed a "back of the envelope" model to estimate the downstream consequences of testing patients with the 21-gene recurrence score versus using clinical risk scores (treating only those at high risk). Four different scenarios were hypothesized

(only two scenarios presented here).

**Scenario 1:**

In this scenario, the GDG made the extreme assumption that almost all women would be treated with chemotherapy if the multigene test would not be used (only 18,4% proportion of women would not be treated, i.e. the proportion of non-treated women among those assigned to the treatment arm in the Sparano trial).

	Chemotherapy for all	Multigene risk strategy	difference
Number of women	1000	1000	
Chemotherapies	816,0	180,0	-636,0
Invasive disease recurrence	77,6	79,7	2,0
Distant metastasis recurrence	25,8	27,2	1,4
Local or distant disease recurrence	38,3	39,0	0,7
Deaths	23,1	23,8	0,7

**Scenario 2**

In this scenario, the GDG made the assumption that, without multigene testing, women would be treated only if the clinical risk is high. The model also assumes that women with low clinical and high genomic risk have no advantage from chemotherapy.

Distribution of the clinical risk within the multigene risk strata are those reported by Sparano et al 2018.

	Chemotherapy for all	Multigene risk strategy	difference
Number of women	1000	1000	
Chemotherapies	816,0	180,0	-636,0
Invasive disease recurrence	77,6	79,7	2,0
Distant metastasis recurrence	25,8	27,2	1,4
Local or distant disease recurrence	38,3	39,0	0,7
Deaths	23,1	23,8	0,7

**The general model assumptions were:**

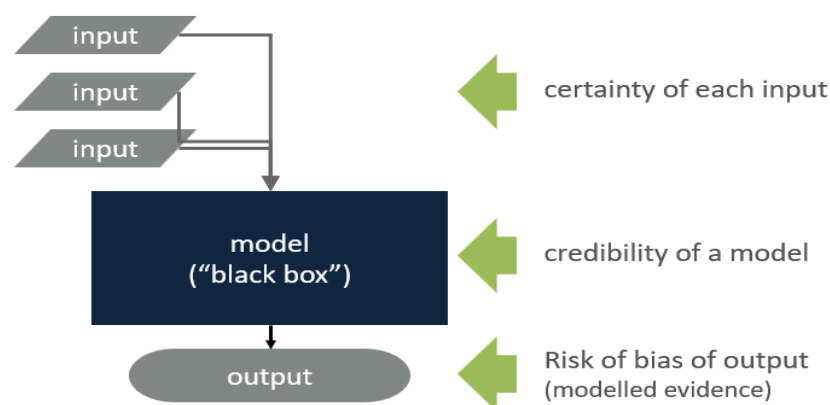
- 1) Results are based on a fixed observation time of 10 years.
  - 2) Distribution according to multigene test is low 14%; intermediate 68%; high 18%, as reported by the authors of the TAILORx trial at recruitment in 2008 before the protocol was modified in order to increase the intermediate risk group recruitment.
  - 3) The observed rate of events at 5 years in the MINDACT trial (Cardoso 2016) will remain constant at 10 years.
  - 4) Rates of events observed in the RCTs were applied to the simulated clinical score arms. It was assumed that basal risk of events for clinical score groups was homogenous between the low and the high clinical risk groups within a given genomic risk group.
- An approximately 40% reduction in the women receiving chemotherapy was considered as a desirable effect.

Finally, and partially based in the previous work, our third publication contributed over two fundamental aspects. First it presents a common framework to incorporate model outputs in health decision-making, including three options: a) developing a model de novo designed specifically to answer the very question at hand, b) searching for an existing model describing the same or a very similar problem and use it “off-the-shelf” or adapt it appropriately to answer the current question, and c) using the results from multiple existing models found in the literature.

We also provide guidance on how to assess the certainty of evidence of modelling evidence using the GRADE approach for either an individual model or across multiple models in the context of a systematic review. We considered that for modelling studies, the certainty of evidence departed from high certainty which could be then downgraded after considering methodological limitations on the GRADE domains. This approach fairly similar to what we used for our first publication which departed from the certainty level pertaining to the input evidence that informed the model parameters.

One distinctive feature of our proposed GRADE approach for modelling studies is how it define the assessment of risk of bias compared to the approach for other type of evidence (i.e. prognosis, diagnostic test). The risk of bias for modelling evidence results from both the credibility of model development and the certainty of evidence for each of model inputs (Figure 3). The credibility of a model development is determined by its structure, calibration, validation, and other factors. While we should assess the certainty of the several types of input data (bodies of evidence) used to populate the model, an efficient alternative to consider is to assess only the input parameters to which the model outputs are the most sensitive.

**Figure 3.** Risk of bias for modelling studies using the GRADE approach



## 7.2 Our results in the context of previous research

Previous initiative has also underlined the relevance of considering modelling evidence when developing clinical practice guidelines. Habbema et al identified areas where modelling evidence might be relevant to guideline development (see introduction section) based on the experience of the USPSTF on issuing recommendations for screening interventions.<sup>16</sup> Additionally, the authors provide advice for the incorporation of modelling evidence. To obtain more robust results, they developed multiple models (5 to 6 per condition) for the same questions in collaboration with the Cancer Intervention and Surveillance Modelling Network (CISNET) for lung, colorectal and breast cancer screening.<sup>16</sup> To compare candidate policies, they consider the outcomes that best capture benefits and harms and used their ratio as a common metric (i.e. number of colonoscopies per life-year gained). Finally, models may lead to different recommendations when used by another guideline groups, for example the USPSTF recommended annual screening for colorectal cancer,<sup>36</sup> whereas the Health Council of the Netherlands recommended a biennial interval based on the similar model evidence.<sup>16</sup>

Regarding our systematic review (empirical and modelling studies) for intervals of mammography screening. The results were consistent with the evidence than informed previous guidelines. The USPSTF, as described above, based their assessment on several models for the US population, concluding that when moving from biennial to annual mammography, there is a small increase in averted deaths but with a large increase of harms.<sup>76</sup> The American Cancer Society included in their review of the evidence an indirect comparison between RCTs and a model study from the CISNET collaboration, concluded that beginning screening with more frequent intervals likely results in a greater mortality reduction but the magnitude is uncertain.<sup>106</sup> In comparison, we included modelling studies for different populations (i.e. US, Canada and European countries) thus we could assess how robust where the results under different assumptions.<sup>46</sup>

The ECIBC guideline issued recommendations on multigene testing to guide adjuvant chemotherapy consistent with other guidelines. NICE issued a conditional recommendation on the use of 21-RS limited to those patients in which the risk of distant recurrence is intermediate using a validated tool such as PREDICT or the Nottingham Prognostic Index and recommended against the use of the 70- signature.<sup>107</sup> The American Society of Clinical Oncology (ASCO) also provided similar recommendations (but with a different strength) using a different methodological approach.<sup>108, 109</sup> Previous guidelines incorporated modelling

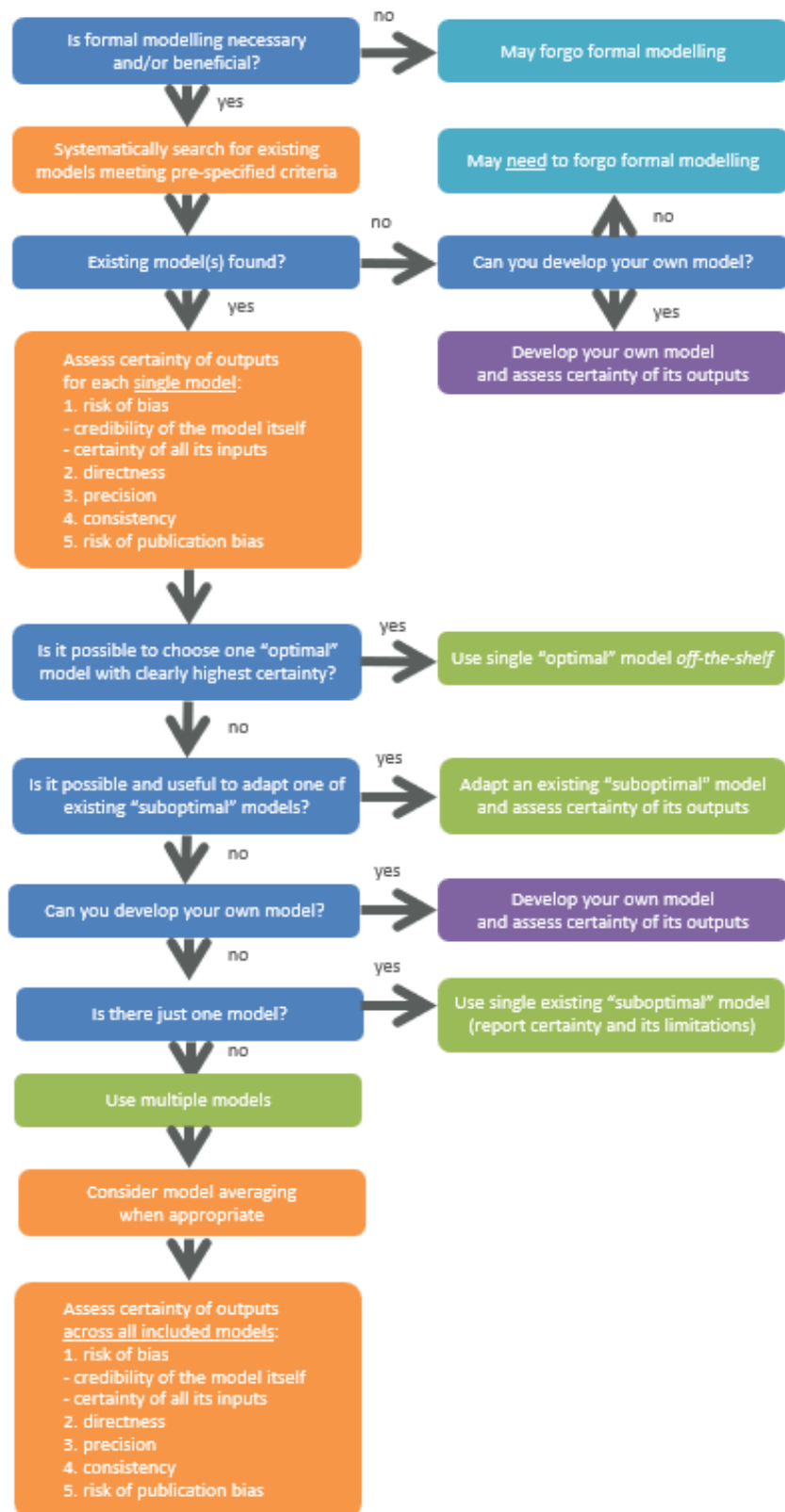


studies but limited to health economic evaluation, sharing important methodological limitations.<sup>110</sup> Our model was intended to be an exploratory tool to help guidelines panellist on their assessment of potential impact of diagnostic tests, and thus there is still a need for more robust modelling analysis for this condition.

Our approach for incorporating modelling evidence into the decision-making process for clinical guidelines is also consistent with some previous proposals but limited to systematic reviews. Kuntz et al, considered decision models as a supportive tool for systematic reviews, and included three basic approaches: a synthesis of previous modelling studies, adaptation of an existing model(s) to complement a systematic review, and the development of a de novo model.<sup>20</sup> They also consider as an ideal scenario to develop a de novo model, but acknowledging that it would be time consuming and costly.<sup>20</sup> Beside of describing similar main scenarios, we suggested a framework to assist guideline developers when considering building a new model, adapting an existing model or using existing models for development of clinical recommendations (Figure 4). This process should be guided by a systematic assessment of the certainty of any identified models as well as the resources available.

The need for development of a GRADE approach for modelling evidence has been previously recognized. The WHO conducted on 2016 a survey on 151 experts from 28 countries (half of them modelers and the other half users of model evidence); around 95% of respondents consider that modelling evidence should inform guidance for public health interventions, and 60% that findings of modelling studies can sometimes provide the same certainty of evidence as empirical studies.<sup>30</sup> This study also provided some initial insights on how to adapt the GRADE approach to modelling studies (see introduction section), that were further developed by our GRADE guidance on the third study of this doctoral thesis.

**Figure 4.** Framework to guide the incorporation of modelling evidence into clinical guideline development.



From: Brozek & Canelo-Aybar et al. (2021)<sup>21</sup>

### 7.3 Limitations and strengths

Among the common cited limitations regarding modelling includes that requires quantification of all input parameters, that some outcomes are not actually quantifiable, that to use “evidence” generated by a model is incongruent with the concept of evidence based in empirical data,<sup>33</sup> or that it is unclear the place of modelling evidence in “hierarchy of evidence”.<sup>33</sup> This criticism shows a misunderstanding about the aim of models which is to offer transparent approach to assist decisionmakers with complex decision in the context of sparse or uncertain empirical evidence.<sup>20</sup> Besides, decisions have to be made even with limited data, and implicit values are always placed on qualitative outcomes;<sup>20</sup> decision analysis provides the tool to assist this process in a transparent manner.<sup>16</sup>

However, to develop a model require substantial skills, time and logistical resources. Even adaptation of models might require model availability of input data relevant expertise and access to the original model. One strength of the framework is that the incorporation of models is a way to use resources more efficiently, relying on an initial systematic search which may identify one or more models meeting pre-specified eligibility criteria, then researchers would assess the certainty of outputs from each model.<sup>21</sup> Depending on this assessment, researchers may be able to use the results of the most direct and lowest risk of bias model or proceed to adapt it. Researchers may consider developing their own model, only when they fail to find a sufficiently direct and low risk of bias model.

Our review for intervals of mammography screening was limited by incomplete reporting of model development in particular regarding the validation process and sensitivity analysis, however the scope was considerable exhaustive and we observed consistent results thorough several modelling studies on different settings. For the decision tree model developed to assess the multigene test downstream consequences we used a fairly pragmatic approach, and despite some assumptions being potentially questionable (i.e. negligible effects in low clinical risk, same effects in studies with different duration of follow-up);<sup>96</sup> we acknowledge these limitations during the panel meetings, and used the model only as a supportive tool to better understand the potential downstream consequences of recommending the multigene test.

Finally, our guidance for the GRADE approach to assess the certainty of modelling evidence will require further efforts to provide detailed guidance on how to apply each of its domain, after testing it on real examples of both systematic reviews as on clinical guidelines. Nevertheless, it builds on the GRADE approach, which is a systematic, transparent, and

widely used method to assess the certainty of evidence across multiple types of bodies of evidence. We have developed it with the participation of experts in the field of modelling in the context of clinical practice guidelines, public health, infectious disease, cost-effectiveness modelling among other disciplines.

#### **7.4 Implications for practice and research**

The adoption of the GRADE approach and the progressive development of the methods to assess the certainty of evidence of model outputs have relevant implications for practice. First, having a systematic approach to ascertain the modelling evidence will improve the rigor of research and transparency, reducing the risk of systematic error with an overall increase in the trustworthiness on systematic reviews and clinical guideline development.<sup>21</sup> Currently several systematic reviews on the impact of control measures for the SARS-COVID-19 pandemic<sup>18, 19, 111, 112</sup> and well as for model based cost effectiveness studies<sup>113, 114</sup> have implemented the GRADE approach for modelling evidence. Additionally, adopting the GRADE approach during the model development has the potential to reduce unnecessary complexity and workload by careful consideration of the most direct evidence as model inputs, optimizing the use of different streams of evidence as model inputs.<sup>21</sup>

Our results from the systematic review of intervals of mammography screening may have different implications for practice depending on the age group, the balance between benefits and harms, and how women value the different outcomes. In the case of multigene testing, the benefits use of 21-RS in lymph node-negative women are probably larger, while the 70-GS seems acceptable only for women at high clinical risk, however there was important uncertainty on the evidence. Given both questions had a very low to low certainty of evidence and the variability of how women value outcomes at stake, guideline panellists would be likely to formulate conditional recommendations, thus a shared decision-making process to carefully explain the pros and cons of each decision is warranted. The European Guidelines on Breast Cancer Screening and Diagnosis issued the following recommendations based on those findings (Table 13 and appendix section).<sup>55</sup>

**Table 13.** Recommendations issued by the European Breast Cancer Guideline for Screening and Diagnosis based in the results of the studies included in the thesis.

Mammography intervals <sup>55</sup>	Multigene testing <sup>96</sup>
<p><b>For women aged 45 to 49 years:</b></p> <ul style="list-style-type: none"> <li>Suggests either biennial or triennial mammography over annual screening (conditional recommendation, very low certainty of evidence)</li> </ul>	<ul style="list-style-type: none"> <li>suggests the use of the 21-RS for lymph node-negative women (conditional recommendation, very low certainty of evidence)</li> </ul>
<p><b>For women aged 50 to 69 years:</b></p> <ul style="list-style-type: none"> <li>recommends against annual mammography screening (strong recommendation, very low certainty of evidence)</li> <li>suggests biennial mammography screening over triennial mammography screening (conditional recommendation, very low certainty of evidence)</li> </ul>	<ul style="list-style-type: none"> <li>suggests the use of the 70-GS for women at high clinical risk (conditional recommendation, low certainty of evidence)</li> <li>recommends not using 70-GS in women at low clinical risk (strong recommendation, low certainty of evidence)</li> </ul>
<p><b>For women aged 70 to 74 years:</b></p> <ul style="list-style-type: none"> <li>recommends against annual mammography screening (strong recommendation, very low certainty of evidence)</li> <li>suggests triennial mammography screening over biennial mammography screening (conditional recommendation, very low certainty of evidence)</li> </ul>	

Along the studies included in this doctoral thesis several priorities of research were identified, many of them pertains to the gaps on the evidence which also lead to the incorporation of modelling evidence while others are related to the refinement of the GRADE approach for this type of evidence (Table 14). Further empirical research on the effectiveness of the different screening intervals would be ideal, although it may require substantial resources. Better evidence for other imagen modalities for breast cancer screening or personalized risk stratification for BC would allow research to conduct more robust modelling studies. For multigene tests longer observation time is required to inform end-clinical outcomes. With respect to the GRADE approach the priorities include refinement of the guidance for each domain based in real examples, adaptation of evidence profiles and summary of findings format to modelling studies, and development of an specific tool to assess the credibility of models that reflects the conceptual GRADE approach.



**Table 14.** Research priorities identified across the thesis's articles

Review on mammography intervals (study 1) <sup>46</sup>
<ul style="list-style-type: none"> <li>• Empirical research on the effectiveness of the different screening intervals due to the current very low certainty of evidence</li> <li>• Cost-effectiveness studies using unitary costs from different settings, and in particular for women aged 45 to 49</li> <li>• Assessment of alternative imaging modalities to mammography</li> <li>• Tailored screening according to personalised risk assessment</li> </ul>
Review and denovo model for multigene testing (study 2) <sup>96</sup>
<ul style="list-style-type: none"> <li>• Exploring in what subgroups the use of 21-RS would have larger anticipated benefits.</li> <li>• Carrying out longer follow-up studies for 70-GS</li> </ul>
GRADE guidance for modelling evidence (study 3) <sup>21</sup>
<ul style="list-style-type: none"> <li>• Developing methods and guidance for using model outputs in health-related decision-making.</li> <li>• Provide more detailed guidance about choosing the “best” model when multiple models are found</li> <li>• Integrating the certainty of evidence from various bodies of evidence with credibility of the model and arriving at the overall certainty in model outputs</li> </ul>

---

## **CONCLUSIONS**

---

## 8. CONCLUSIONS

---

1. This doctoral thesis provided real examples on how to integrate modelling evidence into the guideline development process (either from a systematic review or developing *de-novo* model) and a guidance to assess the certainty of evidence of modelling studies.
2. In women of average BC risk, screening intervals have different trade-offs for each age group. The balance probably favours biennial screening in women 50–69. In younger women, annual screening may have a less favourable balance, while in women aged 70–74 years longer screening intervals may be more favourable (*from the systematic review of empirical and modelling studies*).
3. Testing women with early BC with 21-RS to guide the decision of providing adjuvant chemotherapy would lead to large desirable effects and trivial undesirable effects. For the 70 gene signature test, in women at low clinical risk, there will be no benefits and a very large cost; in high clinical risk population there will be moderate desirable effect and large savings.
4. Modelling evidence is relevant to assist guidelines panel on developing clinical recommendations for settings where the evidence from empirical evidence is limited or unfeasible to develop such as: projecting to a lifetime horizon, assessing complex interventions, or extrapolating results to underrepresented populations.
5. The GRADE evidence profiles formats can be adapted to incorporate modelling evidence. For example, adding modelling evidence as a separate type of studies or considering their inclusion only when empirical evidence is largely indirect or uncertain.
6. During the guideline development process there are three options of how to incorporate modelling evidence: i) to develop a model *de novo* to specifically to answer the clinical question ii) to search for an existing model describing the same or a very similar problem and use it “*off-the-shelf*” or adapt it appropriately. and iii) to use the results from multiple existing models identified in the literature.

7. A GRADE guidance is presented to assess the certainty of evidence of modelling evidence is now available and represent a relevant progress in the methods to incorporate this type of evidence on systematic reviews, health technology assessment, CPG, and overall health decision making.
8. Future research areas include developing further methods and guidance for applying each of the GRADE domains, testing the approach in additional real examples and refine the guidance accordingly, and improve the presentation formats to adequacy display modelling evidence research.

---

## REFERENCES

---



## 9. REFERENCES

---

1. Brauer F. Mathematical epidemiology: Past, present, and future. *Infectious Disease Modelling* 2017; **2**(2): 113-27.
2. Smith DL, Battle KE, Hay SI, Barker CM, Scott TW, McKenzie FE. Ross, macdonald, and a theory for the dynamics and control of mosquito-transmitted pathogens. *PLoS pathogens* 2012; **8**(4): e1002588.
3. Institute of Medicine (US) Committee on Standards for Developing Trustworthy Clinical Practice Guidelines; Graham R, Mancher M, Miller Wolman D, et al., editors. Clinical Practice Guidelines We Can Trust. Washington (DC): National Academies Press (US); 2011. Summary. Available from: <https://www.ncbi.nlm.nih.gov/books/NBK209538/>.
4. World Health Organization. (n.d.). Number of clinical trials by year, location, disease, phase, age and sex of trial participants. World Health Organization. Retrieved May 23, 2022, from <https://www.who.int/observatories/global-observatory-on-health-research-and-development/monitoring/number-of-trial-registrations-by-year-location-disease-and-phase-of-development>
5. Panteli D L-QH, Reichebner C, et al. Clinical Practice Guidelines as a quality strategy. In: Busse R, Klazinga N, Panteli D, et al., editors. Improving healthcare quality in Europe: Characteristics, effectiveness and implementation of different strategies [Internet]. Copenhagen (Denmark): European Observatory on Health Systems and Policies; 2019. (Health Policy Series, No. 53.) 9. Available from: <https://www.ncbi.nlm.nih.gov/books/NBK549283/>.
6. Guidelines International Network. Retrieved May 23, 2022, from <https://g-i-n.net/>
7. Guyatt G, Oxman AD, Akl EA, et al. GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. *Journal of clinical epidemiology* 2011; **64**(4): 383-94.
8. Alonso-Coello P, Schunemann HJ, Moher J, 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.
9. Alonso-Coello P, Oxman AD, Moher J, et al. 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.
10. Brozek JL, Akl EA, Alonso-Coello P, et al. Grading quality of evidence and strength of recommendations in clinical practice guidelines. Part 1 of 3. An overview of the GRADE approach and grading quality of evidence about interventions. *Allergy* 2009; **64**(5): 669-77.
11. The ADAPTE Collaboration. The ADAPTE Process. Resource Toolkit for Guideline Adaptation.; 2009. Report No.: version 2.0. Available from: <http://www.g-i-n.net/document-store/working-groups-documents/adaptation/adapte-resource-toolkit-guideline-adaptation-2-0.pdf>.
12. Schunemann HJ, Wiercioch W, Brozek J, et al. GRADE Evidence to Decision (EtD) frameworks for adoption, adaptation, and de novo development of trustworthy recommendations: GRADE-ADOLOPMENT. *Journal of clinical epidemiology* 2017; **81**: 101-10.
13. Martinez Garcia L, Pardo-Hernandez H, Nino de Guzman E, et al. Development of a prioritisation tool for the updating of clinical guideline questions: the UpPriority Tool protocol. *BMJ open* 2017; **7**(8): e017226.
14. Vernooij RW, Alonso-Coello P, Brouwers M, Martinez Garcia L, CheckUp P. Reporting Items for Updated Clinical Guidelines: Checklist for the Reporting of Updated Guidelines (CheckUp). *PLoS medicine* 2017; **14**(1): e1002207.

15. Dahabreh IJ, Trikalinos TA, Balk EM, Wong JB. Recommendations for the Conduct and Reporting of Modeling and Simulation Studies in Health Technology Assessment. *Annals of internal medicine* 2016; **165**(8): 575-81.
16. Habbema JD, Wilt TJ, Etzioni R, et al. Models in the development of clinical practice guidelines. *Annals of internal medicine* 2014; **161**(11): 812-8.
17. Boyd CM, Darer J, Boulton C, Fried LP, Boulton L, Wu AW. Clinical practice guidelines and quality of care for older patients with multiple comorbid diseases: implications for pay for performance. *Jama* 2005; **294**(6): 716-24.
18. Stratil JM, Biallas RL, Burns J, et al. Non-pharmacological measures implemented in the setting of long-term care facilities to prevent SARS-CoV-2 infections and their consequences: a rapid review. *The Cochrane database of systematic reviews* 2021; **9**: CD015085.
19. Krishnaratne S, Littlecott H, Sell K, et al. Measures implemented in the school setting to contain the COVID-19 pandemic. *The Cochrane database of systematic reviews* 2022; **1**: CD015029.
20. Kuntz K, Sainfort F, Butler M, et al. Decision and Simulation Modeling in Systematic Reviews. Rockville (MD); 2013.
21. Brozek JL, Canelo-Aybar C, Akl EA, et al. GRADE Guidelines 30: the GRADE approach to assessing the certainty of modeled evidence-An overview in the context of health decision-making. *Journal of clinical epidemiology* 2021; **129**: 138-50.
22. Eykhoff P. System identification; parameter and state estimation. Chester: John Wiley & Sons Ltd., 1974.
23. Roberts M, Russell LB, Paltiel AD, et al. Conceptualizing a model: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force--2. *Value in health : the journal of the International Society for Pharmacoeconomics and Outcomes Research* 2012; **15**(6): 804-11.
24. Caro JJ, Briggs AH, Siebert U, Kuntz KM, Force I-SMGRPT. Modeling good research practices--overview: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force-1. *Med Decis Making* 2012; **32**(5): 667-77.
25. Tsoi B, Goeree R, Jegathisawaran J, Tarride JE, Blackhouse G, O'Reilly D. Do different decision-analytic modeling approaches produce different results? A systematic review of cross-validation studies. *Expert review of pharmacoeconomics & outcomes research* 2015; **15**(3): 451-63.
26. Samur S, Klebanoff M, Banken R, et al. Long-term clinical impact and cost-effectiveness of obeticholic acid for the treatment of primary biliary cholangitis. *Hepatology* 2017; **65**(3): 920-8.
27. Smolen HJ, Cohen DJ, Samsa GP, et al. Development, validation, and application of a microsimulation model to predict stroke and mortality in medically managed asymptomatic patients with significant carotid artery stenosis. *Value in health : the journal of the International Society for Pharmacoeconomics and Outcomes Research* 2007; **10**(6): 489-97.
28. Schuetz CA, van Herick A, Alperin P, Peskin B, Hsia J, Gandhi S. Comparing the effectiveness of rosuvastatin and atorvastatin in preventing cardiovascular outcomes: estimates using the Archimedes model. *Journal of medical economics* 2012; **15**(6): 1118-29.
29. Trotter CL, Gay NJ, Edmunds WJ. Dynamic models of meningococcal carriage, disease, and the impact of serogroup C conjugate vaccination. *American journal of epidemiology* 2005; **162**(1): 89-100.
30. Egger M, Johnson L, Althaus C, et al. Developing WHO guidelines: Time to formally include evidence from mathematical modelling studies. *F1000Research* 2017; **6**: 1584.
31. Sculpher M, Fenwick E, Claxton K. Assessing quality in decision analytic cost-effectiveness models. A suggested framework and example of application. *Pharmacoeconomics* 2000; **17**(5): 461-77.

32. DB. P. *Methods for Quantitative Synthesis in Medicine*. 2 ed. New York, NY: Oxford UP; 2000. Meta-Analysis, Decision Analysis, and Cost-Effectiveness Analysis.
33. Sainfort F, Kuntz KM, Gregory S, et al. Adding decision models to systematic reviews: informing a framework for deciding when and how to do so. *Value in health : the journal of the International Society for Pharmacoeconomics and Outcomes Research* 2013; **16**(1): 133-9.
34. Luhn M, Prediger B, Neugebauer EAM, Mathes T. Systematic reviews of health economic evaluations: A structured analysis of characteristics and methods applied. *Research synthesis methods* 2019; **10**(2): 195-206.
35. Canelo-Aybar C, Nieto W, Vasquez V, Alva C, Chavez F. Systematic reviews of decision analytical models: an umbrella review of characteristics, quality criteria and methods applied. (non-published data). 2022.
36. Force USPST. Screening for colorectal cancer: U.S. Preventive Services Task Force recommendation statement. *Annals of internal medicine* 2008; **149**(9): 627-37.
37. Zuber AG, Lansdorf-Vogelaar I, Knudsen AB, Wilschut J, van Ballegooijen M, Kuntz KM. Evaluating test strategies for colorectal cancer screening: a decision analysis for the U.S. Preventive Services Task Force. *Annals of internal medicine* 2008; **149**(9): 659-69.
38. Porgo TV, Norris SL, Salanti G, et al. The use of mathematical modeling studies for evidence synthesis and guideline development: A glossary. *Research synthesis methods* 2019; **10**(1): 125-33.
39. Basu S, Andrews JR, Poolman EM, et al. Prevention of nosocomial transmission of extensively drug-resistant tuberculosis in rural South African district hospitals: an epidemiological modelling study. *Lancet* 2007; **370**(9597): 1500-7.
40. Schunemann HJ, Mustafa RA, Brozek J, et al. GRADE guidelines: 22. The GRADE approach for tests and strategies-from test accuracy to patient-important outcomes and recommendations. *Journal of clinical epidemiology* 2019; **111**: 69-82.
41. Schunemann HJ, Wiercioch W, Etzeandia I, et al. Guidelines 2.0: systematic development of a comprehensive checklist for a successful guideline enterprise. *CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne* 2014; **186**(3): E123-42.
42. Schunemann H BJ, Guyatt G, Oxman A, editors. GRADE handbook for grading quality of evidence and strength of recommendations. Updated October 2013. The GRADE Working Group, 2013. Available from [guidelinedevelopment.org/handbook](http://guidelinedevelopment.org/handbook).
43. Balshem H, Helfand M, Schunemann HJ, et al. GRADE guidelines: 3. Rating the quality of evidence. *Journal of clinical epidemiology* 2011; **64**(4): 401-6.
44. Philips Z, Bojke L, Sculpher M, Claxton K, Golder S. Good practice guidelines for decision-analytic modelling in health technology assessment: a review and consolidation of quality assessment. *Pharmacoeconomics* 2006; **24**(4): 355-71.
45. Jaime Caro J, Eddy DM, Kan H, et al. Questionnaire to assess relevance and credibility of modeling studies for informing health care decision making: an ISPOR-AMCP-NPC Good Practice Task Force report. *Value in health : the journal of the International Society for Pharmacoeconomics and Outcomes Research* 2014; **17**(2): 174-82.
46. Canelo-Aybar C, Posso M, Montero N, et al. Benefits and harms of annual, biennial, or triennial breast cancer mammography screening for women at average risk of breast cancer: a systematic review for the European Commission Initiative on Breast Cancer (ECIBC). *British journal of cancer* 2022; **126**(4): 673-88.
47. Schunemann HJ, Mustafa R, Brozek J, et al. GRADE Guidelines: 16. GRADE evidence to decision frameworks for tests in clinical practice and public health. *Journal of clinical epidemiology* 2016; **76**: 89-98.
48. Schunemann HJ, Oxman AD, Brozek J, et al. Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. *Bmj* 2008; **336**(7653): 1106-10.

49. Iorio A, Spencer FA, Falavigna M, et al. Use of GRADE for assessment of evidence about prognosis: rating confidence in estimates of event rates in broad categories of patients. *Bmj* 2015; **350**: h870.
50. Hooijmans CR, de Vries RBM, Ritskes-Hoitinga M, et al. Facilitating healthcare decisions by assessing the certainty in the evidence from preclinical animal studies. *PLoS One* 2018; **13**(1): e0187271.
51. Brunetti M, Shemilt I, Pregno S, et al. GRADE guidelines: 10. Considering resource use and rating the quality of economic evidence. *Journal of clinical epidemiology* 2013; **66**(2): 140-50.
52. Zhang Y, Alonso-Coello P, Guyatt GH, et al. GRADE Guidelines: 19. Assessing the certainty of evidence in the importance of outcomes or values and preferences-Risk of bias and indirectness. *Journal of clinical epidemiology* 2019; **111**: 94-104.
53. Zhang Y, Coello PA, Guyatt GH, et al. GRADE guidelines: 20. Assessing the certainty of evidence in the importance of outcomes or values and preferences-inconsistency, imprecision, and other domains. *Journal of clinical epidemiology* 2019; **111**: 83-93.
54. Schunemann HJ, Lerda D, Dimitrova N, et al. Methods for Development of the European Commission Initiative on Breast Cancer Guidelines: Recommendations in the Era of Guideline Transparency. *Annals of internal medicine* 2019; **171**(4): 273-80.
55. Schunemann HJ, Lerda D, Quinn C, et al. Breast Cancer Screening and Diagnosis: A Synopsis of the European Breast Guidelines. *Annals of internal medicine* 2019; **172**(1): 46-56.
56. Higgins JP, Altman DG, Gotzsche PC, et al. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *Bmj* 2011; **343**: d5928.
57. Sterne JA, Hernan MA, Reeves BC, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. *Bmj* 2016; **355**: i4919.
58. Guyatt GH, Oxman AD, Schunemann HJ, Tugwell P, Knottnerus A. GRADE guidelines: a new series of articles in the Journal of Clinical Epidemiology. *Journal of clinical epidemiology* 2011; **64**(4): 380-2.
59. Guyatt GH, Oxman AD, Vist GE, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *Bmj* 2008; **336**(7650): 924-6.
60. Duffy SW BR. Long term mortality results from the UK screening frequency trial *EJC Supplements* 2008; **6**(7): 48.
61. Breast Screening Frequency Trial G. The frequency of breast cancer screening: results from the UKCCCR Randomised Trial. United Kingdom Co-ordinating Committee on Cancer Research. *European journal of cancer* 2002; **38**(11): 1458-64.
62. Braithwaite D, Zhu W, Hubbard RA, et al. Screening outcomes in older US women undergoing multiple mammograms in community practice: does interval, age, or comorbidity score affect tumor characteristics or false positive rates? *Journal of the National Cancer Institute* 2013; **105**(5): 334-41.
63. Coldman AJ, Phillips N, Olivotto IA, Gordon P, Warren L, Kan L. Impact of changing from annual to biennial mammographic screening on breast cancer outcomes in women aged 50-79 in British Columbia. *Journal of medical screening* 2008; **15**(4): 182-7.
64. Dittus K, Geller B, Weaver DL, et al. Impact of mammography screening interval on breast cancer diagnosis by menopausal status and BMI. *Journal of general internal medicine* 2013; **28**(11): 1454-62.
65. Goel A, Littenberg B, Burack RC. The association between the pre-diagnosis mammography screening interval and advanced breast cancer. *Breast cancer research and treatment* 2007; **102**(3): 339-45.
66. Hubbard RA, Kerlikowske K, Flowers CI, Yankaskas BC, Zhu W, Miglioretti DL. Cumulative probability of false-positive recall or biopsy recommendation after 10 years of screening mammography: a cohort study. *Annals of internal medicine* 2011; **155**(8): 481-92.

67. Hunt KA, Rosen EL, Sickles EA. Outcome analysis for women undergoing annual versus biennial screening mammography: a review of 24,211 examinations. *AJR American journal of roentgenology* 1999; **173**(2): 285-9.
68. Kerlikowske K, Zhu W, Hubbard RA, et al. Outcomes of screening mammography by frequency, breast density, and postmenopausal hormone therapy. *JAMA internal medicine* 2013; **173**(9): 807-16.
69. Klemi PJ, Toikkanen S, Rasanen O, Parvinen I, Joensuu H. Mammography screening interval and the frequency of interval cancers in a population-based screening. *British journal of cancer* 1997; **75**(5): 762-6.
70. Miglioretti DL, Zhu W, Kerlikowske K, et al. Breast Tumor Prognostic Characteristics and Biennial vs Annual Mammography, Age, and Menopausal Status. *JAMA oncology* 2015; **1**(8): 1069-77.
71. O'Meara ES, Zhu W, Hubbard RA, et al. Mammographic screening interval in relation to tumor characteristics and false-positive risk by race/ethnicity and age. *Cancer* 2013; **119**(22): 3959-67.
72. Parvinen I, Chiu S, Pylkanen L, et al. Effects of annual vs triennial mammography interval on breast cancer incidence and mortality in ages 40-49 in Finland. *British journal of cancer* 2011; **105**(9): 1388-91.
73. White E, Miglioretti DL, Yankaskas BC, et al. Biennial versus annual mammography and the risk of late-stage breast cancer. *Journal of the National Cancer Institute* 2004; **96**(24): 1832-9.
74. Sanderson M, Levine RS, Fadden MK, et al. Mammography Screening Among the Elderly: A Research Challenge. *The American journal of medicine* 2015; **128**(12): 1362 e7-14.
75. McGuinness JE, Ueng W, Trivedi MS, et al. Factors Associated with False Positive Results on Screening Mammography in a Population of Predominantly Hispanic Women. *Cancer Epidemiol Biomarkers Prev* 2018; **27**(4): 446-53.
76. Mandelblatt JS, Stout NK, Schechter CB, et al. Collaborative Modeling of the Benefits and Harms Associated With Different U.S. Breast Cancer Screening Strategies. *Annals of internal medicine* 2016; **164**(4): 215-25.
77. Gunsoy NB, Garcia-Closas M, Moss SM. Estimating breast cancer mortality reduction and overdiagnosis due to screening for different strategies in the United Kingdom. *British journal of cancer* 2014; **110**(10): 2412-9.
78. Miglioretti DL, Lange J, van den Broek JJ, et al. Radiation-Induced Breast Cancer Incidence and Mortality From Digital Mammography Screening: A Modeling Study. *Annals of internal medicine* 2016; **164**(4): 205-14.
79. Trentham-Dietz A, Kerlikowske K, Stout NK, et al. Tailoring Breast Cancer Screening Intervals by Breast Density and Risk for Women Aged 50 Years or Older: Collaborative Modeling of Screening Outcomes. *Annals of internal medicine* 2016; **165**(10): 700-12.
80. Tsunematsu M, Kakehashi M. An analysis of mass screening strategies using a mathematical model: comparison of breast cancer screening in Japan and the United States. *Journal of epidemiology* 2015; **25**(2): 162-71.
81. van Ravesteyn NT, Miglioretti DL, Stout NK, et al. Tipping the balance of benefits and harms to favor screening mammography starting at age 40 years: a comparative modeling study of risk. *Annals of internal medicine* 2012; **156**(9): 609-17.
82. Yaffe MJ, Mainprize JG. Risk of radiation-induced breast cancer from mammographic screening. *Radiology* 2011; **258**(1): 98-105.
83. Yaffe MJ, Mittmann N, Lee P, et al. Clinical outcomes of modelling mammography screening strategies. *Health reports* 2015; **26**(12): 9-15.
84. Vilapriyo E, Forne C, Carles M, et al. Cost-effectiveness and harm-benefit analyses of risk-based screening strategies for breast cancer. *PLoS One* 2014; **9**(2): e86858.

85. Mittmann N, Stout NK, Tosteson ANA, Trentham-Dietz A, Alagoz O, Yaffe MJ. Cost-effectiveness of mammography from a publicly funded health care system perspective. *CMAJ Open* 2018; **6**(1): E77-E86.
86. Arnold M, Pfeifer K, Quante AS. Is risk-stratified breast cancer screening economically efficient in Germany? *PLoS One* 2019; **14**(5): e0217213.
87. van den Broek JJ, van Ravesteyn NT, Mandelblatt JS, et al. Comparing CISNET Breast Cancer Models Using the Maximum Clinical Incidence Reduction Methodology. *Med Decis Making* 2018; **38**(1\_suppl): 112S-25S.
88. Lee SJ, Li X, Huang H, Zelen M. The Dana-Farber CISNET Model for Breast Cancer Screening Strategies: An Update. *Med Decis Making* 2018; **38**(1\_suppl): 44S-53S.
89. van den Broek JJ, van Ravesteyn NT, Heijnsdijk EA, de Koning HJ. Simulating the Impact of Risk-Based Screening and Treatment on Breast Cancer Outcomes with MISCAN-Fadia. *Med Decis Making* 2018; **38**(1\_suppl): 54S-65S.
90. Schechter CB, Near AM, Jayasekera J, Chandler Y, Mandelblatt JS. Structure, Function, and Applications of the Georgetown-Einstein (GE) Breast Cancer Simulation Model. *Med Decis Making* 2018; **38**(1\_suppl): 66S-77S.
91. Huang X LY, Song J, Berry DA. . The MD Anderson CISNET model for estimating benefits of adjuvant therapy and screening mammography for breast cancer: an update. . *Medical Decision Making* 2016.
92. Plevritis SK, Sigal BM, Salzman P, Rosenberg J, Glynn P. A stochastic simulation model of U.S. breast cancer mortality trends from 1975 to 2000. *J Natl Cancer Inst Monogr* 2006; (36): 86-95.
93. Alagoz O, Ergun MA, Cevik M, et al. The University of Wisconsin Breast Cancer Epidemiology Simulation Model: An Update. *Med Decis Making* 2018; **38**(1\_suppl): 99S-111S.
94. Carles M, Vilaprinyo E, Cots F, et al. Cost-effectiveness of early detection of breast cancer in Catalonia (Spain). *BMC cancer* 2011; **11**: 192.
95. Lehman CD, Arao RF, Sprague BL, et al. National Performance Benchmarks for Modern Screening Digital Mammography: Update from the Breast Cancer Surveillance Consortium. *Radiology* 2017; **283**(1): 49-58.
96. Giorgi Rossi P, Lebeau A, Canelo-Aybar C, et al. Recommendations from the European Commission Initiative on Breast Cancer for multigene testing to guide the use of adjuvant chemotherapy in patients with early breast cancer, hormone receptor positive, HER-2 negative. *British journal of cancer* 2021; **124**(9): 1503-12.
97. Cardoso F, van't Veer LJ, Bogaerts J, et al. 70-Gene Signature as an Aid to Treatment Decisions in Early-Stage Breast Cancer. *The New England journal of medicine* 2016; **375**(8): 717-29.
98. Sparano JA, Gray RJ, Makower DF, et al. Adjuvant Chemotherapy Guided by a 21-Gene Expression Assay in Breast Cancer. *The New England journal of medicine* 2018; **379**(2): 111-21.
99. Albain KS, Barlow WE, Shak S, et al. Prognostic and predictive value of the 21-gene recurrence score assay in postmenopausal women with node-positive, oestrogen-receptor-positive breast cancer on chemotherapy: a retrospective analysis of a randomised trial. *The Lancet Oncology* 2010; **11**(1): 55-65.
100. Paik S, Tang G, Shak S, et al. Gene expression and benefit of chemotherapy in women with node-negative, estrogen receptor-positive breast cancer. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 2006; **24**(23): 3726-34.
101. Knauer M, Mook S, Rutgers EJ, et al. The predictive value of the 70-gene signature for adjuvant chemotherapy in early breast cancer. *Breast cancer research and treatment* 2010; **120**(3): 655-61.



102. Kaltenthaler E, Tappenden P, Paisley S. Reviewing the evidence to inform the population of cost-effectiveness models within health technology assessments. *Value in health : the journal of the International Society for Pharmacoeconomics and Outcomes Research* 2013; **16**(5): 830-6.
103. Paisley S. Identification of Evidence for Key Parameters in Decision-Analytic Models of Cost Effectiveness: A Description of Sources and a Recommended Minimum Search Requirement. *PharmacoEconomics* 2016; **34**(6): 597-608.
104. Foroutan F, Guyatt G, Zuk V, et al. GRADE Guidelines 28: Use of GRADE for the assessment of evidence about prognostic factors: rating certainty in identification of groups of patients with different absolute risks. *Journal of clinical epidemiology* 2020; **121**: 62-70.
105. Bilcke J, Beutels P, Brisson M, Jit M. Accounting for methodological, structural, and parameter uncertainty in decision-analytic models: a practical guide. *Med Decis Making* 2011; **31**(4): 675-92.
106. Myers ER, Moorman P, Gierisch JM, et al. Benefits and Harms of Breast Cancer Screening: A Systematic Review. *Jama* 2015; **314**(15): 1615-34.
107. NICE National Institute for Health Care and Excellence (2018). Tumour profiling tests to guide adjuvant chemotherapy decisions in early breast cancer. Diagnostic Guidance DG34. <https://www.nice.org.uk/guidance/dg34/resources/tumour-profiling-tests-to-guide-adjuvant-chemotherapy-decisions-in-earlybreast-cancer-pdf-1053750722245> (2020).
108. Andre F, Ismaila N, Henry NL, et al. Use of Biomarkers to Guide Decisions on Adjuvant Systemic Therapy for Women With Early-Stage Invasive Breast Cancer: ASCO Clinical Practice Guideline Update-Integration of Results From TAILORx. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 2019; **37**(22): 1956-64.
109. Krop I, Ismaila N, Stearns V. Use of Biomarkers to Guide Decisions on Adjuvant Systemic Therapy for Women With Early-Stage Invasive Breast Cancer: American Society of Clinical Oncology Clinical Practice Focused Update Guideline Summary. *Journal of oncology practice* 2017; **13**(11): 763-6.
110. Wang SY, Dang W, Richman I, Mougalian SS, Evans SB, Gross CP. Cost-Effectiveness Analyses of the 21-Gene Assay in Breast Cancer: Systematic Review and Critical Appraisal. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 2018; **36**(16): 1619-27.
111. Burns J, Movsisyan A, Stratil JM, et al. International travel-related control measures to contain the COVID-19 pandemic: a rapid review. *The Cochrane database of systematic reviews* 2021; **3**: CD013717.
112. Movsisyan A, Burns J, Biallas R, et al. Travel-related control measures to contain the COVID-19 pandemic: an evidence map. *BMJ open* 2021; **11**(4): e041619.
113. Zhang Y, Begum HA, Grewal H, et al. Cost-effectiveness of diagnostic strategies for venous thromboembolism: a systematic review. *Blood advances* 2022; **6**(2): 544-67.
114. Gajic-Veljanoski O, Li C, Schaink AK, et al. Noninvasive Fetal RhD Blood Group Genotyping: A Systematic Review of Economic Evaluations. *Journal of obstetrics and gynaecology Canada : JOGC = Journal d'obstetrique et gynecologie du Canada : JOGC* 2021; **43**(12): 1416-25 e5.

---

## APENDIX

---

## 10. APENDIX

---

### Other publications related to the doctoral thesis.

- **Article 1: Schünemann HJ, Lerda D, Quinn C, Follmann M, Alonso-Coello P, Rossi PG, Lebeau A, Nyström L, Broeders M, Ioannidou-Mouzaka L, Duffy SW, Borisch B, Fitzpatrick P, Hofvind S, Castells X, Giordano L, Canelo-Aybar C, Warman S, Mansel R, Sardanelli F, Parmelli E, Gräwingholt A, Saz-Parkinson Z; European Commission Initiative on Breast Cancer (ECIBC) Contributor Group. Breast Cancer Screening and Diagnosis: A Synopsis of the European Breast Guidelines. Ann Intern Med. 2020 Jan 7;172(1):46-56**

## Breast Cancer Screening and Diagnosis: A Synopsis of the European Breast Guidelines

Holger J. Schünemann, MD, PhD, MSc; Donata Lerda, PhD; Cecily Quinn, MD; Markus Follmann, MD, MPH, MSc; Pablo Alonso-Coello, MD, PhD; Paolo Giorgi Rossi, PhD; Annette Lebeau, MD; Lennarth Nyström, PhD; Mireille Broeders, PhD; Lydia Ioannidou-Mouzaka, MD; Stephen W. Duffy, BSc, MSc, CStat; Bettina Borisch, MD; Patricia Fitzpatrick, MD; Solveig Hofvind, PhD; Xavier Castells, MD, PhD; Livia Giordano, MD; Carlos Canelo-Aybar, MD, MSc; Sue Warman, MEd; Robert Mansel, MD; Francesco Sardanelli, MD; Elena Parmelli, PhD; Axel Gräwingholt, MD; and Zuleika Saz-Parkinson, PhD; for the European Commission Initiative on Breast Cancer (ECIBC) Contributor Group\*

**Description:** The European Commission Initiative for Breast Cancer Screening and Diagnosis guidelines (European Breast Guidelines) are coordinated by the European Commission's Joint Research Centre. The target audience for the guidelines includes women, health professionals, and policymakers.

**Methods:** An international guideline panel of 28 multidisciplinary members, including patients, developed questions and corresponding recommendations that were informed by systematic reviews of the evidence conducted between March 2016 and December 2018. GRADE (Grading of Recommendations Assessment, Development and Evaluation) Evidence to Decision frameworks were used to structure the process and minimize the influence of competing interests by enhancing transparency. Questions and recommendations, expressed as strong or conditional, focused on outcomes that matter to women and provided a rating of the certainty of evidence.

**Recommendations:** This synopsis of the European Breast Guidelines provides recommendations regarding organized screening programs for women aged 40 to 75 years who are at average risk. The recommendations address digital mammography screening and the addition of hand-held ultrasonography, automated breast ultrasonography, or magnetic resonance imaging compared with mammography alone. The recommendations also discuss the frequency of screening and inform decision making for women at average risk who are recalled for suspicious lesions or who have high breast density.

Ann Intern Med. 2020;172:46-56. doi:10.7326/M19-2125

Annals.org

For author affiliations, see end of text.

This article was published at Annals.org on 26 November 2019.

\* For members of the ECIBC Contributor Group, see the Supplement (available at Annals.org).

Despite intensified efforts by the European Council since 2003, the implementation of organized, population-based mammography screening is not uniform across Europe and depends greatly on the policies in place in different countries, the organization of health care, and available resources (1). Since the last edition of the European Guidelines on Breast Cancer Screening and Diagnosis was published in 2006 (2), new evidence regarding breast cancer and innovation in guideline methodology prompted the European Commission Initiative on Breast Cancer (ECIBC) to develop new evidence-based recommendations (in short, the European Breast Guidelines).

This article provides a synopsis of 15 key recommendations selected from the European Breast Guidelines, coordinated by the European Commission's Joint Research Centre and developed by an international guideline development group (GDG). These guidelines inform women, health professionals, and policymakers about important questions related to organized mammography screening and breast cancer diagnosis, but recommendations may apply in contexts in which orga-

nized screening programs are not in place. The recommendations primarily address women at average risk for breast cancer without increased risk due to genetic predisposition (mutations in *BRCA1* and *BRCA2*), reproductive history, or race/ethnicity. However, women with a family history, who may have a higher-than-average risk, are included in the ECIBC recommendations. Some recommendations also focus on women with high breast density and suspicious lesions on screening. The corresponding evidence reviews and recommendations are kept up to date and are available for adoption and adaptation at <https://ecibc.jrc.ec.europa.eu/recommendations>.

### GUIDELINE DEVELOPMENT AND REVIEW PROCESS

The European Commission adheres to methods for producing trustworthy guidelines (3-6), which we described in detail previously (7). In brief, the European Commission authorized new systematic reviews, or syntheses of existing ones, up to March 2016 for earlier recommendations and to December 2018 for later, more recent recommendations. This evidence informed the criteria in the GRADE (Grading of Recommendations Assessment, Development and Evaluation) Evidence to Decision (EtD) frameworks that the GDG, guided by 4 coauthors and vice chairs, used to develop the recommendations (7-10). Each recommendation is linked to the full online EtD containing references, ex-

#### See also:

Editorial comment .....	65
Web-Only Supplement	



planations (including considerations for implementation, monitoring, and research priorities), and judgments that were developed with GRADE's official app GRADEpro ([www.gradepro.org](http://www.gradepro.org)) (7).

## RECOMMENDATIONS

The Supplement Table (available at [Annals.org](http://Annals.org)) lists all 40 questions and recommendations addressed by the group as of May 2019; the first 15 recommendations listed in the table are those addressed in this synopsis. The table includes the strength (strong or conditional) and certainty-of-evidence ratings and the dates of the last pertinent literature searches. The GDG took a programmatic population perspective, suggesting that strong recommendations in this context may be adopted as policies in most situations (11). Conditional recommendations suggest that policymaking will require substantial debate and involvement of various stakeholders. The implications of the recommendations for women and clinicians are supported by more specific, linked recommendations focusing on communication and shared decision making.

### Should Organized Mammography Screening in Women Be Used?

The GDG considered women in the following age groups: 40 to 44, 45 to 49, 50 to 69, and 70 to 74 years. Evidence from some systematic reviews applied to all age groups for 1 or more EtD criteria. For example, mammography screening does not seem to create anxiety in women who are given a clear result after a mammogram. However, women recalled for further testing reported transient or long-term anxiety (from 6 months to 3 years after recall), but this was not consistent across studies (12-14). Women generally consider these undesirable effects acceptable (low certainty of evidence), and a systematic review suggested that women place a relatively low value on the psychosocial and physical effects of false-positive results and overdiagnosis; however, some studies raised concerns about whether women fully understand the resulting implications (15).

### Organized Mammography Screening in Women Aged 40 to 44 Years or 45 to 49 Years

**Recommendation 1.** For asymptomatic women aged 40 to 44 years with an average risk for breast cancer, the ECIBC's GDG suggests not implementing organized mammography screening (conditional recommendation, moderate certainty of evidence; EtD available at <http://bit.ly/2pf8I9M>).

**Recommendation 2.** For asymptomatic women aged 45 to 49 years with an average risk for breast cancer, the ECIBC's GDG suggests mammography screening over no mammography screening, in the context of an organized screening program (conditional recommendation, moderate certainty of evidence; EtD available at <http://bit.ly/2Pn1HZx>).

Eight randomized controlled trials (RCTs) of invitation to mammography screening provided breast cancer mortality data from 348 478 women younger

than 50 years (16-22), and 4 reviews of observational studies evaluated relevant outcomes (12-14, 23). Organized mammography screening probably reduces breast cancer mortality (16-22) and may reduce the risk for breast cancer stage IIA or higher (17, 18, 22, 24-28). The incidence of breast cancer and mortality increases with age, and the GDG extrapolated that the absolute health benefits are greater in women aged 45 to 49 than those aged 40 to 44 years.

Data from 5 available trials in women aged 40 to 74 years suggest an increase in the rate of mastectomy (19, 29-32), although the GDG was concerned that these results might be misleading because of lead time. One RCT suggests a rate of 12.4% (95% CI, 9.9% to 14.9%) to 22.7% (CI, 18.4% to 27.0%) for overdiagnosis, depending on whether a population or an individual woman perspective is taken (27). The number of false-positives depends on the age of first screening, and women aged 40 to 44 years also have a greater radiation risk than older women.

The balance of desirable versus undesirable health effects for starting screening at age 40 probably favors no screening (the GDG judged that the undesirable health effects are large and the desirable ones small). However, for the 45- to 49-year age group, the higher breast cancer incidence and mortality compared with women between the ages of 40 and 44, as well as observational evidence showing a greater benefit in this age group (33), led the GDG to judge that the balance of health effects probably favors screening, although the required resources for screening likely differ across settings (34, 35).

### Organized Mammography Screening in Women Aged 50 to 69 Years

**Recommendation 3.** For asymptomatic women aged 50 to 69 years with an average risk for breast cancer, the ECIBC's GDG recommends mammography screening over no mammography screening, in the context of an organized screening program (strong recommendation, moderate certainty of evidence; EtD available at <http://bit.ly/2qNKE91>).

On the basis of data from 249 930 women aged 50 to 69 years from 6 RCTs, invitation to organized mammography screening reduces breast cancer mortality (17, 19-22, 36) and may reduce the risk for breast cancer stage IIA or higher (17, 22, 24-26, 37). Five trials describe increased rates of mastectomy in women between ages 40 and 74 (19, 29-32), with concerns about lead-time bias similar to those for the younger age group. Pooled estimates from 2 RCTs suggest overdiagnosis rates of 10.1% (CI, 8.6% to 11.6%) and 17.3% (CI, 14.7% to 20.0%) (37, 38).

The cost-effectiveness studies probably favored screening, but this would vary across countries (34, 39-41). The GDG determined that screening in this age group has a net health benefit, and other EtD criteria were generally in favor of implementing organized mammography screening. Thus, despite uncertainty about the relative importance of outcomes or values, the GDG made a strong recommendation for organized screening but emphasizes that all invited women



## CLINICAL GUIDELINE

## A Synopsis of the European Breast Guidelines

should receive clear information about the desirable and undesirable effects to make informed decisions.

#### Organized Mammography Screening in Women Aged 70 to 74 Years

**Recommendation 4.** For asymptomatic women aged 70 to 74 years with an average risk for breast cancer, the ECIBC's GDG suggests mammography screening over no mammography screening, in the context of an organized screening program (conditional recommendation, moderate certainty of evidence; EtD available at <http://bit.ly/31KjCMA>).

According to 2 RCTs of invitation to mammography screening in 18 233 women aged 70 years and older (19, 21), organized mammography screening reduces breast cancer mortality, the risk for breast cancer stage IIA or higher, and detection of tumors larger than 50 mm (25).

Five trials in women aged 40 to 74 years described increased mastectomy rates (19, 29–32). Concerns have been raised about lead-time bias, the small number of women aged 70 to 74 years included for the outcome of mastectomy, and the available data for overdiagnosis being derived exclusively from women aged 50 to 69 years for an overall judgment of probable net health benefit. Other EtD criteria also were generally in favor of implementing organized mammography screening in this age group.

#### How Often Should Women Attend an Organized Mammography Screening Program?

##### Women Aged 45 to 49 Years

**Recommendation 5.** For asymptomatic women aged 45 to 49 years with an average risk for breast cancer, the ECIBC's GDG suggests either biennial or triennial mammography over annual screening in the context of an organized screening program (conditional recommendation, very low certainty of evidence; EtD available at <http://bit.ly/32O1faP>).

##### Women Aged 50 to 69 Years

**Recommendations 6 and 7.** For asymptomatic women aged 50 to 69 years with an average risk for breast cancer, the ECIBC's GDG recommends against annual mammography screening (strong recommendation, very low certainty of evidence; EtD available at <http://bit.ly/2B1zNzj>) and suggests biennial mammography screening over triennial mammography screening in the context of an organized screening program (conditional recommendation, very low certainty of evidence; EtD available at <http://bit.ly/31QCUCi>).

##### Women Aged 70 to 74 Years

**Recommendations 8 and 9.** For asymptomatic women aged 70 to 74 years with an average risk for breast cancer, the ECIBC's GDG recommends against annual mammography screening (strong recommendation, very low certainty of evidence; EtD available at <http://bit.ly/342qJS0>) and suggests triennial mammography screening over biennial mammography screening

in the context of an organized screening program (conditional recommendation, very low certainty of evidence; EtD available at <http://bit.ly/2JpK1su>).

The GDG compared annual, biennial, and triennial screening intervals in women for whom the GDG either strongly (ages 50 to 69 years) or conditionally (ages 45 to 49 and 70 to 74 years) recommended screening (Table 1). Evidence exists from RCTs to compare annual with triennial screening in women aged 50 to 69 years (42) and from observational studies (43–46) for a broader age range. To fill gaps in the direct evidence, the GDG used evidence from indirect comparisons of annual (18, 20, 47) or biennial (19, 48) screening compared with no screening, as well as the results of modeling studies (44, 49, 50). The GDG also conducted its own simple modeling—for example, calculating events by subtracting the estimated outcome rates in women aged 45 to 69 years (or 70 to 74 years) from those aged 50 to 69 years (or 70 to 74 years)—and assumed that effects were incremental to those found for women aged 50 to 69 years (or 70 to 74 years) at screening.

The benefits resulting from more rather than less frequent screening differed across age groups but suggest that for all age groups, annual screening may reduce breast cancer mortality compared with biennial or triennial screening. Compared with biennial screening, the incidence of stage IIB to IV breast cancer and interval cancer seemed lower with annual screening (51–53). More quality-adjusted life-years seemed to be gained with annual than biennial or triennial screening (44, 49). When biennial was compared with triennial screening, the reported benefits were similar in all age groups, except for detection of stage IIB to IV breast cancer in women aged 50 to 69 years, which favored biennial screening.

Harms also differed across age groups but showed similar patterns. Annual screening showed increased overdiagnosis rates, more false-positive results (in some comparisons, >30% more), and more suggestions for follow-up with biopsies for false-positive results (in some comparisons, >5% more) across age groups compared with biennial or triennial screening (43, 44, 49, 52, 54). Biennial screening probably leads to more overdiagnosis, false-positive results, and suggestions for follow-up with biopsies for false-positive results than triennial screening, but the differences become smaller with increasing age (44, 45). Radiation-induced breast cancer and higher rates with biennial or triennial screening of radiation-induced breast cancer deaths probably result from annual (6 in 100 000 women) and biennial screening (4 in 100 000 women) compared with triennial screening (50).

#### What Tests Should Be Used to Screen for Breast Cancer?

The following 2 recommendations about digital breast tomosynthesis (DBT), originally made in April



**Table 1.** Multiple-Intervention Comparison of Desirable and Undesirable Consequences of Annual, Biennial, and Triennial Mammography Screening for Women Aged 45 to 49, 50 to 69, and 70 to 74 Years

Evaluation Criteria	Screening Intervals for Women Aged 45 to 49 Years		
	Annual vs. Triennial	Triennial vs. Biennial	Annual vs. Biennial
Certainty of evidence	Very low	Very low	Very low
Balance of health effects	Probably favors triennial screening	Probably favors biennial screening	Probably favors biennial screening
Resources required	Large costs	Moderate savings	Moderate costs
Cost-effectiveness	Probably favors triennial screening	Probably favors triennial screening	Probably favors biennial screening
Equity	Varies	Varies	Varies
Acceptability	Varies	Varies	Varies
Feasibility	Varies	Yes, compared with biennial	Varies
Overall judgment	The GDG judged that biennial or triennial screening provided the most net desirable consequences compared with annual screening. Biennial screening probably provides more net desirable health consequences than triennial, but costs are lower for triennial screening programs.		

Evaluation Criteria	Screening Intervals for Women Aged 50 to 69 Years	
	Annual vs. Triennial	Triennial vs. Biennial
Certainty of evidence	Very low	Very low
Balance of health effects	Probably favors triennial screening	Probably favors biennial screening
Resources required	Large costs with annual screening	Moderate savings with biennial screening
Cost-effectiveness	Does not favor either	Does not favor either
Equity	Varies	Varies
Acceptability	Varies	Varies
Feasibility	Probably no, compared with triennial	Yes, compared with biennial
Overall judgment	The GDG judged that the net desirable consequences of annual screening are much smaller than those of triennial screening, largely because of the harms from more frequent screening (a strong recommendation against annual screening resulted). The GDG judged that triennial screening has less net desirable consequences than biennial, but the panel was not as certain (a conditional recommendation resulted). The GDG decided by logic that biennial also has more net desirable consequences than annual screening, and it did not produce a detailed EtD framework.	

Evaluation Criteria	Screening Intervals for Women Aged 70 to 74 Years		
	Annual vs. Biennial	Annual vs. Triennial	Triennial vs. Biennial
Certainty of evidence	Very low	Very low	Very low
Balance of effects	Probably favors biennial screening	Probably favors triennial screening	Does not favor either the intervention or the comparison
Resources required	Large costs with annual	Moderate costs with annual	Moderate savings with triennial
Cost-effectiveness	Favors biennial	No included studies	Probably favors triennial
Equity	Varies	Varies	Varies
Acceptability	Probably no, compared with biennial	Probably no, compared with triennial	Probably yes, compared with biennial
Feasibility	Probably no, compared with biennial	Probably no, compared with triennial	Yes, compared with biennial
Overall judgment	The GDG judged that biennial and triennial screening provide similar net desirable consequences and both of these intervals have more net desirable consequences than annual screening intervals.		

EtD = Evidence to Decision; GDG = guideline development group.

2016, were updated and changed in November 2018.

#### **Should Screening With DBT (Including Synthesized 2-Dimensional Images) Versus Digital Mammography Be Used for Early Detection of Breast Cancer in Asymptomatic Women?**

**Recommendation 10.** For asymptomatic women with an average risk for breast cancer, the ECIBC's GDG suggests screening with digital mammography over DBT, in the context of an organized screening program (conditional recommendation, very low certainty of evidence; EtD available at <http://bit.ly/2pRtw1G>). Because the GDG made a strong recommendation for screening at ages 50 to 69 years, this applies specifically to this age group.

We found 9 relevant observational studies (55–63), but they did not measure the outcomes of breast can-

cer mortality, cancer stage, and quality of life. Screening with DBT increased breast cancer detection compared with digital mammography (55–57, 61, 62). No differences in interval cancer detection rate, recall rate, or false-positive recall were found between DBT and digital mammography (55–58, 61–63).

The resources needed to move to DBT were considered moderate by the GDG, not only because of the greater costs of the machines but also because of the human resources required. One observational study (59) reported that radiologists' reading time would double for DBT compared with digital mammography, but staff costs may vary depending on the country. The GDG emphasized that research on direct outcomes (namely, other-cause mortality, breast cancer mortality, radiation-induced cancer, and quality of life) is not yet available, leading to uncertainty in the balance of health effects from using DBT in screening programs.

## CLINICAL GUIDELINE

A Synopsis of the European Breast Guidelines

**Should Screening Using DBT (Including Synthesized 2-Dimensional Images) in Addition to Digital Mammography Versus Digital Mammography Alone Be Used for Early Detection of Breast Cancer in Asymptomatic Women?**

**Recommendation 11.** For asymptomatic women with an average risk for breast cancer, the ECIBC's GDG suggests screening with digital mammography alone over screening with DBT in addition to digital mammography, in the context of an organized screening program (conditional recommendation, very low certainty of evidence; EtD available at <http://bit.ly/33aQf6V>).

We found 1 RCT (64) and 10 observational studies (55–60, 65–71) that were relevant. Screening with DBT in addition to digital mammography increased the cancer detection rate and detection of invasive cancer compared with digital mammography alone (55–58, 64–66, 69). No differences were found in recall rate (55, 56, 58, 64–66, 69), but in 4 of the observational studies the rate of false-positive recalls was increased when both techniques were combined, although the RCT (64) showed no differences. The GDG agreed that the effect would vary depending on the baseline rate. Despite about a 2-fold increase in radiation dose with use of both DBT and digital mammography, the GDG determined that the absolute increase in radiation-induced cancer was probably small (58–60, 64).

The resources needed to adopt DBT plus digital mammography were considered large because of the higher costs of the machines and the necessary human resources (72). For instance, radiologists' reading time would at least double by using both techniques (77 to 191 seconds) compared with digital mammography alone (33 to 67 seconds) (56, 59, 73). Although the GDG could not determine whether using DBT in addition to digital mammography in screening programs provided a net health benefit, it concluded that, overall, the undesirable consequences were greater than the desirable ones.

**What Tests Should Be Used to Screen for Breast Cancer in Women With Dense Breast Tissue?**

The GDG answered 4 questions about whether a woman whose mammogram shows no breast cancer but who has dense breast tissue should have another mammogram or other tests, such as DBT, magnetic resonance imaging (MRI), or ultrasonography (automated or hand-held). The DBT question currently is being updated, so only the other 3 questions are described in detail here.

**Tailored Screening With Automated Breast Ultrasonography**

**Recommendation 12.** For asymptomatic women with high mammographic breast density and negative mammography results, in the context of an organized screening program, the ECIBC's GDG suggests not implementing tailored screening with automated breast ultrasonography (ABUS) over mammography screening alone (conditional recommendation, very low certainty of evidence; EtD available at <http://bit.ly/341Kg4V>).

We found 3 observational studies reporting the effect on breast cancer detection and recall rates of additional screening with ABUS after a negative mammography result (74–76). The addition of ABUS after a negative mammography result increased the number of breast cancer cases detected. However, interaction may exist between risk factors other than breast density and detection rate; therefore, absolute or relative effects may not be comparable. The GDG expressed concern about the link between higher detection rate and mortality because of the lack of evidence for the outcome of breast cancer mortality. Two studies suggested an increase in recall rate with ABUS (74, 75). The GDG determined that the balance of health effects favors neither ABUS after mammography nor mammography alone, and other EtD criteria generally were in favor of not implementing additional screening with ABUS.

**Tailored Screening With Hand-Held Ultrasonography Based on High Mammographic Breast Density**

**Recommendation 13.** For asymptomatic women with high mammographic breast density and a negative mammography result, in the context of an organized screening program, the ECIBC's GDG suggests not implementing tailored screening with hand-held ultrasound (HHUS) over mammography screening alone where such is not already the practice (conditional recommendation, low certainty of evidence; EtD available at <http://bit.ly/366cEVx>).

Additional screening with HHUS after a negative mammography result increased the number of breast cancer cases detected compared with mammography alone in 1 randomized and 5 observational studies (77–82). Because of a lack of evidence about the anticipated effects on mortality and other outcomes, the GDG could not determine what the desirable effects would be.

We found no evidence of undesirable effects of adding HHUS after a mammogram. The GDG considered indirect evidence suggesting that the lifetime incremental cost for biennial screening with supplemental HHUS is \$560 per woman aged 50 to 74 years and the incremental cost-effectiveness ratio per quality-adjusted life-year gained is equal to \$238 550 in purchasing power parity in the United States (83). The GDG determined that the balance of health effects favors neither HHUS after mammography nor mammography alone, so the additional resources needed to implement HHUS led the GDG to advise against adding HHUS for these women.

**Tailored Screening With MRI Based on High Mammographic Breast Density**

**Recommendation 14.** For asymptomatic women with high mammographic breast density and a negative mammography result, in the context of an organized screening program, the ECIBC's GDG suggests not implementing tailored screening with MRI over mammography screening alone (conditional recommendation,



Table 2. Recommendations for Breast Cancer Screening for Average-Risk Women\*

Guideline, Year (Reference)	Age, y	Direction and Strength of the Recommendation (if Provided)	Age to Stop Screening Mammography	Screening Interval
ACOG, 2017 (99)	40 (discuss; offer if chosen by SDM) 50-74 (start screening if not previously started)	Discuss and offer if chosen In favor	75 y	Every 1 or 2 y
ACP, 2019 (100)†	40-49	No recommendation made, only discussion should be held	75 y with life expectancy <10 y	Every 2 y
ACS, 2015 (101)	50-74 40-44 (discuss; offer if chosen by SDM)	Offer screening Discuss and offer if chosen	Life expectancy <10 y	Every 1 y for age 45-54 y and every 2 y for age ≥55 y
ACR, 2017 (102)	45 (start screening)	In favor	None	Every 1 y
NCCN, 2018 (103)	40 (start screening)	In favor	None	Every 1 y
WHO, 2014 (104)	50-75	In favor	75 y	Every 2 y
USPSTF, 2016 (105)	40-49 50-75	Discuss and offer if chosen In favor	75 y	Every 2 y
CTFPHC, 2018 (106)‡	40-49 50-69 70-74	Suggest against Suggest in favor Suggest against	75 y	Every 2-3 y
ECIBC, 2019	40-45 45-49 50-69 70-74	Suggest against   Suggest in favor   Recommend   Suggest in favor, organized mammography screening	74 y	Not applicable§ Every 2-3 y Every 2 y Every 3 y

ACOG = American College of Obstetricians and Gynecologists; ACP = American College of Physicians; ACR = American College of Radiology; ACS = American Cancer Society; CTFPHC = Canadian Task Force on Preventive Health Care; ECIBC = European Commission Initiative on Breast Cancer; NCCN = National Comprehensive Cancer Network; SDM = shared decision making; USPSTF = U.S. Preventive Services Task Force; WHO = World Health Organization.

\* Modified and updated from Oaseem and colleagues (100).

† The ACP did not produce a guideline but a guidance statement; no systematic reviews were conducted, but existing guidelines were reviewed to formulate ungraded statements rather than recommendations.

‡ The CTFPHC guideline addressed only women aged 40-74 y.

§ If implemented, follow recommendations for women aged 45-49 y, every 2-3 y.

|| SDM should take place in organized programs, applicable to all ECIBC recommendations.

very low certainty of evidence; EtD available at <http://bit.ly/32PMDaK>).

We found 5 observational studies reporting on rates of breast cancer detection and recall (84-88). Additional testing with MRI markedly increased the breast cancer detection rate compared with mammography alone, raising concerns about overdiagnosis; no evidence was found for mortality or other related outcomes. The GDG discussed the importance of false-positives and interval cancer cases in particular, as well as possible side effects of the contrast medium used in MRI-based screening.

Although the GDG found no evidence regarding resources and cost-effectiveness, it assumed that the costs of MRI equipment and examinations are much higher than those of digital mammography. The GDG determined that MRI after mammography in women with high mammographic breast density probably results in net harm, and after also considering the increased costs, the group advised against additional testing with MRI for these women.

#### What Test Should Be Used for Diagnosis in Average-Risk Women Recalled Because of Suspicious Lesions at Mammography Screening?

**Recommendation 15.** The ECIBC's GDG suggests using DBT over diagnostic mammography projections in women at average risk for breast cancer recalled for suspicious lesions at mammography screening

(conditional recommendation, moderate certainty of test accuracy data; EtD available at <http://bit.ly/31KV0mD>).

We found 10 studies (72, 89-97) reporting the accuracy of DBT compared with assessment mammography for diagnosis in women recalled because of suspicious lesions at mammography screening. Digital breast tomosynthesis leads to more true-positives (patients correctly diagnosed with breast cancer), fewer false-negatives (patients incorrectly classified as not having breast cancer), more true-negatives (women without breast cancer), and fewer false-positives (women incorrectly assumed to have breast cancer). Although the GDG found no evidence regarding the consequences of these accuracy results on clinical outcomes, the group discussed the possible concern about radiation dose in DBT. Only 1 study reported radiation dose (a surrogate outcome to assess the risk for radiation-induced breast cancer), and the GDG judged that side effects of DBT compared with assessment mammography (including magnification) were likely to be trivial (91).

The GDG concluded that DBT probably confers a net health benefit, and although the DBT device is much more expensive than the equipment needed for magnification mammography, information for other EtD criteria also generally favored using DBT for diagnosis in women recalled for suspicious lesions at mammography screening.

## CLINICAL GUIDELINE

## A Synopsis of the European Breast Guidelines

## DISCUSSION

In developing the European Breast Guidelines, the ECIBC used a rigorous approach to produce recommendations on breast cancer screening and diagnosis for women. The guidelines include recommendations that address the use of various tests, including DBT, MRI, ABUS, and HHUS, for women who have suspicious lesions on mammography screening or who have dense breast tissue. The use of some tests, such as DBT, in women with high breast density are not addressed in this synopsis, but updates that incorporate emerging pertinent evidence and related recommendations are under way.

The strengths of the guidelines include their adherence to requirements for trustworthy development (4, 6, 98), including the public and transparent display of all evidence, considerations, and judgments for use by women, health care professionals, policymakers, and researchers (<https://ecibc.jrc.ec.europa.eu/recommendations>). Previously we described limitations of our guidelines related to the lack of high-certainty evidence for some recommendations, the absence of formal modeling, conflicts of interest, and process issues (7). We believe these limitations are balanced by the recommendations' transparency, which allows for scientific discourse and comparison with other guidelines.

Table 2 shows that our key recommendations on screening in women younger than 50 years generally agree with guidelines from the American College of Obstetricians and Gynecologists (99), American College of Physicians (100), and American Cancer Society (101), which suggest shared decision making. However, our recommendations are less strong and favor wider screening intervals than those of the American College of Radiology (102) and the National Comprehensive Cancer Network (103) (Table 2). For the other age groups, the recommendations agree with those of the World Health Organization (104) and U.S. Preventive Services Task Force (105) but not with those of the Canadian Task Force for Preventive Health Care (CTFPHC) (106). The CTFPHC also used the GRADE EtD approach, allowing a more detailed exploration of the differences. The key difference is the CTFPHC's recommendation against screening in women until age 49 and after age 69. We believe this is a result of the CTFPHC attaching a higher value to potential harms; more concerns about risk of bias, leading to lower certainty of the evidence; and greater importance attached to outcomes for which less information was available. This in turn led the CTFPHC to assign overall lower certainty. The ECIBC's GDG carefully analyzed the existing data and supplemented the RCTs when available with observational studies and had no serious concerns about risk of bias in the trials overall (see explanations in the evidence profile at <http://bit.ly/2qNKE91>). In contrast to the CTFPHC, the ECIBC's GDG also did not have concerns about inconsistency in trial results, making the GDG more confident in the recommendation for women aged 50 to 69 years.

The feasibility of implementing a recommendation, the acceptability of that recommendation, the required resources, and the associated values are often context dependent. Some countries have started or intend to adapt or adopt specific recommendations in Europe (Bulgaria, Czech Republic, Denmark, Estonia, Germany, Italy, Norway, and Slovakia) and outside Europe (Bahrain, Chile, China, and Tunisia) using the EtD frameworks and the GRADE-ADOLOPMENT (GRADE EtD frameworks for adoption, adaptation, and de novo development of trustworthy recommendations) methodology (107).

In summary, this synopsis presents and summarizes the rationale for 15 key recommendations of the European Breast Guidelines. The complete set of recommendations (Supplement Table) provides advice on additional issues, such as how to communicate with vulnerable populations about screening options, how to inform women about results, the use of decision aids, how to work up calcifications, whether to use clip marking for core needle biopsies, and whether mammograms require double reading.

From McMaster University, Hamilton, Ontario, Canada (H.J.S.); European Commission, Joint Research Centre, Ispra, Italy (D.L., E.P., Z.S.); St. Vincent's University Hospital, Dublin, Ireland (C.Q.); German Cancer Society, Berlin, Germany (M.F.); Iberoamerican Cochrane Center, Barcelona, Spain (P.A.); Azienda Unità Sanitaria Locale-IRCCS di Reggio Emilia, Reggio Emilia, Italy (P.G.R.); Private Group Practice for Pathology, Lübeck, Germany (A.L.); Umeå University, Umeå, Sweden (L.N.); Radboud University Medical Centre, Nijmegen, the Netherlands (M.B.); University of Athens Medical School, Athens, Greece (L.I.); Queen Mary University of London, London, United Kingdom (S.W.D.); University of Geneva, Geneva, Switzerland (B.B.); National Screening Service, Dublin, Ireland (P.F.); Cancer Registry of Norway, Oslo, Norway (S.H.); IMIM (Hospital del Mar Medical Research Institute), Barcelona, Spain (X.C.); CPO-Piedmont - AOU Città della Salute e della Scienza, Torino, Italy (L.G.); Iberoamerican Cochrane Center, Barcelona, Spain (C.C.); Langford, North Somerset, United Kingdom (S.W.); Cardiff University, Cardiff, United Kingdom (R.M.); Università degli Studi di Milano, Milan, Italy (F.S.); and Radiologie am Theater, Paderborn, Germany (A.G.).

**Acknowledgment:** The authors thank Jesús López Alcalde who coordinated the JRC methods for the ECIBC guidelines between 2013 and 2016, Chris de Wolf, Roberto D'Amico, and Peter Rabe for their contributions to the ECIBC guidelines as previous Joint Research Centre colleagues or retired GDG members.

**Financial Support:** By the European Commission.

**Disclosures:** Members of the GDG do not receive financial compensation for their work but are reimbursed by the European Commission for travel-related expenses for the meetings organized by the Joint Research Centre. Dr. Schünemann reports that the European Commission had chosen to use GRADE as one of the core methods for its guidelines before involving Dr. Schünemann. He was invited to participate in the guideline development as a methodologist and was elected



## A Synopsis of the European Breast Guidelines

by the ECIBC GDG as its cochair. He is also cochair of the GRADE working group and has codeveloped its methodology and tools, was commissioned by the National Academy of Sciences to write the background reports for the Institute of Medicine standards for trustworthy guideline development with coauthors, has conducted Cochrane reviews (currently is director of Cochrane Canada), and is a member of the Board of Trustees of the Guidelines International Network. He has not received direct financial payments for ECIBC work but has received travel support and is under contract from the European Commission for a project relating to other guideline methods. Dr. Quinn is the chair of the European Working Group for Breast Screening Pathology (EWGBSP). Various companies have provided some sponsorship to the EWGBSP for group meetings. Dr. Alonso-Coello reports that his institution received payments from the European Commission to develop the systematic reviews informing the recommendations. He coordinated the systematic review team informing the guidelines. He is a member of the GRADE guidance group of the GRADE working group and a member of the board of the Guidelines International Network. He has contributed to the development of some of the methodology and tools. Dr. Giorgi Rossi reports that he published opinions about the superiority of public, organized, population-based screening programs instead of opportunistic and private screening, according to the European Commission recommendations 2003/878/EC. He is on the steering committee of MyPeBS (My Personal Breast Screening), a European multicentric trial to compare the effectiveness of personalized screening programs and standard protocols, and of the RETomo and MAITA trials, comparing digital mammography and DBT in breast cancer screening. Dr. Lebeau reports grants and reimbursement for travel-related expenses related to consultancy from Roche Pharma and Novartis Oncology, and grants from BioNTech Diagnostics, outside the submitted work. Dr. Lebeau reports that she is chair of the Breast Pathology Working Group of the German S3 Guidelines for the Early Detection, Diagnosis, Treatment and Follow-up of Breast Cancer; a member of the Scientific Advisory Council for the Cooperation Alliance Mammography (Kooperationsgemeinschaft Mammographie GBR), Germany; a member of the certification commission "breast cancer centres" as a representative of the German Society of Pathology and the Federal Association of German Pathologists; and a board member of the German Society of Pathology. Dr. Hofvind reports permanent employment as a researcher at the Cancer Registry of Norway, independent of her job as administrative leader of BreastScreen Norway. Dr. Canelo-Aybar reports that his institution received payments from the European Commission to develop the systematic reviews informing the recommendations. He is a member of the GRADE Working Group. Dr. Sardanelli is responsible for the department of radiology performing mammographic screening at the IRCCS Policlinico San Donato, Milan, Italy. He is a member of the executive board of the European Society of Breast Imaging and codirector of the Breast MRI training course run by this society; director of the European Network for Assessment of Imaging in Medicine, joint initiative of the European Institute for Biomedical Imaging Research; editor-in-chief of *European Radiology Experimental*; and a recipient of research grants from Bracco, Bayer, and General Electric. Dr. Sardanelli is not a member of the GDG but did participate in formulating the recommendations. Dr. Parmelli is employed by the European Commission. Dr. Gräwingholt is head of the mammography screening cen-

Annals.org

## CLINICAL GUIDELINE

ter Paderborn, consultant radiologist for screening programs in Switzerland, and consultant radiologist for Hellenic School of Senology. Dr. Saz-Parkinson is employed by the European Commission, coordinating the ECIBC's GDG. Authors not named here have disclosed no conflicts of interest. Disclosures can also be viewed at [www.acponline.org/authors/icmje/ConflictOfInterestForms.do?msNum=M19-2125](http://www.acponline.org/authors/icmje/ConflictOfInterestForms.do?msNum=M19-2125).

**Corresponding Author:** Holger J. Schünemann, MD, PhD, MSc, McMaster University Health Sciences Centre, Room 2C16, 1280 Main Street West, Hamilton, Ontario L8N 4K1, Canada (e-mail, [schuneh@mcmaster.ca](mailto:schuneh@mcmaster.ca)); and Zuleika Saz-Parkinson, PhD, European Commission, Joint Research Centre, Via Enrico Fermi 2749, TP 127, Ispra VA, 21027, Italy (e-mail, [zuleika.saz-parkinson@ec.europa.eu](mailto:zuleika.saz-parkinson@ec.europa.eu)).

Current author addresses and author contributions are available at [Annals.org](http://Annals.org).

## References

1. Deandrea S, Molina-Barceló A, Uluturk A, et al. Presence, characteristics and equity of access to breast cancer screening programmes in 27 European countries in 2010 and 2014. Results from an international survey. *Prev Med*. 2016;91:250-263. [PMID: 27527575] doi:10.1016/j.ypmed.2016.08.021
2. Perry N, Broeders M, DeWolf C, et al. European Guidelines for Quality Assurance in Breast Cancer Screening and Diagnosis. 4th ed. Luxembourg: Office for Official Publications of the European Communities; 2006.
3. Institute of Medicine. Clinical Practice Guidelines We Can Trust. Washington, DC: National Academies Press; 2011.
4. Oxman AD, Fretheim A, Schünemann HJ; SURE. Improving the use of research evidence in guideline development: introduction. *Health Res Policy Syst*. 2006;4:12. [PMID: 17116254]
5. Qaseem A, Forland F, Macbeth F, et al; Board of Trustees of the Guidelines International Network. Guidelines International Network: toward international standards for clinical practice guidelines. *Ann Intern Med*. 2012;156:525-31. [PMID: 22473437] doi:10.7326/0003-4819-156-7-201204030-00009
6. Schünemann HJ, Wiercioch W, Etxeandia I, et al. Guidelines 2.0: systematic development of a comprehensive checklist for a successful guideline enterprise. *CMAJ*. 2014;186:E123-42. [PMID: 24344144] doi:10.1503/cmaj.131237
7. Schünemann HJ, Lerda D, Dimitrova N, et al; European Commission Initiative on Breast Cancer Contributor Group. Methods for development of the European Commission Initiative on Breast Cancer guidelines: recommendations in the era of guideline transparency. *Ann Intern Med*. 2019. [PMID: 31330534] doi:10.7326/M18-3445
8. Alonso-Coello P, Oxman AD, Moher J, 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. [PMID: 27365494] doi:10.1136/bmj.i2089
9. Alonso-Coello P, Schünemann HJ, Moher J, et al; GRADE Working Group. GRADE Evidence to Decision (EtD) frameworks: a systematic and transparent approach to making well informed healthcare choices. 1: Introduction. *BMJ*. 2016;353:i2016. [PMID: 27353417] doi:10.1136/bmj.i2016
10. Schünemann HJ, Mustafa R, Brozek J, et al; GRADE Working Group. GRADE Guidelines: 16. GRADE evidence to decision frameworks for tests in clinical practice and public health. *J Clin Epidemiol*. 2016;76:89-98. [PMID: 26931285] doi:10.1016/j.jclinepi.2016.01.032
11. Andrews J, Guyatt G, Oxman AD, et al. GRADE guidelines: 14. Going from evidence to recommendations: the significance and pre-

Annals of Internal Medicine • Vol. 172, No. 1 • 7 January 2020 53



## CLINICAL GUIDELINE

## A Synopsis of the European Breast Guidelines

- sentation of recommendations. *J Clin Epidemiol*. 2013;66:719-25. [PMID: 23312392] doi:10.1016/j.jclinepi.2012.03.013
12. Bond M, Pavey T, Welch K, et al. Systematic review of the psychological consequences of false-positive screening mammograms. *Health Technol Assess*. 2013;17:1-170, v-vi. [PMID: 23540978] doi:10.3310/hta17130
13. Brett J, Bankhead C, Henderson B, et al. The psychological impact of mammographic screening. A systematic review. *Psychooncology*. 2005;14:917-38. [PMID: 15786514]
14. Salz T, Richman AR, Brewer NT. Meta-analyses of the effect of false-positive mammograms on generic and specific psychosocial outcomes. *Psychooncology*. 2010;19:1026-34. [PMID: 20882572] doi:10.1002/pon.1676
15. Mathioudakis AG, Salakari M, Pyllkanen L, et al. Systematic review on women's values and preferences concerning breast cancer screening and diagnostic services. *Psychooncology*. 2019;28:939-947. [PMID: 30812068] doi:10.1002/pon.5041
16. Miller AB, Baines CJ, To T, et al. Canadian national breast screening study: 1. breast cancer detection and death rates among women aged 40 to 49 years. *CMAJ*. 1992;147:1459-76. [PMID: 1423087]
17. Bjurstam NG, Bjørneld LM, Duffy SW. Updated results of the gothenburg trial of mammographic screening. *Cancer*. 2016;122:1832-5. [PMID: 27061821] doi:10.1002/cncr.29975
18. Moss SM, Wale C, Smith R, et al. Effect of mammographic screening from age 40 years on breast cancer mortality in the UK Age trial at 17 years' follow-up: a randomised controlled trial. *Lancet Oncol*. 2015;16:1123-1132. [PMID: 26206144] doi:10.1016/S1470-2045(15)00128-X
19. Nyström L, Andersson I, Bjurstam N, et al. Long-term effects of mammography screening: updated overview of the Swedish randomised trials. *Lancet*. 2002;359:909-19. [PMID: 11918907]
20. Shapiro S. Periodic screening for breast cancer: the HIP randomized controlled trial. *health insurance plan. J Natl Cancer Inst Monogr*. 1997;27-30. [PMID: 9709271]
21. Tabar L, Duffy SW, Yen MF, et al. All-cause mortality among breast cancer patients in a screening trial: support for breast cancer mortality as an end point. *J Med Screen*. 2002;9:159-62. [PMID: 12518005]
22. Habbema JD, van Oortmarssen GJ, van Putten DJ, et al. Age-specific reduction in breast cancer mortality by screening: an analysis of the results of the Health Insurance Plan of Greater New York study. *J Natl Cancer Inst*. 1986;77:317-20. [PMID: 3461193]
23. Hofvind S, Ponti A, Patnick J, et al. EUNICE Project and Euroscreen Working Groups. False-positive results in mammographic screening for breast cancer in Europe: a literature review and survey of service screening programmes. *J Med Screen*. 2012;19 Suppl 1:57-66. [PMID: 22972811]
24. Bjurstam N, Bjørneld L, Warwick J, et al. The Gothenburg breast screening trial. *Cancer*. 2003;97:2387-96. [PMID: 12733136]
25. Tabar L, Fagerberg G, Chen HH, et al. Efficacy of breast cancer screening by age. New results from the Swedish Two-County Trial. *Cancer*. 1995;75:2507-17. [PMID: 7736395]
26. Chu KC, Smart CR, Tarone RE. Analysis of breast cancer mortality and stage distribution by age for the Health Insurance Plan clinical trial. *J Natl Cancer Inst*. 1988;80:1125-32. [PMID: 3411625]
27. Miller AB, To T, Baines CJ, et al. The Canadian National Breast Screening Study-1: breast cancer mortality after 11 to 16 years of follow-up. A randomized screening trial of mammography in women age 40 to 49 years. *Ann Intern Med*. 2002;137:305-12. [PMID: 12204013]
28. Moss S, Waller M, Anderson TJ, et al; Trial Management Group. Randomised controlled trial of mammographic screening in women from age 40: predicted mortality based on surrogate outcome measures. *Br J Cancer*. 2005;92:955-60. [PMID: 15726103]
29. Miller AB. The costs and benefits of breast cancer screening. *Am J Prev Med*. 1993 May-Jun;9:175-80. [PMID: 8347369]
30. Andersson I, Aspegren K, Janzon L, et al. Mammographic screening and mortality from breast cancer: the Malmö mammographic screening trial. *BMJ*. 1988;297:943-8. [PMID: 3142562]
31. Tabar L, Chen HH, Duffy SW, et al. Primary and adjuvant therapy, prognostic factors and survival in 1053 breast cancers diagnosed in a trial of mammography screening. *Jpn J Clin Oncol*. 1999;29:608-16. [PMID: 10721943]
32. Frisell J. Mammographic screening for breast cancer [thesis]. Stockholm: Stockholm University; 1989.
33. Hellquist BN, Duffy SW, Abdsaleh S, et al. Effectiveness of population-based service screening with mammography for women ages 40 to 49 years: evaluation of the Swedish Mammography Screening in Young Women (SCRY) cohort. *Cancer*. 2011;117:714-22. [PMID: 20882563] doi:10.1002/cncr.25650
34. Sankatsing VD, Heijnsdijk EA, van Luijt PA, et al. Cost-effectiveness of digital mammography screening before the age of 50 in the Netherlands. *Int J Cancer*. 2015;137:1990-9. [PMID: 25895135] doi:10.1002/ijc.29572
35. Madan J, Rawdin A, Stevenson M, et al. A rapid-response economic evaluation of the UK NHS Cancer Reform Strategy breast cancer screening program extension via a plausible bounds approach. *Value Health*. 2010 Mar-Apr;13:215-21. [PMID: 19878494] doi:10.1111/j.1524-4733.2009.00667.x
36. Miller AB, Baines CJ, To T, et al. Canadian national breast screening study: 2. breast cancer detection and death rates among women aged 50 to 59 years. *CMAJ*. 1992;147:1477-88. [PMID: 1423088]
37. Miller AB, To T, Baines CJ, et al. Canadian National Breast Screening Study-2: 13-year results of a randomized trial in women aged 50-59 years. *J Natl Cancer Inst*. 2000;92:1490-9. [PMID: 10995804]
38. Zackrisson S, Andersson I, Janzon L, et al. Rate of over-diagnosis of breast cancer 15 years after end of Malmö mammographic screening trial: follow-up study. *BMJ*. 2006;332:689-92. [PMID: 16517548]
39. Pharoah PD, Sewell B, Fitzsimmons D, et al. Cost effectiveness of the NHS breast screening programme: life table model. *BMJ*. 2013;346:f2618. [PMID: 23661112] doi:10.1136/bmj.f2618
40. Carles M, Vilapriyo E, Cots F, et al. Cost-effectiveness of early detection of breast cancer in Catalonia (Spain). *BMC Cancer*. 2011;11:192. [PMID: 21605383] doi:10.1186/1471-2407-11-192
41. Rojnik K, Naversnik K, Mateovic-Rojnik T, et al. Probabilistic cost-effectiveness modeling of different breast cancer screening policies in Slovenia. *Value Health*. 2008 Mar-Apr;11:139-48. [PMID: 18380626] doi:10.1111/j.1524-4733.2007.00223.x
42. Breast Screening Frequency Trial Group. The frequency of breast cancer screening: results from the UKCCCR randomised trial. United Kingdom co-ordinating committee on cancer research. *Eur J Cancer*. 2002;38:1458-64. [PMID: 12110490]
43. O'Meara ES, Zhu W, Hubbard RA, et al. Mammographic screening interval in relation to tumor characteristics and false-positive risk by race/ethnicity and age. *Cancer*. 2013;119:3959-67. [PMID: 24037812] doi:10.1002/cncr.28310
44. Vilapriyo E, Forné C, Carles M, et al; Interval Cancer (INCA) Study Group. Cost-effectiveness and harm-benefit analyses of risk-based screening strategies for breast cancer. *PLoS One*. 2014;9:e86858. [PMID: 24498285] doi:10.1371/journal.pone.0086858
45. Yaffe MJ, Mittmann N, Lee P, et al. Clinical outcomes of modeling mammography screening strategies. *Health Rep*. 2015;26:9-15. [PMID: 26676234]
46. Kerlikowske K, Zhu W, Hubbard RA, et al; Breast Cancer Surveillance Consortium. Outcomes of screening mammography by frequency, breast density, and postmenopausal hormone therapy. *JAMA Intern Med*. 2013;173:807-16. [PMID: 23552817] doi:10.1001/jamainternmed.2013.307
47. Miller AB, Wall C, Baines CJ, et al. Twenty five year follow-up for breast cancer incidence and mortality of the Canadian National Breast Screening Study: randomised screening trial. *BMJ*. 2014;348:g366. [PMID: 24519768] doi:10.1136/bmj.g366
48. Tabar L, Vitak B, Chen HH, et al. The Swedish Two-County Trial twenty years later. Updated mortality results and new insights from long-term follow-up. *Radiol Clin North Am*. 2000;38:625-51. [PMID: 10943268]



## A Synopsis of the European Breast Guidelines

## CLINICAL GUIDELINE

49. Mandelblatt JS, Stout NK, Schechter CB, et al. Collaborative modeling of the benefits and harms associated with different U.S. breast cancer screening strategies. *Ann Intern Med*. 2016;164:215-25. [PMID: 26756606] doi:10.7326/M15-1536
50. Miglioretti DL, Lange J, van den Broek JJ, et al. Radiation-induced breast cancer incidence and mortality from digital mammography screening: a modeling study. *Ann Intern Med*. 2016;164:205-14. [PMID: 26756460] doi:10.7326/M15-1241
51. Miglioretti DL, Zhu W, Kerlikowske K, et al. Breast Cancer Surveillance Consortium. Breast tumor prognostic characteristics and biennial vs annual mammography, age, and menopausal status. *JAMA Oncol*. 2015;1:1069-77. [PMID: 26501844] doi:10.1001/jamaoncol.2015.3084
52. Braithwaite D, Zhu W, Hubbard RA, et al. Breast Cancer Surveillance Consortium. Screening outcomes in older US women undergoing multiple mammograms in community practice: does interval, age, or comorbidity score affect tumor characteristics or false positive rates? *J Natl Cancer Inst*. 2013;105:334-41. [PMID: 23385442] doi:10.1093/jnci/djs645
53. Hunt KA, Rosen EL, Sickles EA. Outcome analysis for women undergoing annual versus biennial screening mammography: a review of 24,211 examinations. *AJR Am J Roentgenol*. 1999;173:285-9. [PMID: 10430120]
54. Dittus K, Geller B, Weaver DL, et al. Breast Cancer Surveillance Consortium. Impact of mammography screening interval on breast cancer diagnosis by menopausal status and BMI. *J Gen Intern Med*. 2013;28:1454-62. [PMID: 23760741] doi:10.1007/s11606-013-2507-0
55. Romero Martin S, Raya Povedano JL, Cara García M, et al. Prospective study aiming to compare 2D mammography and tomosynthesis + synthesized mammography in terms of cancer detection and recall. From double reading of 2D mammography to single reading of tomosynthesis. *Eur Radiol*. 2018;28:2484-2491. [PMID: 29294150] doi:10.1007/s00330-017-5219-8
56. Skaane P, Bandos AI, Gullien R, et al. Comparison of digital mammography alone and digital mammography plus tomosynthesis in a population-based screening program. *Radiology*. 2013;267:47-56. [PMID: 23297332] doi:10.1148/radiol.12121373
57. Bernardi D, Macaskill P, Pellegrini M, et al. Breast cancer screening with tomosynthesis (3D mammography) with acquired or synthetic 2D mammography compared with 2D mammography alone (STORM-2): a population-based prospective study. *Lancet Oncol*. 2016;17:1105-1113. [PMID: 27345635] doi:10.1016/S1470-2045(16)30101-2
58. Ciatto S, Houssami N, Bernardi D, et al. Integration of 3D digital mammography with tomosynthesis for population breast-cancer screening (STORM): a prospective comparison study. *Lancet Oncol*. 2013;14:583-9. [PMID: 23623721] doi:10.1016/S1470-2045(13)70134-7
59. Wallis MG, Moa E, Zanca F, et al. Two-view and single-view tomosynthesis versus full-field digital mammography: high-resolution X-ray imaging observer study. *Radiology*. 2012;262:788-96. [PMID: 22274840] doi:10.1148/radiol.11103514
60. Paulis LE, Lobbes MB, Lalji UC, et al. Radiation exposure of digital breast tomosynthesis using an antiscatter grid compared with full-field digital mammography. *Invest Radiol*. 2015;50:679-85. [PMID: 26011823] doi:10.1097/RLI.0000000000000168
61. Hofvind S, Hovda T, Holen AS, et al. Digital breast tomosynthesis and synthetic 2D mammography versus digital mammography: evaluation in a population-based screening program. *Radiology*. 2018;287:787-794. [PMID: 29494322] doi:10.1148/radiol.2018171361
62. Zackrisson S, Lång K, Rosso A, et al. One-view breast tomosynthesis versus two-view mammography in the Malmö Breast Tomosynthesis Screening Trial (MBTST): a prospective, population-based, diagnostic accuracy study. *Lancet Oncol*. 2018;19:1493-1503. [PMID: 30322817] doi:10.1016/S1470-2045(18)30521-7
63. Bahl M, Gaffney S, McCarthy AM, et al. Breast cancer characteristics associated with 2D digital mammography versus digital breast tomosynthesis for screening-detected and interval cancers. *Radiology*. 2018;287:49-57. [PMID: 29272213] doi:10.1148/radiol.2017171148
64. Pattacini P, Nitrosi A, Giorgi Rossi P, et al. RETomo Working Group. Digital mammography versus digital mammography plus tomosynthesis for breast cancer screening: the Reggio Emilia tomosynthesis randomized trial. *Radiology*. 2018;288:375-385. [PMID: 29869961] doi:10.1148/radiol.2018172119
65. Skaane P, Bandos AI, Eben EB, et al. Two-view digital breast tomosynthesis screening with synthetically reconstructed projection images: comparison with digital breast tomosynthesis with full-field digital mammographic images. *Radiology*. 2014;271:655-63. [PMID: 24484063] doi:10.1148/radiol.13131391
66. Skaane P, Bandos AI, Gullien R, et al. Prospective trial comparing full-field digital mammography (FFDM) versus combined FFDM and tomosynthesis in a population-based screening programme using independent double reading with arbitration. *Eur Radiol*. 2013;23:2061-71. [PMID: 23553585] doi:10.1007/s00330-013-2820-3
67. Lång K, Andersson I, Rosso A, et al. Performance of one-view breast tomosynthesis as a stand-alone breast cancer screening modality: results from the Malmö Breast Tomosynthesis Screening Trial, a population-based study. *Eur Radiol*. 2016;26:184-90. [PMID: 25929946] doi:10.1007/s00330-015-3803-3
68. Lång K, Nergård M, Andersson I, et al. False positives in breast cancer screening with one-view breast tomosynthesis: An analysis of findings leading to recall, work-up and biopsy rates in the Malmö Breast Tomosynthesis Screening Trial. *Eur Radiol*. 2016;26:3899-3907. [PMID: 26943342]
69. Houssami N, Macaskill P, Bernardi D, et al. Breast screening using 2D-mammography or integrating digital breast tomosynthesis (3D-mammography) for single-reading or double-reading—evidence to guide future screening strategies. *Eur J Cancer*. 2014;50:1799-1807. [PMID: 24746887] doi:10.1016/j.ejca.2014.03.017
70. Houssami N, Bernardi D, Caumo F, et al. Interval breast cancers in the screening with tomosynthesis or standard mammography (STORM) population-based trial. *Breast*. 2018;38:150-153. [PMID: 29328943] doi:10.1016/j.breast.2018.01.002
71. Skaane P, Sebuodegård S, Bandos AI, et al. Performance of breast cancer screening using digital breast tomosynthesis: results from the prospective population-based Oslo Tomosynthesis Screening Trial. *Breast Cancer Res Treat*. 2018;169:489-496. [PMID: 29429017] doi:10.1007/s10549-018-4705-2
72. Gilbert FJ, Tucker L, Gillan MG, et al. Accuracy of digital breast tomosynthesis for depicting breast cancer subgroups in a UK retrospective reading study (TOMMY trial). *Radiology*. 2015;277:697-706. [PMID: 26176654] doi:10.1148/radiol.2015142566
73. Bernardi D, Ciatto S, Pellegrini M, et al. Application of breast tomosynthesis in screening: incremental effect on mammography acquisition and reading time. *Br J Radiol*. 2012;85:e1174-8. [PMID: 23175484] doi:10.1259/bjr/19385909
74. Kelly KM, Dean J, Comulada WS, et al. Breast cancer detection using automated whole breast ultrasound and mammography in radiographically dense breasts. *Eur Radiol*. 2010;20:734-42. [PMID: 19727744] doi:10.1007/s00330-009-1588-y
75. Brem RF, Tabár L, Duffy SW, et al. Assessing improvement in detection of breast cancer with three-dimensional automated breast US in women with dense breast tissue: the SonoInsight Study. *Radiology*. 2015;274:663-73. [PMID: 25329763] doi:10.1148/radiol.14132832
76. Giuliano V, Giuliano C. Improved breast cancer detection in asymptomatic women using 3D-automated breast ultrasound in mammographically dense breasts. *Clin Imaging*. 2013 May-Jun;37:480-6. [PMID: 23116728] doi:10.1016/j.clinimag.2012.09.018
77. Ohuchi N, Suzuki A, Sobue T, et al. J-START investigator groups. Sensitivity and specificity of mammography and adjunctive ultrasonography to screen for breast cancer in the Japan Strategic Anti-cancer Randomized Trial (J-START): a randomised controlled trial. *Lancet*. 2016;387:341-348. [PMID: 26547101] doi:10.1016/S0140-6736(15)00774-6
78. De Felice C, Savelli S, Angeletti M, et al. Diagnostic utility of combined ultrasonography and mammography in the evaluation of



## CLINICAL GUIDELINE

## A Synopsis of the European Breast Guidelines

- women with mammographically dense breasts. *J Ultrasound*. 2007;10:143-51. [PMID: 23396266] doi:10.1016/j.jus.2007.05.001
79. Kolb TM, Lichy J, Newhouse JH. Comparison of the performance of screening mammography, physical examination, and breast US and evaluation of factors that influence them: an analysis of 27,825 patient evaluations. *Radiology*. 2002;225:165-75. [PMID: 12355001]
80. Korpraphong P, Limswarn P, Tangcharoensathien W, et al. Improving breast cancer detection using ultrasonography in asymptomatic women with non-fatty breast density. *Acta Radiol*. 2014;55:903-8. [PMID: 24103915] doi:10.1177/0284185113507711
81. Venturini E, Losio C, Panizza P, et al. Tailored breast cancer screening program with microdose mammography, US, and MR Imaging: short-term results of a pilot study in 40-49-year-old women. *Radiology*. 2013;268:347-55. [PMID: 23579052] doi:10.1148/radiol.13122278
82. Corsetti V, Ferrari A, Gherardi M, et al. Role of ultrasonography in detecting mammographically occult breast carcinoma in women with dense breasts. *Radiol Med*. 2006;111:440-8. [PMID: 16683089]
83. Sprague BL, Stout NK, Schechter C, et al. Benefits, harms, and cost-effectiveness of supplemental ultrasonography screening for women with dense breasts. *Ann Intern Med*. 2015;162:157-66. [PMID: 25486550] doi:10.7326/M14-0692
84. Berg WA, Zhang Z, Lehrer D, et al; ACRIN 6666 Investigators. Detection of breast cancer with addition of annual screening ultrasound or a single screening MRI to mammography in women with elevated breast cancer risk. *JAMA*. 2012;307:1394-404. [PMID: 22474203] doi:10.1001/jama.2012.388
85. Kuhl CK, Schrading S, Strobil K, et al. Abbreviated breast magnetic resonance imaging (MRI): first postcontrast subtracted images and maximum-intensity projection: a novel approach to breast cancer screening with MRI. *J Clin Oncol*. 2014;32:2304-10. [PMID: 24958821] doi:10.1200/JCO.2013.52.5386
86. Kuhl CK, Strobil K, Bieling H, et al. Supplemental breast MR imaging screening of women with average risk of breast cancer. *Radiology*. 2017;283:361-370. [PMID: 28221097] doi:10.1148/radiol.2016161444
87. Kriege M, Brekelmans CT, Boetes C, et al; Magnetic Resonance Imaging Screening Study Group. Efficacy of MRI and mammography for breast-cancer screening in women with a familial or genetic predisposition. *N Engl J Med*. 2004;351:427-37. [PMID: 15282350]
88. Chen SQ, Huang M, Shen YY, et al. Application of abbreviated protocol of magnetic resonance imaging for breast cancer screening in dense breast tissue. *Acad Radiol*. 2017;24:316-320. [PMID: 27916594] doi:10.1016/j.acra.2016.10.003
89. Whelehan P, Heywang-Kobrunner SH, Vinnicombe SJ, et al. Clinical performance of Siemens digital breast tomosynthesis versus standard supplementary mammography for the assessment of screen-detected soft-tissue abnormalities: a multi-reader study. *Clin Radiol*. 2017;72:95.e9-95.e15. [PMID: 27737763] doi:10.1016/j.crad.2016.08.011
90. Waldherr C, Cerny P, Altermatt HJ, et al. Value of one-view breast tomosynthesis versus two-view mammography in diagnostic workup of women with clinical signs and symptoms and in women recalled from screening. *AJR Am J Roentgenol*. 2013;200:226-31. [PMID: 23255766] doi:10.2214/AJR.11.8202
91. Tagliafico A, Astengo D, Cavagnetto F, et al. One-to-one comparison between digital spot compression view and digital breast tomosynthesis. *Eur Radiol*. 2012;22:539-44. [PMID: 21987214] doi:10.1007/s00330-011-2305-1
92. Poplack SP, Tosteson TD, Kogel CA, et al. Digital breast tomosynthesis: initial experience in 98 women with abnormal digital screening mammography. *AJR Am J Roentgenol*. 2007;189:616-23. [PMID: 17715109]
93. Michell MJ, Iqbal A, Wasan RK, et al. A comparison of the accuracy of film-screen mammography, full-field digital mammography, and digital breast tomosynthesis. *Clin Radiol*. 2012;67:976-81. [PMID: 22625656] doi:10.1016/j.crad.2012.03.009
94. Heywang-Kobrunner SH, Hacker A, Jansch A, et al; German Reader Team. Use of single-view digital breast tomosynthesis (DBT) and ultrasound vs. additional views and ultrasound for the assessment of screen-detected abnormalities: German multi-reader study. *Acta Radiol*. 2018;59:782-788. [PMID: 28929783] doi:10.1177/0284185117732600
95. Heywang-Kobrunner S, Jaensch A, Hacker A, et al. Value of digital breast tomosynthesis versus additional views for the assessment of screen-detected abnormalities - a first analysis. *Breast Care (Basel)*. 2017;12:92-97. [PMID: 28559765] doi:10.1159/000456649
96. Brandt KR, Craig DA, Hoskins TL, et al. Can digital breast tomosynthesis replace conventional diagnostic mammography views for screening recalls without calcifications? A comparison study in a simulated clinical setting. *AJR Am J Roentgenol*. 2013;200:291-8. [PMID: 23345348] doi:10.2214/AJR.12.8881
97. Cornford EJ, Turnbull AE, James JJ, et al. Accuracy of GE digital breast tomosynthesis vs supplementary mammographic views for diagnosis of screen-detected soft-tissue breast lesions. *Br J Radiol*. 2016;89:20150735. [PMID: 26559441] doi:10.1259/bjr.20150735
98. Woolf S, Schünemann HJ, Eccles MP, et al. Developing clinical practice guidelines: types of evidence and outcomes; values and economics, synthesis, grading, and presentation and deriving recommendations. *Implement Sci*. 2012;7:61. [PMID: 22762158] doi:10.1186/1748-5908-7-61
99. Practice bulletin no. 179 summary: breast cancer risk assessment and screening in average-risk women. *Obstet Gynecol*. 2017;130:241-243. [PMID: 28644328] doi:10.1097/AOG.0000000000002151
100. Qaseem A, Lin JS, Mustafa RA, et al; Clinical Guidelines Committee of the American College of Physicians. Screening for breast cancer in average-risk women: a guidance statement from the American College of Physicians. *Ann Intern Med*. 2019;170:547-560. [PMID: 30959525] doi:10.7326/M18-2147
101. Oeffinger KC, Fontham ET, Etzioni R, et al; American Cancer Society. Breast cancer screening for women at average risk: 2015 guideline update from the American Cancer Society. *JAMA*. 2015;314:1599-614. [PMID: 26501536] doi:10.1001/jama.2015.12783
102. Mainiero MB, Moy L, Baron P, et al; Expert Panel on Breast Imaging. ACR Appropriateness Criteria® breast cancer screening. *J Am Coll Radiol*. 2017;14:5383-5390. [PMID: 29101979] doi:10.1016/j.jacr.2017.08.044
103. National Comprehensive Cancer Network. NCCN Clinical Practice Guidelines in Oncology: Breast Cancer. Version 2.2018. Accessed at [www.nccn.org/professionals/physician\\_gls/pdf/breast\\_screening.pdf](http://www.nccn.org/professionals/physician_gls/pdf/breast_screening.pdf) on 7 June 2018.
104. World Health Organization. WHO Position Paper on Mammography Screening. Accessed at [www.who.int/cancer/publications/mammography\\_screening/en](http://www.who.int/cancer/publications/mammography_screening/en) on 20 June 2018.
105. Nelson HD, Fu R, Cantor A, et al. Effectiveness of breast cancer screening: systematic review and meta-analysis to update the 2009 U.S. Preventive Services Task Force recommendation. *Ann Intern Med*. 2016;164:244-55. [PMID: 26756588] doi:10.7326/M15-0969
106. Klarenbach S, Sims-Jones N, Lewin G, et al; Canadian Task Force on Preventive Health Care. Recommendations on screening for breast cancer in women aged 40-74 years who are not at increased risk for breast cancer. *CMAJ*. 2018;190:E1441-E1451. [PMID: 30530611] doi:10.1503/cmaj.180463
107. Schünemann HJ, Wiercioch W, Brozek J, 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. [PMID: 27713072] doi:10.1016/j.jclinepi.2016.09.009

**Current Author Addresses:** Dr. Schünemann: Michael G. De-Groote Cochrane Canada Centre and McMaster GRADE Centre, McMaster University 1280 Main Street West, Hamilton, Ontario L8N 4K1, Canada.  
 Drs. Lerda, Parmelli, and Saz-Parkinson: European Commission, Joint Research Centre, Via Enrico Fermi 2749, Ispra VA, 21027, Italy.  
 Dr. Quinn: St. Vincent's University Hospital, 96 Merrion Road, Elm Park, Dublin, D04 T6F4, Ireland.  
 Dr. Follmann: German Cancer Society, Kuno-fischer-straße 8, Berlin, 14057, Germany.  
 Drs. Alonso-Coello and Canelo-Aybar: Iberoamerican Cochrane Center, Biomedical Research Institute (IIB Sant Pau-CIBERESP), Sant Antoni Maria Claret 167, Barcelona, 8025, Spain.  
 Dr. Giorgi Rossi: Azienda Unità Sanitaria Locale-IRCCS di Reggio Emilia, Via Amendola 2, Reggio Emilia, 42122, Italy.  
 Dr. Lebeau: University Medical Center Hamburg-Eppendorf, Martinistraße 52, Hamburg, 20246, Germany.  
 Dr. Nyström: Umeå University, 90 187 Umeå, Sweden.  
 Dr. Broeders: Radboud University Medical Centre, Geert Grooteplein 21, Nijmegen, 6525 EZ, the Netherlands.  
 Dr. Ioannidou-Mouzaka: Leto Gynecological-Surgical and Obstetrical Clinic, 18, Avenue Kifissias, 11526 Athens, Greece.  
 Mr. Duffy: Centre for Cancer Prevention, Queen Mary University of London, Charterhouse Square, London, EC1M 6BQ, United Kingdom.  
 Dr. Borisch: Institute of Global Health, University of Geneva, Chemin des Mines 9, Geneva, 1202, Switzerland.  
 Dr. Fitzpatrick: National Screening Service, Kings Inns House, 200 Parnell Street, Dublin, D01 A3Y8, Ireland.  
 Dr. Hofvind: Cancer Registry of Norway, Ullernchausseen 64, 0379 Oslo, Norway.  
 Dr. Castells: IMIM (Hospital del Mar Medical Research Institute), Carrer del Dr. Aiguader, 88, Barcelona, 8003, Spain.  
 Dr. Giordano: CPO Piedmont-AOU Città della Salute e della Scienza, via Cavour 31, Turin, 10131, Italy.  
 Mrs. Warman: Havyatt Lodge, Havyatt Road, Langford, North Somerset, BS40 5DD, United Kingdom.  
 Dr. Mansel: Cardiff University, The Gables, Monmouth, NP25 3PA, United Kingdom.  
 Dr. Sardanelli: Università degli Studi di Milano, Via Morandi 30, San Donato Milanese, Milan, 20097, Italy.  
 Dr. Gräwingholt: Radiologie am Theater, Neuer Platz 4, Padernborn, NRW, 33098, Germany.

**Author Contributions:** Conception and design: H.J. Schünemann, D. Lerda, Z. Saz-Parkinson.  
 Analysis and interpretation of the data: H.J. Schünemann, C. Quinn, M. Follmann, P. Alonso-Coello, P. Giorgi Rossi, A. Lebeau, M. Broeders, S.W. Duffy, P. Fitzpatrick, S. Hofvind, C. Canelo-Aybar, S. Warman, E. Parmelli, A. Gräwingholt, Z. Saz-Parkinson.  
 Drafting of the article: H.J. Schünemann, Z. Saz-Parkinson.  
 Critical revision for important intellectual content: H.J. Schünemann, C. Quinn, M. Follmann, P. Alonso-Coello, P. Giorgi Rossi, A. Lebeau, L. Nyström, M. Broeders, L. Ioannidou-Mouzaka, B. Borisch, X. Castells, C. Canelo-Aybar, S. Warman, R. Mansel, F. Sardanelli, E. Parmelli, A. Gräwingholt, L. Giordano, Z. Saz-Parkinson.  
 Final approval of the article: H.J. Schünemann, D. Lerda, C. Quinn, M. Follmann, P. Alonso-Coello, P. Giorgi Rossi, A. Lebeau, L. Nyström, M. Broeders, L. Ioannidou-Mouzaka, S.W. Duffy, B. Borisch, P. Fitzpatrick, S. Hofvind, X. Castells, L. Giordano, C. Canelo-Aybar, S. Warman, R. Mansel, F. Sardanelli, E. Parmelli, A. Gräwingholt, Z. Saz-Parkinson.  
 Provision of study materials or patients: H.J. Schünemann, D. Lerda, P. Alonso-Coello, Z. Saz-Parkinson.  
 Statistical expertise: H.J. Schünemann, P. Giorgi Rossi, L. Nyström, S.W. Duffy, C. Canelo-Aybar.  
 Administrative, technical, or logistic support: H.J. Schünemann, P. Alonso-Coello, B. Borisch, D. Lerda, E. Parmelli, Z. Saz-Parkinson.  
 Collection and assembly of data: H.J. Schünemann, P. Alonso-Coello, A. Lebeau, P. Fitzpatrick, C. Canelo-Aybar, S. Warman, Z. Saz-Parkinson.