

GRADE-Ansatz zur Beurteilung der Aussagesicherheit modellierter Evidenz

IQWiG im Dialog, Köln

June 16, 2023

Holger Schünemann, MD, MSc, PhD, FRCPC

Professor of Medicine and Clinical Epidemiology

McMaster University, Humanitas University

schuneh@mcmaster.ca

 [@schunemann_mac](https://twitter.com/schunemann_mac)



Disclosures

- No direct financial conflicts
- GRADE Working Group Co-Chair
 - GRADEpro GRADE's official app
- Cochrane Canada - Director
- Guidelines International Network – board, chair elect
- Research grants from Canadian Institutes of Health Research (CIHR, FRN VR4-172741), EC, the WHO & ASH
- Views expressed my own

GRADE working group

GIN
Guidelines
International
Network

 **Cochrane**
Canada



Objectives

- Very, very brief background on GRADE
- GRADE certainty assessment for models – concept paper on models and how we got there
- My worries about rating certainty in model outputs

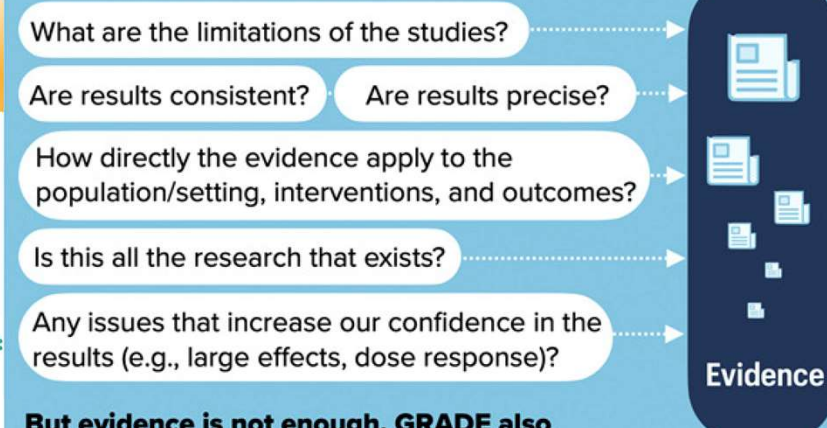
Based on perspective of user of models or mini-modeller

What is GRADE

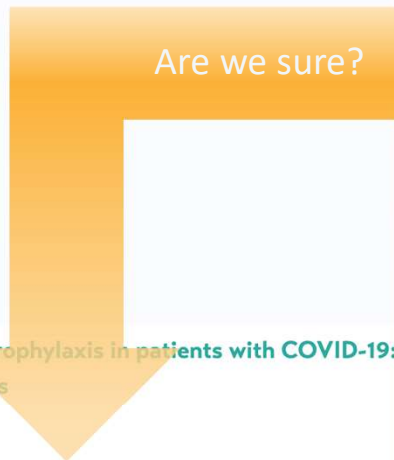
Journal of Clinical Epidemiology
Volume 127, November 2020, Pages 202-207



GRADE is a transparent and structured approach that helps us judge how certain we are about a body of evidence by addressing questions such as:



But evidence is not enough. GRADE also offers a framework to help reaching decisions and making recommendations, by evaluating:



Certainty of evidence
⊕○○○ Very low

Recommendation strength
✔ conditional



American Society of Hematology living guidelines on the use of anticoagulation for thromboprophylaxis in patients with COVID-19: January 2022 update on the use of therapeutic- intensity anticoagulation in acutely ill patients
Source: American Society of Hematology (ASH)


Intent: Treatment and rehabilitation
The ASH (American Society of Hematology) guideline panel suggests using therapeutic-intensity over prophylactic-intensity anticoagulation for patients with COVID-19-related acute illness who do not have suspected or confirmed VTE or another indication for anticoagulation.

More desirable than undesirable consequences?

Summary of findings table and certainty rating

Should any DOAC vs. LMWH be used for for VTE prophylaxis in acutely ill hospitalized medical patients?

Author(s) of the Evidence Profile: Ignacio Neumann, Juan Jose Yepes-Nuñez, Wojtek Wiercioch, Holger Schünemann

GRADE evidence profile Summary of Findings table Evidence to Decision framework Interactive Summary of Findings			
Outcomes	Absolute Effect With LMWH With any DOAC	Relative effect (95% CI)	Certainty of the evidence GRADE
▶ Mortality Follow-up: 10 days			
▶ Pulmonary Embolism – representing the moderate marker state Follow-up: 10 days			
▶ Proximal Deep Vein Thrombosis – representing the moderate marker state Follow-up: 10 days			
▶ Distal Deep Vein Thrombosis – representing the moderate distal DVT marker state Follow-up: 10 days			
▼ Major bleeding Follow-up: 10 days <input checked="" type="radio"/> Study population <input type="radio"/> Moderate	<div style="display: flex; justify-content: space-around; align-items: center;"> <div style="text-align: center;">2 per 1000</div> <div style="text-align: center;">3 per 1000</div> <div style="text-align: center;">  </div> </div> <p style="text-align: center;"> Difference: 1 more per 1000 patients (95% CI: 0 to 4 more per 1000 patients) Based on data from 21821 patients in 3 studies </p>	<div style="text-align: center;"> RR 1.7 (1.02 to 2.82) </div>	<div style="text-align: center;"> ⊕⊕⊕⊕ HIGH </div>
▶ Gastrointestinal bleeding Follow-up: 10 days			
▶ Heparin-Induced Thrombocytopenia (HIT)			

Summary of findings table and certainty rating

Should any DOAC vs. LMWH be used for for VTE prophylaxis in acutely ill hospitalized medical patients?

Author(s) of the Evidence Profile: Ignacio Neumann, Juan Jose Yepes-Nuñez, Wojtek Wiercioch, Holger Schönemann

GRADE evidence profile

Summary of Findings table

Evidence to Decision framework

Interactive Summary of Findings

Outcomes

Absolute Effect
With LMWH
With any DOAC

Relative effect
(95% CI)

Certainty of the evidence
GRADE

▶ **Mortality** Follow-up: 10 days

▶ **Pulmonary Embolism – representing the moderate marker state** Follow-up: 10 days

▶ **Proximal Deep Vein Thrombosis – representing the moderate marker state** Follow-up: 10 days

▶ **Distal Deep Vein Thrombosis – representing the moderate distal DVT marker state** Follow-up: 10 days

▼ **Major bleeding**

Follow-up: 10 days

Study population

Moderate

12
per 1000

20
per 1000



Difference: 8 more per 1000 patients
(95% CI: 0 to 22 more per 1000 patients)
Based on data from 21821 patients in 3 studies

RR 1.7
(1.02 to 2.82)

⊕⊕⊕⊕

HIGH

Same?

▶ **Gastrointestinal bleeding** Follow-up: 10 days

▶ **Heparin-Induced Thrombocytopenia (HIT)**

Model

MODEL

- **simple**

baseline risk \times relative risk reduction = risk difference

- **sophisticated**

nutritional reference values, economic, system dynamics, ...

**Certainty in modelled
evidence**

**Really concerned about
certainty of model output**



ELSEVIER



Journal of Clinical Epidemiology 129 (2021) 138–150

Journal of
Clinical
Epidemiology

GRADE SERIES

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

^aDepartment of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, Ontario, Canada

^bDepartment of Medicine, McMaster University, Hamilton, Ontario, Canada

^cMcMaster GRADE Centre & Michael DeGroot Cochrane Canada Centre, McMaster University, Hamilton, Ontario, Canada

^dDepartment 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

^eIberoamerican Cochrane Center, Biomedical Research Institute (IIB Sant Pau-CIBERESP), Barcelona, Spain

^fDepartment of Internal Medicine, American University of Beirut, Beirut, Lebanon

^gToronto Health Economics and Technology Assessment (THETA) Collaborative, Toronto, Ontario, Canada

^hNational Toxicology Program, National Institute of Environmental Health Sciences, Durham, NC, USA

ⁱDepartment of Veterinary Integrative Biosciences, Texas A&M University, College Station, TX, USA

^jSchool of Pharmacy and Chemistry, Liverpool John Moores University, Liverpool, UK

^kCenter for Evidence-Based Medicine and Health Outcome Research, Morsani College of Medicine, University of South Florida, Tampa, Florida, USA

^lInstitute for Education and Research, Hospital Moinhos de Vento, Porto Alegre, Rio Grande do Sul, Brazil

^mICF International, Durham, NC, USA

ⁿInstitute of Health & Wellbeing, University of Glasgow, Glasgow, UK

^oDepartment of Immunization, Vaccines and Biologicals, World Health Organization, Geneva, Switzerland

^pClinical Epidemiology and Medical Technology Assessment, Maastricht University Medical Centre+, Maastricht, the Netherlands

^qLaKind Associates, LLC, Catonsville, MD, USA

^rDepartment of Epidemiology and Public Health, University of Maryland School of Medicine, Baltimore, MD, USA

^sDepartment of Clinical Epidemiology, Biostatistics, and Bioinformatics, Academic Medical Center, University of Amsterdam, Amsterdam, the Netherlands

^tDepartment of Surgery, University of California Davis, Sacramento, CA, USA

^uDepartment of Medicine, Department of Veterans Affairs, Northern California Health Care System, Mather, CA, USA

^vDivision of Infection, Immunity and Respiratory Medicine, University Hospital of South Manchester, University of Manchester, Manchester, UK

^wInstitute for Evidence in Medicine, Medical Center, University of Freiburg, Freiburg-am-Breisgau, Germany

^xCochrane Germany, Freiburg-am-Breisgau, Germany

^yNational Center for Environmental Assessment, U.S. Environmental Protection Agency, Washington, DC, USA

^zDepartment of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, Ontario, Canada

^{aa}Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, Ontario, Canada

^{ab}Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, Ontario, Canada

^{ac}Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, Ontario, Canada

^{ad}Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, Ontario, Canada

^{ae}Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, Ontario, Canada

^{af}Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, Ontario, Canada

^{ag}Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, Ontario, Canada

^{ah}Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, Ontario, Canada

Main points

- simple or sophisticated – all assessments of the certainty of effects (on health outcomes and cost or resources) are based on MODELS
- certainty of outputs ← certainty of inputs & credibility of model
- 4 scenarios:
 1. develop a **new model**
 2. use off-the-shelf or adapt an **existing model**
 3. use results from **multiple existing models**
 4. forgo modelling
- Most models produce very low certainty in the output

Background

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 did not consider statistical models used to estimate the associations between measured variables (e.g., proportional hazards models or models used for meta-analysis).

Existing approaches to estimating the trustworthiness in the model output...

... lacked clarity regarding sources of uncertainty that may arise from model inputs and from the uncertainty about a model itself.

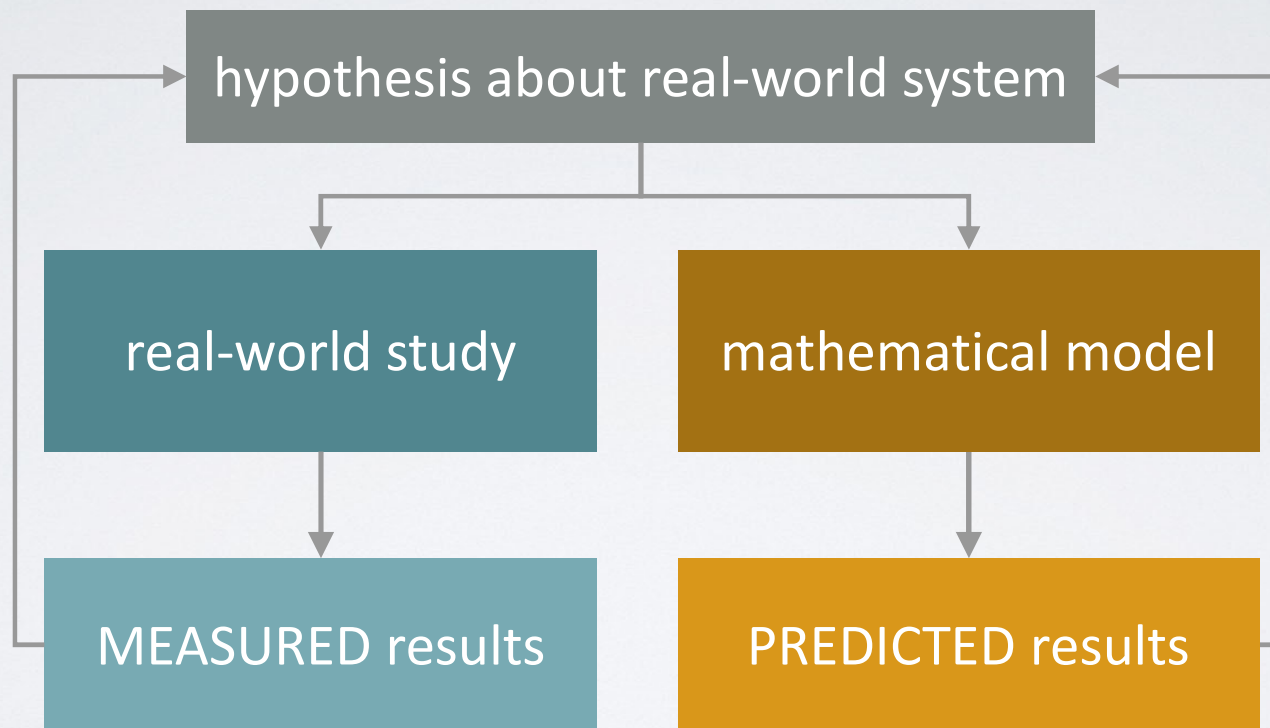
Tools to assess credibility of models...

... address only selected aspects, such as statistical reproducibility of data, the quality of reporting, or a combination of reporting with aspects of good modeling practices

Methods for the GRADE concept on models

- Multi-disciplinary across disciplines in health: economics, toxicology, pharmacology, guidelines, systematic reviews
- Review of existing methods
- Multiple workshops
- Consensus methods to arrive at conclusion

SCIENTIFIC METHOD



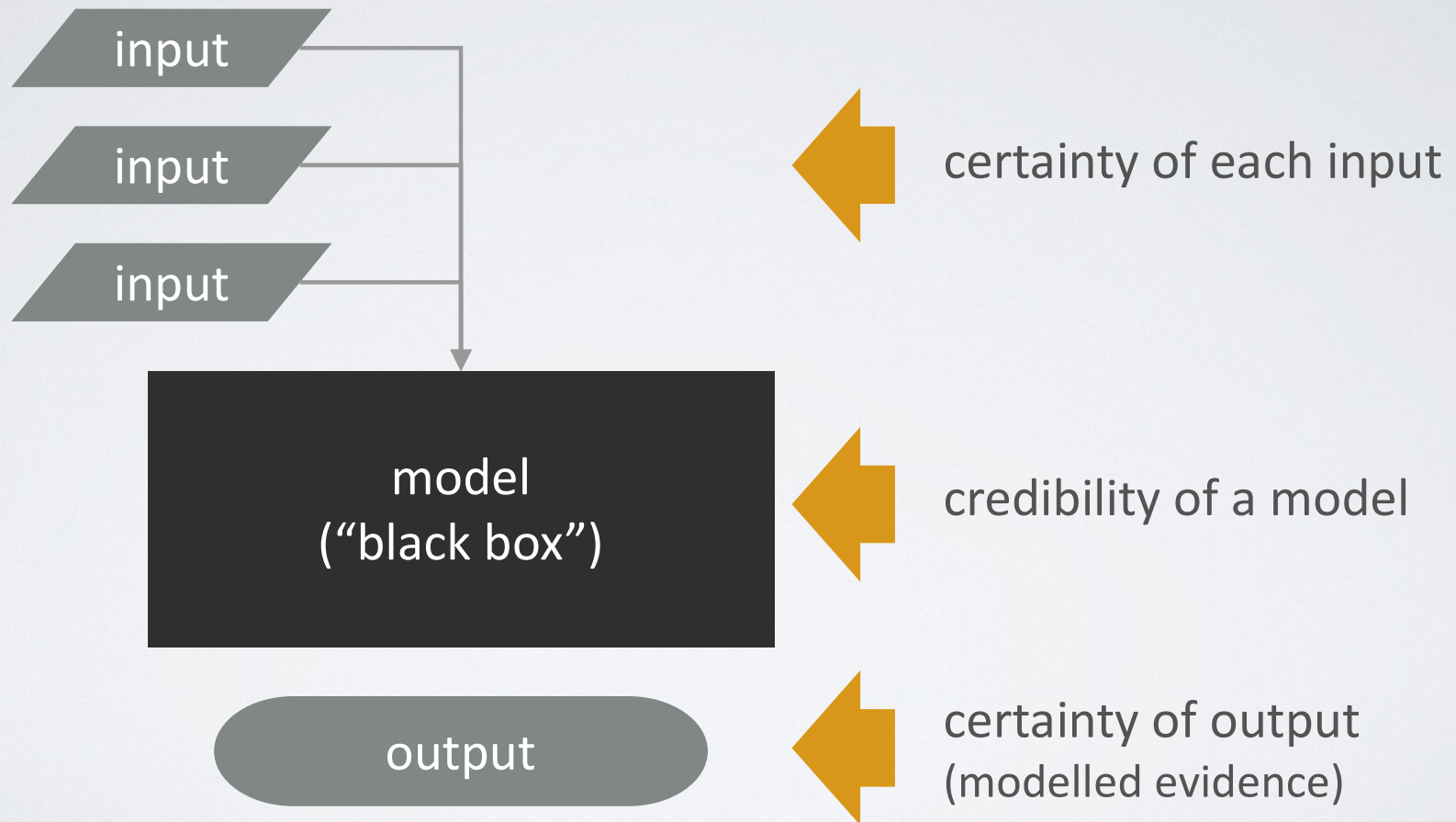


i.e. not back-of-the-envelope



FORMAL QUANTITATIVE **MATHEMATICAL** MODEL

CERTAINTY OF EVIDENCE



SOURCES OF EVIDENCE FOR A MODEL

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

Quality (may refer to many concepts, thus alternative terms are preferred to reduce confusion)

Certainty of model outputs

Alternative terms:

- certainty of modeled evidence
- quality of evidence
- quality of model output
- strength of evidence
- confidence in model outputs

Certainty of model inputs

Alternative term:

- quality of model inputs

Credibility of a model

Alternative terms:

- quality of a model
- risk of bias in a model
- validity of a model

Quality of reporting

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.

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).

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.

To avoid confusion and keep with terminology used by modeling community [7], we suggest using the term *credibility* rather than *quality* 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.

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:

- relevance
- external validity
- applicability
- generalizability
- transferability
- translatability

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.

Directness is the term used in the GRADE approach because each of the alternatives has been used usually in a narrower meaning.

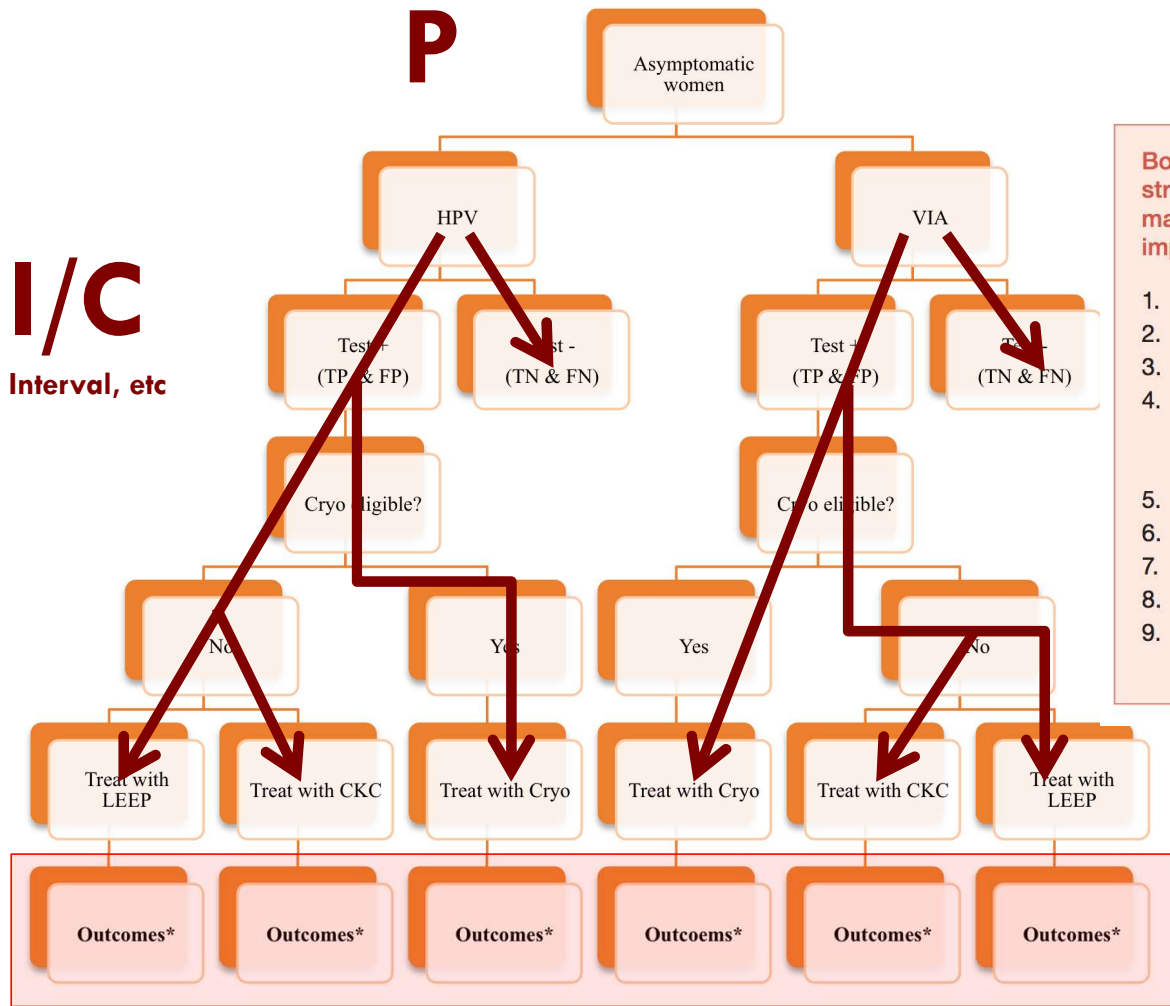
WHO guidelines

WHO guidelines for screening and treatment of precancerous lesions for cervical cancer prevention



1. Should HPV or VIA be used to screen for CIN 2+?

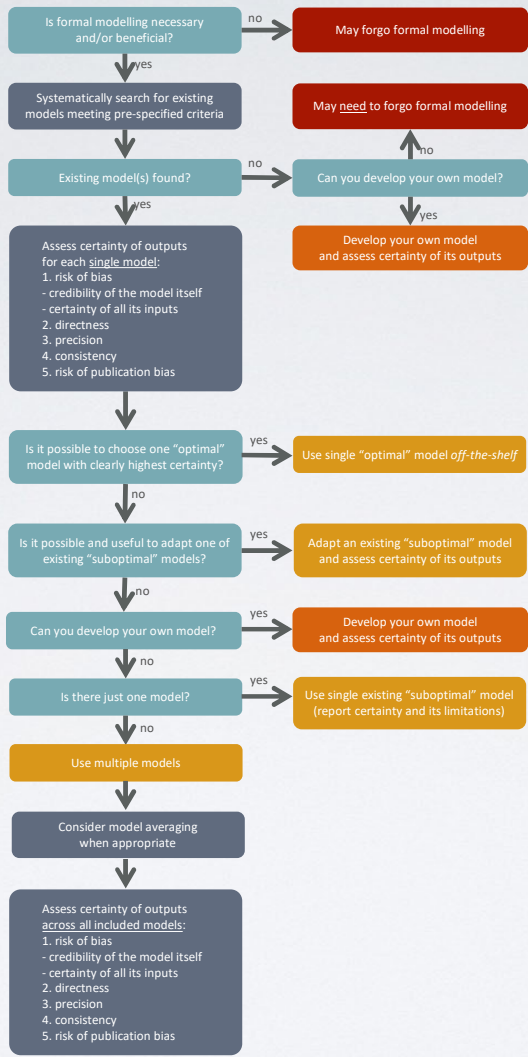
Screen – treat strategies



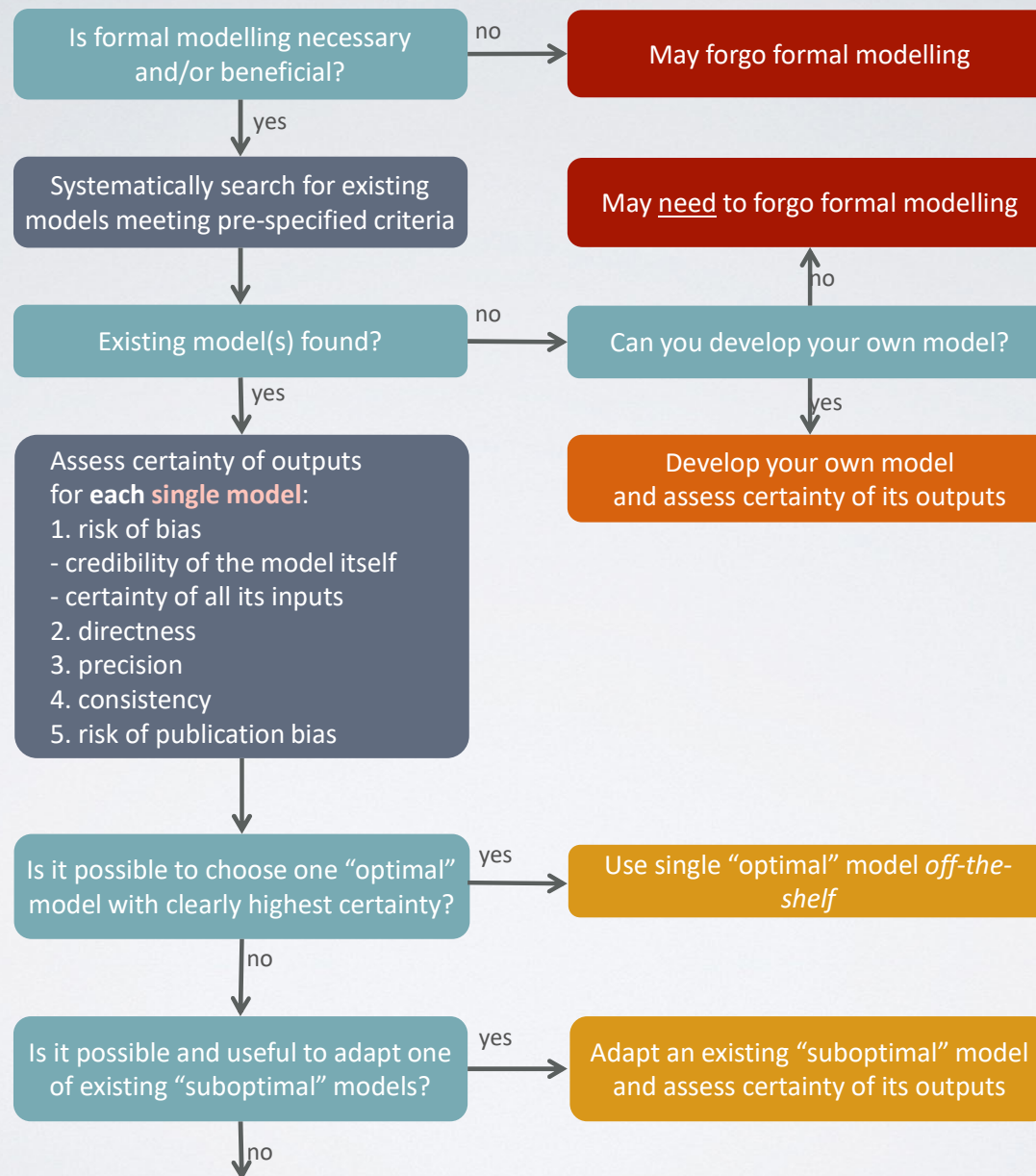
- Box 2: Outcomes for screen-and-treat strategies identified as important for making recommendations (in order of importance)**
1. Mortality from cervical cancer
 2. Cervical cancer incidence
 3. Detected CIN2, CIN3
 4. Major infections (requiring hospital admission and antibiotics, e.g. pelvic inflammatory disease)
 5. Maternal bleeding
 6. Premature delivery
 7. Fertility
 8. Identification of STIs (benefit)
 9. Minor infections (requiring outpatient treatment only)

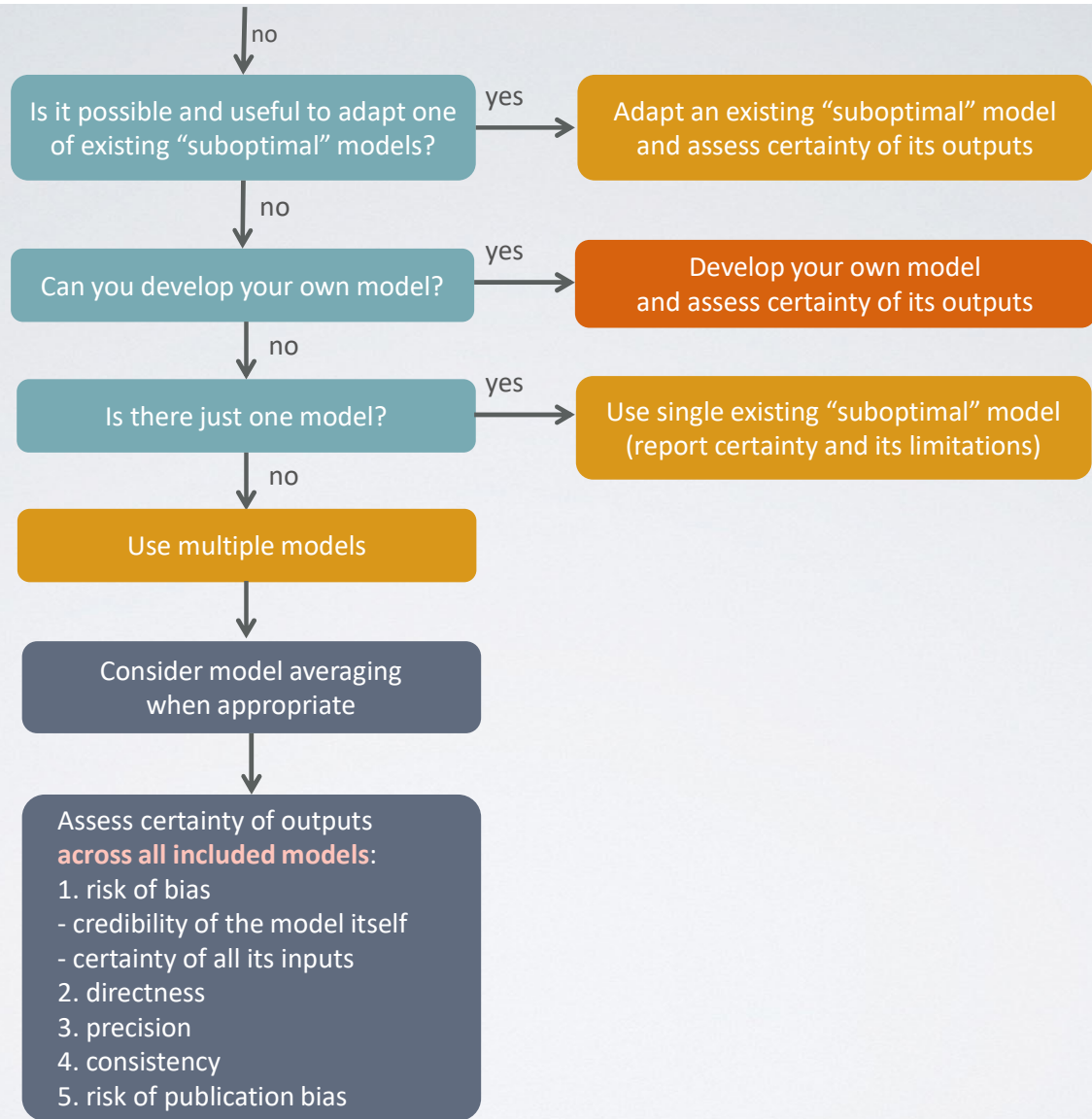


*Outcomes are: Mortality from cervical cancer, Rate of cervical cancer detection, Rate of CIN 2 & 3 detection, major bleeding, premature delivery, infertility, STI detection, major infections, and minor infections



HOW TO USE MODELS





Is it possible and useful to adapt one of existing "suboptimal" models?

yes

Adapt an existing "suboptimal" model and assess certainty of its outputs

Can you develop your own model?

yes

Develop your own model and assess certainty of its outputs

Is there just one model?

yes

Use single existing "suboptimal" model (report certainty and its limitations)

Use multiple models

Consider model averaging when appropriate

Assess certainty of outputs across all included models:

1. risk of bias
 - credibility of the model itself
 - certainty of all its inputs
2. directness
3. precision
4. consistency
5. risk of publication bias

CONCEPTUALIZATION

*Researches should start by designing a **conceptualization of the problem and the ideal target model** that would best represent the actual phenomenon they are considering*

CERTAINTY OF EVIDENCE



decrease certainty

risk of bias

indirectness

inconsistency

imprecision

publication bias



increase certainty

large effect

dose response

opposing residual
confounding

FINAL CERTAINTY ASSESSMENT

Certainty assessment							№ of patients		Effect		Certainty	Importance
№ of studies	Study design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations	annual mammography screening	triennial mammography screening	Relative (95% CI)	Absolute (95% CI)		
Breast cancer death averted												
1	modelling studies	very serious	not serious	not serious	not serious	none	90	47	--	43 more (from 40 more to 45 more) per 10.000	⊕○○○ LOW	CRITICAL
Overdiagnosis												
1	modeling studies	very serious	not serious	not serious	not serious	none	80	58	--	22 more (from 17 more to 28 more) per 10.000	⊕○○○ LOW	CRITICAL

TABLE 7 Summary of available evidence on economic impact

7.B Cost utility										
Comparison: Dupilumab vs. endoscopic sinus surgery (ESS) plus postoperative medical therapy										
Quality assessment							Summary of resources and costs			
No. of studies	Study design	Limitations	Inconsistency	Indirectness	Imprecision	Publication bias	Incremental cost per patient	Incremental effect per patient	ICER	Quality
<i>ICER per QALY</i>										
1 ¹	Cost utility, Markov model	Serious ^{a,b}	Not serious	Serious ^{c,d}	Not serious	Not serious	From 50,437\$ to 536,420 \$ (lifetime horizon)	From 9.80 to 8.95 QALYs (lifetime horizon)	Dominated by the ESS	⊕⊕○○ Low

Abbreviations: \$, US Dollar;ESS, endoscopic sinus surgery; ICER, Incremental cost-effectiveness ratio; QALY, quality-adjusted life-years.

Explanations:

^aThe study did not describe the validation/calibration procedures during Markov model development.

^bInput parameters that informed the model were of moderate certainty of evidence.

^cThe studies were performed in the USA. The results may not be applicable to other countries.

^dSome patients might have benefit from treatment of concomitant asthma or might not be willing to undergo surgical procedures. These scenarios were not included in the model.

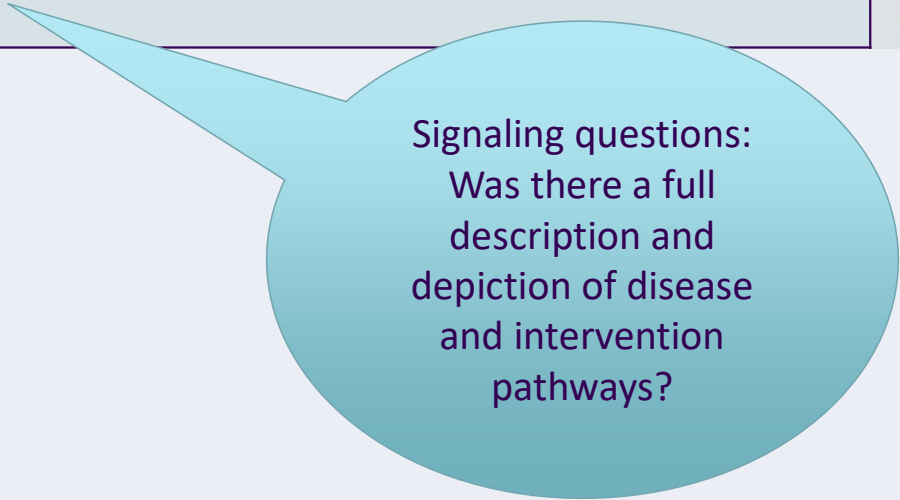
¹Scangas et al.⁵⁰

WHO guidelines

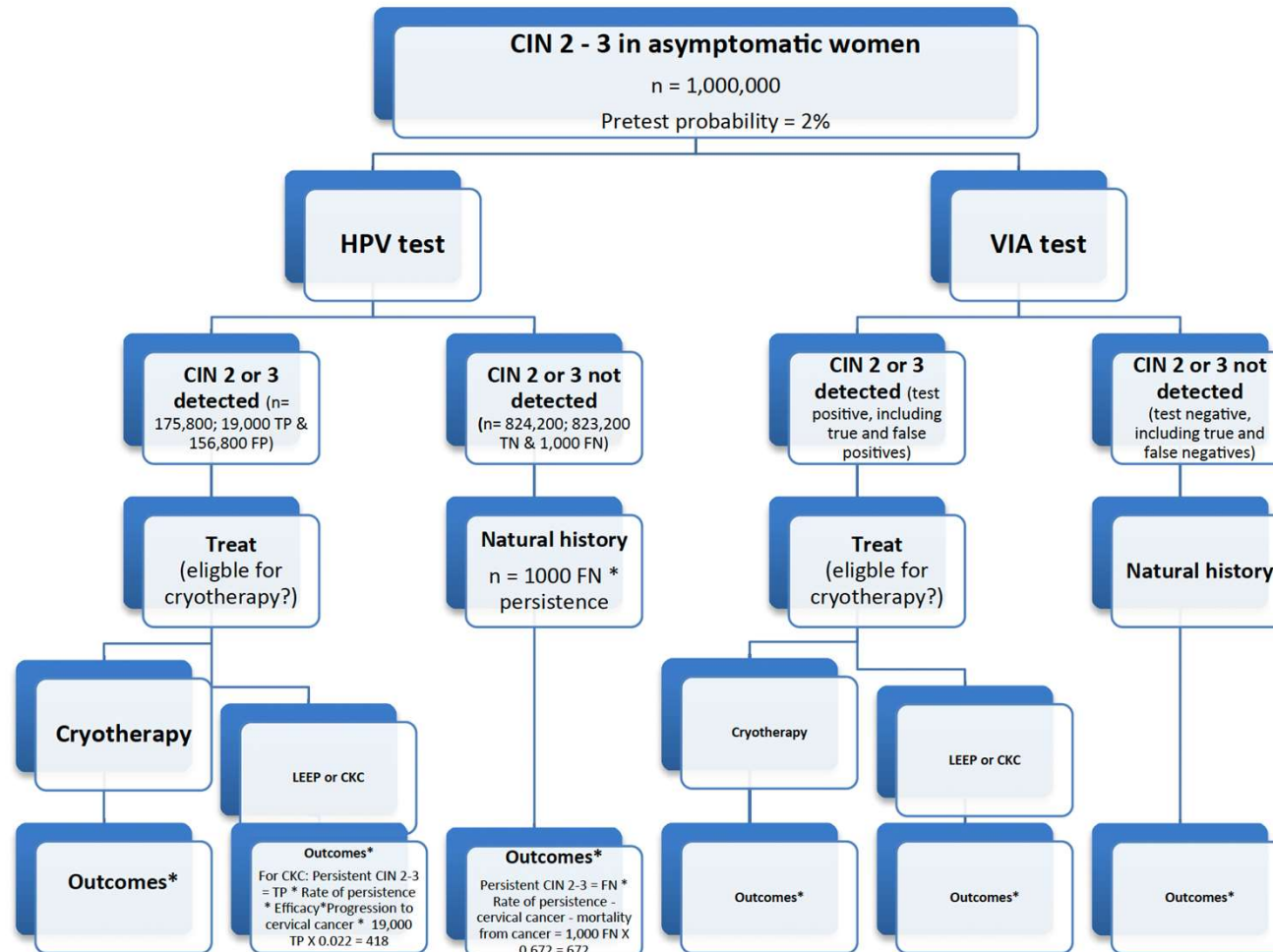
WHO guidelines for screening and treatment of precancerous lesions for cervical cancer prevention



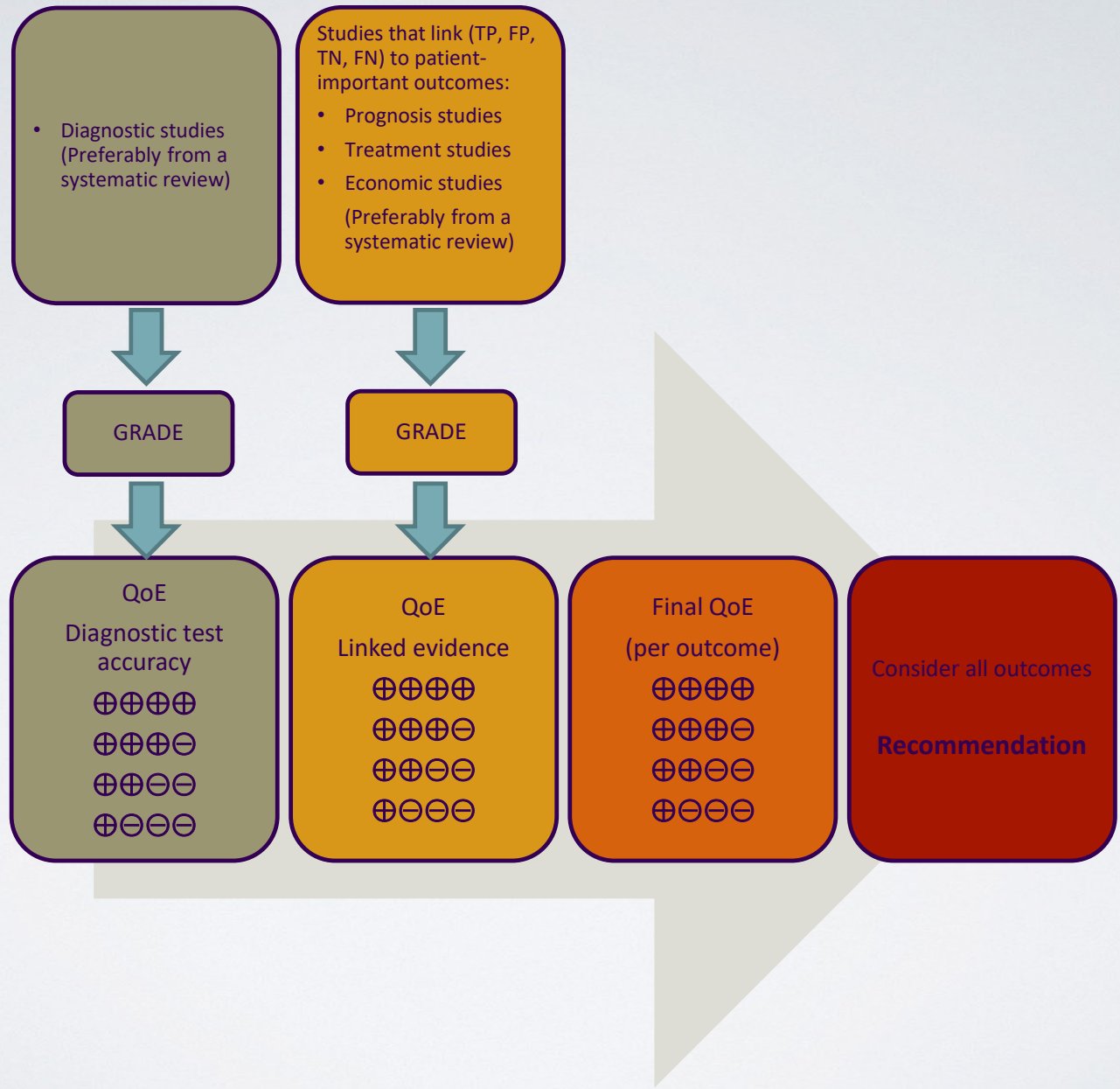
Domains of modeling requiring evaluation by guideline panel	What is being evaluated (produced)
Structure	PI(E)CO analytical framework - Graphical representation Description of model characteristics (e.g. annual vs biannual screening for cervical cancer)



Signaling questions:
Was there a full description and depiction of disease and intervention pathways?



Domains of modeling requiring evaluation by guideline panel	What is being evaluated (produced)
Structure	PICO analytical framework - Graphical representation Description of model characteristics (e.g. annual vs biannual screening)
Input	Certainty in the evidence (GRADE) in evidence profiles <ul style="list-style-type: none"> • Prognostic information • Test accuracy <ul style="list-style-type: none"> • Effects of interventions (as part of the pathways described) <ul style="list-style-type: none"> • Link(ed), indirect evidence <ul style="list-style-type: none"> • Resources • Values and preferences
Calculation	Summary of findings/evidence profiles Evidence to Decision Frameworks



SR – SENSITIVITY & SPECIFICITY

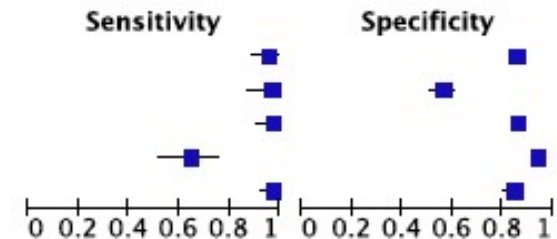
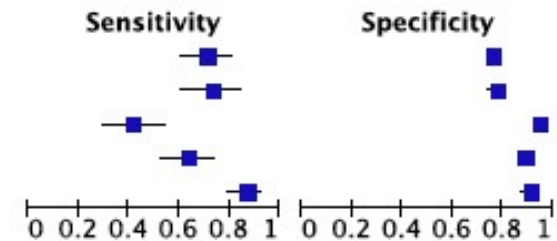
Reference standard: colposcopy plus biopsy

VIA for CIN2+

Study	TP	FP	FN	TN	Sensitivity	Specificity
Belinson 2001 VIA direct	61	461	25	1450	0.71 [0.60, 0.80]	0.76 [0.74, 0.78]
De Vuyst	44	133	16	460	0.73 [0.60, 0.84]	0.78 [0.74, 0.81]
Qiao	29	127	41	2191	0.41 [0.30, 0.54]	0.95 [0.94, 0.95]
Shastri	54	454	31	3470	0.64 [0.52, 0.74]	0.88 [0.87, 0.89]
Sodhani 2006	91	34	14	333	0.87 [0.79, 0.93]	0.91 [0.87, 0.93]

HPV (>1 pg/ml) for CIN2+

Study	TP	FP	FN	TN	Sensitivity	Specificity
Belinson 2001 VIA direct	82	282	4	1629	0.95 [0.89, 0.99]	0.85 [0.84, 0.87]
De Vuyst	52	204	2	258	0.96 [0.87, 1.00]	0.56 [0.51, 0.60]
Qiao	68	334	2	1984	0.97 [0.90, 1.00]	0.86 [0.84, 0.87]
Shastri	45	226	25	3250	0.64 [0.52, 0.75]	0.93 [0.93, 0.94]
Sodhani 2006	102	58	3	309	0.97 [0.92, 0.99]	0.84 [0.80, 0.88]



Outcome	Study design	DTA QoE	Effect per 1000 patients/year for pre-test probability of 5%		Importance
			HPV	VIA	
True positives (patients with CIN2+)	Cross sectional and cohort studies	⊕⊕⊕⊕ high	48 (42 to 49)	35 (27 to 41)	CRITICAL
TP absolute difference			13 more		
True negatives (patients without CIN2+)	Cross sectional and cohort studies	⊕⊕⊕⊖ moderate	798 (684 to 865)	827 (751 to 874)	CRITICAL
TN absolute difference			29 less		
False positives (patients incorrectly classified as having CIN2+)	Cross sectional and cohort studies	⊕⊕⊕⊖ moderate	152 (86 to 266)	123 (76 to 200)	CRITICAL
FP absolute difference			29 more		
False negatives (patients incorrectly classified as not having CIN2+)	Cross sectional and cohort studies	⊕⊕⊕⊕ high	2 (1 to 8)	15 (10 to 23)	CRITICAL
FN absolute difference			13 less		

Domains of modeling requiring evaluation by guideline panel	What is being evaluated (produced)
Structure	PICO analytical framework - Graphical representation Description of model characteristics (e.g. annual vs biannual screening)
Input	Certainty in the evidence (GRADE) in evidence profiles <ul style="list-style-type: none"> • Prognostic information • Test accuracy • Effects of interventions (as part of the pathways described) • Link(ed), indirect evidence • Resources • Values and preferences
Calculation	Summary of findings/evidence profiles Evidence to Decision Frameworks
Process	Involvement of (appropriate) panel members at relevant stages Sign off on PICO analytical framework

SUMMARY OF FINDINGS – PATIENT IMPORTANT OUTCOMES SOPHISTICATED MODEL

BENEFITS & HARMS	<p>Overall, are the anticipated desirable effects large?</p> <p>No <input type="checkbox"/> Yes <input checked="" type="checkbox"/> X Compared to no treatment</p>	<p>Summary of Findings Table based on patient-important outcomes</p> <p>HPV: sensitivity 95% (84 to 98), specificity 84% (72 to 91) VIA: sensitivity 69% (54 to 81), specificity 87% (79 to 92)</p> <table border="1"> <thead> <tr> <th rowspan="2">Outcomes</th> <th colspan="6">Events in the screen-treat strategies for patient important outcomes (numbers presented per 1,000,000 patients)*</th> </tr> <tr> <th>HPV +/- CKC</th> <th>HPV +/- LEEP</th> <th>HPV +/- Cryo</th> <th>VIA +/- CKC</th> <th>VIA +/- LEEP</th> <th>VIA +/- Cryo</th> </tr> </thead> <tbody> <tr> <td>Mortality from cervical cancer</td> <td>6</td> <td>7</td> <td>7</td> <td>9</td> <td>10</td> <td>10</td> </tr> <tr> <td>Cervical Cancer Incidence</td> <td>14</td> <td>15</td> <td>15</td> <td>19</td> <td>21</td> <td>21</td> </tr> <tr> <td>CIN2-3 recurrence</td> <td>125</td> <td>190</td> <td>166</td> <td>565</td> <td>612</td> <td>595</td> </tr> <tr> <td>Major bleeding</td> <td>16546</td> <td>2340</td> <td>117</td> <td>13071</td> <td>3250</td> <td>740</td> </tr> <tr> <td>Premature delivery</td> <td>741</td> <td>646</td> <td>625</td> <td>691</td> <td>615</td> <td>599</td> </tr> <tr> <td>Major infections</td> <td>1351</td> <td>105</td> <td>104</td> <td>1068</td> <td>97</td> <td>82</td> </tr> <tr> <td>Minor infections</td> <td>18487</td> <td>1894</td> <td>1826</td> <td>14605</td> <td>1789</td> <td>1442</td> </tr> </tbody> </table>	Outcomes	Events in the screen-treat strategies for patient important outcomes (numbers presented per 1,000,000 patients)*						HPV +/- CKC	HPV +/- LEEP	HPV +/- Cryo	VIA +/- CKC	VIA +/- LEEP	VIA +/- Cryo	Mortality from cervical cancer	6	7	7	9	10	10	Cervical Cancer Incidence	14	15	15	19	21	21	CIN2-3 recurrence	125	190	166	565	612	595	Major bleeding	16546	2340	117	13071	3250	740	Premature delivery	741	646	625	691	615	599	Major infections	1351	105	104	1068	97	82	Minor infections	18487	1894	1826	14605	1789	1442	<ul style="list-style-type: none"> The focus is on patient important outcomes that are calculated on the basis of the pretest probability, diagnostic test accuracy, presumed natural history of disease, anticipated frequency of outcomes related to the disease, treatment efficacy and reported as patient or population outcomes Agreement that should screen over no screen. Mortality and cancer the desirable effects are large. The undesirable effects are small ' except for the CKC groups.
	Outcomes			Events in the screen-treat strategies for patient important outcomes (numbers presented per 1,000,000 patients)*																																																													
			HPV +/- CKC	HPV +/- LEEP	HPV +/- Cryo	VIA +/- CKC	VIA +/- LEEP	VIA +/- Cryo																																																									
Mortality from cervical cancer	6	7	7	9	10	10																																																											
Cervical Cancer Incidence	14	15	15	19	21	21																																																											
CIN2-3 recurrence	125	190	166	565	612	595																																																											
Major bleeding	16546	2340	117	13071	3250	740																																																											
Premature delivery	741	646	625	691	615	599																																																											
Major infections	1351	105	104	1068	97	82																																																											
Minor infections	18487	1894	1826	14605	1789	1442																																																											
<p>Overall, are the anticipated undesirable effects small?</p> <p>No <input type="checkbox"/> Yes <input checked="" type="checkbox"/> X Except for CKC</p>	<p>Overall, is there certainty of the link between the diagnostic test accuracy information and the consequences?</p> <p>Very uncertain <input type="checkbox"/> uncertain <input checked="" type="checkbox"/> X Moderately certain <input type="checkbox"/> Certain <input type="checkbox"/> Very certain <input type="checkbox"/></p> <p><i>This certainty is high if there is moderate or high quality evidence indicating that treatment has clear consequences for patient important outcomes.</i></p>	<ul style="list-style-type: none"> There is uncertainty in the link because we used data from the DTA studies, treatment efficacy studies and natural progression to inform the decision analysis model and estimate the effects on patient important outcome. These studies were conducted in different populations, 																																																															

VERY LOW CERTAINTY IN PEOPLE IMPORTANT OUTCOMES

Outcomes	Events in the screen-treat strategies for patient important outcomes (numbers presented per 1,000,000 patients)*						
	HPV +/- CKC	HPV +/- LEEP	HPV +/- Cryo	VIA +/- CKC	VIA +/- LEEP	VIA +/- Cryo	NO screen ¹⁰
Mortality from cervical cancer ¹	20	30	30	81	88	88	250
Cervical Cancer Incidence ²	28	43	43	112	124	124	350
CIN2-3 recurrence ³	1088	1677	1677	4328	4762	4762	13400
Undetected CIN2-3 (FN)	1000			6000			
Major bleeding ⁴	1511	397	60	1210	318	48	0
Premature delivery ⁵	712	575	610	670	560	588	500
Infertility ⁶	-	-	-	-	-	-	0
Major infections ⁷	156	225	24	125	180	19	0
Minor infections ⁸	1649	1061	1139	1321	850	913	0
Unnecessarily treated (FP)	157000			127000			-
Cancer found at one time screening ⁹	2454			3168			-

Domains of modeling requiring evaluation by guideline panel	What is being evaluated (produced)
Structure	PICO analytical framework - Graphical representation Description of model characteristics (e.g. annual vs biannual screening) – part of Evidence to Decision (EtD) Framework
Input	Certainty in the evidence (GRADE) summarized in evidence profiles for: <ul style="list-style-type: none"> • Prognostic information • Test accuracy • Effects of interventions (as part of the pathways described) • Link(ed), indirect evidence • Resources • Values and preferences
Calculation	Summary of findings/evidence profiles EtD Frameworks
Process	Involvement of (appropriate) panel members at relevant stages Sign off on PICO analytical framework Agreement with input variables COI management Documentation EtD Frameworks (GRADE) Certainty in the evidence for the recommendation (GRADE)

SUMMARY

GRADE provides a framework for using evidence from models in health decision-making and the assessment of certainty of evidence from a model or models.

More operationalization needed!

My worry: model outputs, if carefully evaluated, will rarely result in high certainty

Thank you!

EUROPEAN BREAST GUIDELINES

The benefits and harms of population-wide mammography screening have been long debated. Modelling study to evaluate the impact of screening frequency and age range on breast cancer mortality reduction and overdiagnosis

OUTCOMES

- Breast cancer deaths (averted)
- Overdiagnosis

Keywords: mammography; breast cancer; screening; mortality reduction; overdiagnosis; Markov model

Estimating breast cancer mortality reduction and overdiagnosis due to screening for different strategies in the United Kingdom

N B Gunsoy^{*1}, M Garcia-Closas¹ and S M Moss²

¹*Division of Genetics and Epidemiology, Institute of Cancer Research, Sutton, Surrey, UK* and ²*Centre for Cancer Prevention, Queen Mary University of London, Wolfson Institute of Preventive Medicine, London, UK*

CONCEPTUALIZING (I)

The problem

- Long term diseases, several pre-clinical stages
- In-situ lesions: detection?, regression?
- Efficacy of screening exams / chemotherapy effects
- Age of screening, intervals

CONCEPTUALIZING (II)

The model

- Type of model: decision tree vs. markov model
- Subgroups of populations (pre-clinical states, breast density, risk factors)
- Horizon time: months vs. Years (lifetime)

RISK OF BIAS: CERTAINTY OF INPUTS

Supplementary Table 2 - Model parameters; estimated breast cancer incidence in the absence of screening and mortality from competing cause (per 1,000)

Age group	Invasive breast cancer incidence	<i>in situ</i> breast cancer incidence	Mortality from competing causes
	Cohort born in 1935 (Estimated)		(2008) ¹
40-44	0.96	0.07	0.97
45-49	1.45	0.09	1.49
50-54	1.66	0.08	2.50
55-59	2.02	0.10	3.83
60-64	2.54	0.14	6.06
65-69	3.24	0.16	9.95
70-74	3.86	0.22	17.09
75-79	4.22	0.27	30.87
80-84	4.58	0.33	57.68

¹ (Office for National Statistics, www.ons.gov.uk)

Supplementary Table 1 - Model parameters: uptake, sensitivity, specificity of screening

Age group	Mammography screening Sensitivity ²	Uptake at invitation ¹		
		First	Subsequent	
			attender	non-attender
	(%)	(%)	(%)	(%)
40-44	80	67	69	43
45-49	83	67	69	43
50-54	91	73	84	33
55-59	91		83	18
60-64	93		82	9
65-69	93		79	6
70-79	94		69	3

¹ (NHS Breast Screening Programme, 2012)
² (Gunsoy *et al.*, 2012, Moss *et al.*, 1993, Carney *et al.*, 2003)
(Brekelmans *et al.*, 1996, Duffy *et al.*, 1997)

RISK OF BIAS: CERTAINTY OF INPUTS

- Breast cancer incidence: surveillance data for the population of interest (quality of the register?)
- Mammography accuracy: five studies (no systematic review)

RISK OF BIAS: CERTAINTY OF INPUTS

Accuracy (no systematic review)

Among the 5 studies, 3 are empirical accuracy studies

- Risk of bias: a different reference standard for positive and negative results to the test? (biopsy vs clinical follow-up)
- Indirectness: film mammography (compared to digital mammography?)
- Inconsistency: ---
- Imprecision: no pooled data

Interval Cancers and Sensitivity in the Screening Centres of the UK Trial of Early Detection of Breast Cancer

S.M. Moss, D.A. Coleman, R. Ellman, J. Chamberlain, A.P.M. Forrest, A.E. Kirkpatrick, B.A. Thomas and J.L. Price

The incidence rates of interval cancers following a negative breast screen in two screening centres which offered women aged 45-64 annual screening by mammography and/or clinical examination are examined. Sensitivity of screening is estimated by comparing the incidence rate of interval cancers with that expected in the absence of screening, and the results are compared with those from alternative methods of calculation. The incidence rate of cancers diagnosed within 12 months of a negative screen by examination was reduced by 70% for women aged 45-54, and 84% for women aged 55-64. The incidence rate of interval cancers was substantially lower than in other studies from this that sensitivity in the UK trial was substantially lower than in other studies.

Eur J Cancer, Vol. 29A, No. 2, pp. 255-258, 1993.

Journal of Epidemiology and Community Health 1996;50:68-71

Age specific sensitivity and sojourn time in a breast cancer screening programme (DOM) in The Netherlands: a comparison of different methods

Cecile T M Brekelmans, Paul Westers, Joop A J Faber, Petra H M Peeters, Hubertine J A Collette

Individual and Combined Effects of Age, Breast Density, and Hormone Replacement Therapy Use on the Accuracy of Mammography

Patricia A. Carney, PhD; Diana L. Miglioretti, PhD; Bonnie C. Yankaskas, PhD; Karla Kerlikowsky, PhD; Carole M. Rutter, PhD; Berta M. Geller, EdD; Linn A. Abraham, MS; Steven H. Taplin, MD; Mark Ungchaisri, PhD; Gary Litterer, PhD; and Rachel Ballard-Barbash, MD, MPH

Background: The relationships among breast density, age, and use of hormone replacement therapy (HRT) in breast cancer detection have not been fully evaluated.

Objective: To determine how breast density, age, and use of HRT individually and in combination affect the accuracy of screening mammography.

Design: Prospective cohort study.

Setting: 7 population-based mammography registries in North Carolina; New Mexico; New Hampshire; Vermont; Colorado; Seattle, Washington; and San Francisco, California.

Participants: 329 495 women 40 to 89 years of age who had 463 372 screening mammograms from 1996 to 1998; 2223 women received a diagnosis of breast cancer.

Abstract

Study objective - To estimate age dependent sensitivity and sojourn time in a breast cancer screening programme by different methods.

Population and methods - The study population comprised women par-

a high tumour growth rate in young women, which implicates a short sojourn time.²⁻¹¹

There are several ways of estimating sensitivity. In a previous publication we used the so called classic method - the proportion of all cancers detected within a certain time after screening in relation to those detected at

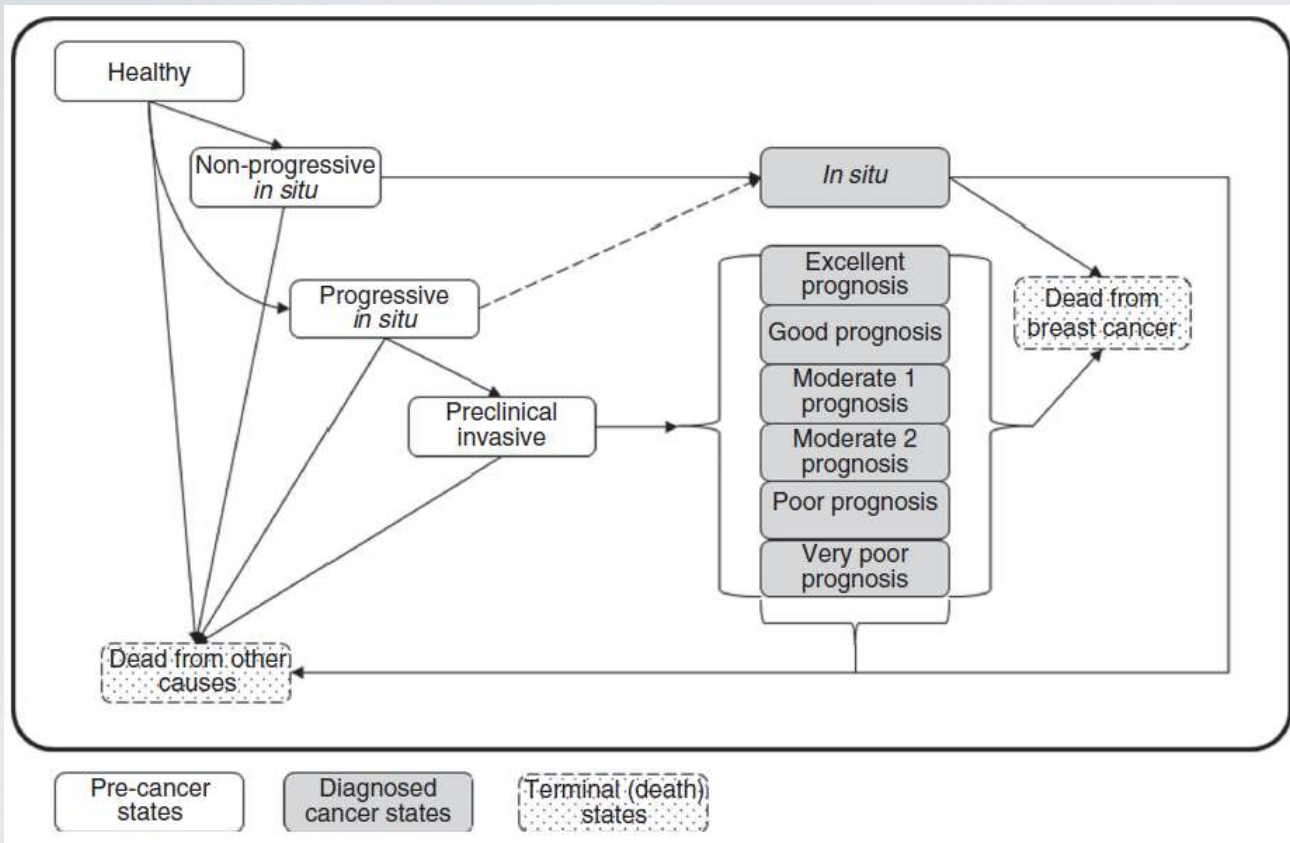
Results: Adjusted sensitivity ranged from 62.9% in women with extremely dense breasts to 87.0% in women with almost entirely fatty breasts; adjusted sensitivity increased with age from 68.6% in women 40 to 44 years of age to 83.3% in women 80 to 89 years of age. Adjusted specificity increased from 89.1% in women with extremely dense breasts to 96.9% in women with almost entirely fatty breasts. In women who did not use HRT, adjusted specificity increased from 91.4% in women 40 to 44 years of age to 94.4% in women 80 to 89 years of age. In women who used HRT, adjusted specificity was about 91.7% for all ages.

Conclusions: Mammographic breast density and age are important predictors of the accuracy of screening mammography. Although HRT use is not an independent predictor of accuracy, it probably affects accuracy by increasing breast density.

RISK OF BIAS: MODEL CREDIBILITY

Design

- Markov model (monthly cycle, several frequency scenarios)
- Assumptions: *...progressive in situ breast cancers could not be diagnosed in the absence of screening, that non-progressive in situ breast cancers were detectable clinically in the absence of screening, and that all preclinical invasive breast cancers had a mandatory progressive in situ precursor.*
- Adjuvant treatment effect for ER/HER2-specific subtypes not included
- Time horizon: 40 to 80 years



The 13-state Markov simulation model for the evaluation of mammography screening

RISK OF BIAS: MODEL CREDIBILITY (II)

Validity

- *Fit of estimated invasive and in situ breast incidence in a cohort of women born in 1935 in the UK to that observed in the UK*
- *Breast cancer and competing **mortality rates could not be validated***

Analysis

- *We performed a series of sensitivity analyses to assess the impact of higher and lower uptake of screening, and of uncertainty in estimates of sensitivity and mean sojourn time.*

RISK OF BIAS: FINAL ASSESSMENT

Certainty of the inputs + Model credibility = Risk of bias
(very low?) (concerns for validation?) (very serious?)

INDIRECTNESS

Outputs

- Breast cancer mortality
- Life years gained
- Overdiagnosis

Input data

- UK from national statistics

INCONSISTENCY

- One model study
- No further information about input inconsistency

IMPRECISION

Results (x10,000 women)

Increasing frequency from triennial to annual 47-73		
Breast cancer mortality reduction	20.3 (19.3, 21.3)	43 (40, 45)
Breast cancer overdiagnosis	2.0 (1.5, 2.5)	22 (17, 28)
Ratio (per case overdiagnosed)		1.9 (1.5, 2.6)
Abbreviation: CI= confidence interval. Numbers in this table do not match increments in Tables 3, 4, and 5 as denominators are different (value in the comparator rather than value for no screening).		

Sensitivity analysis

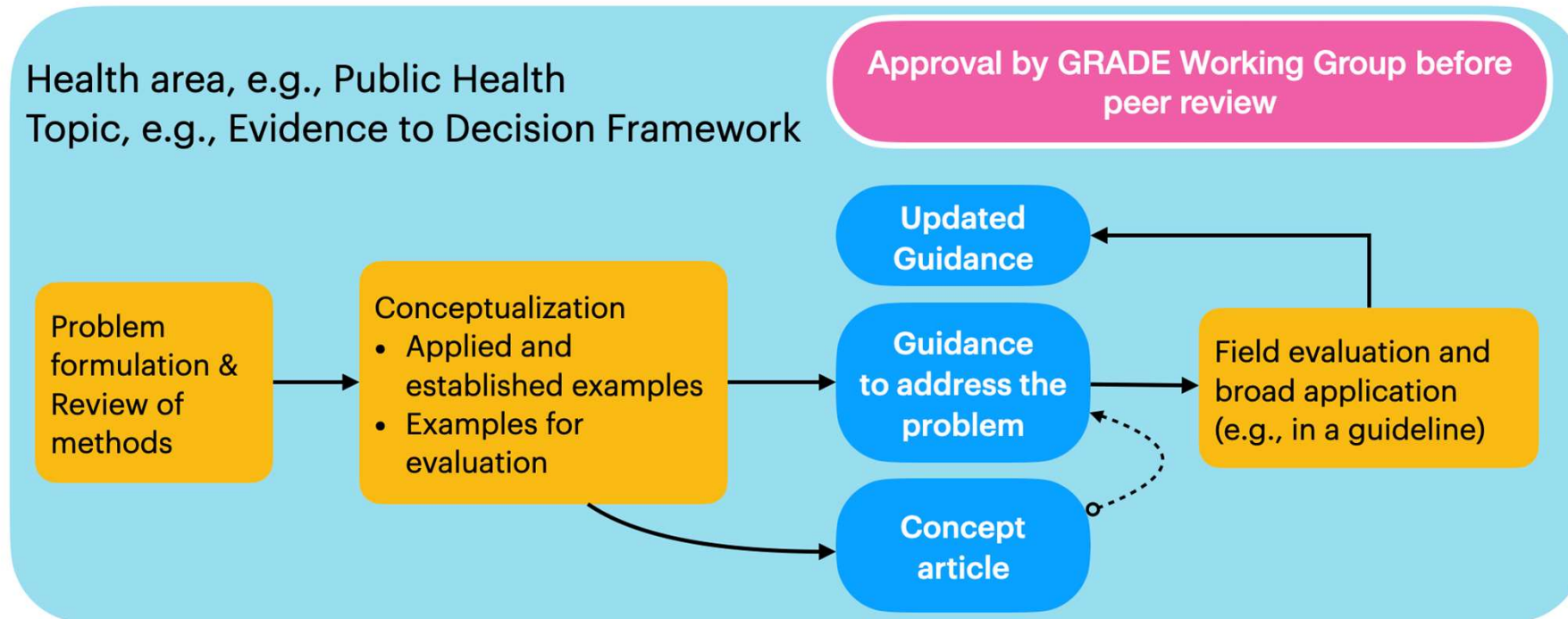
Table 6. Breast cancer mortality reduction and overdiagnosis due to screening women every three years in women aged 47-73 in a cohort of British women followed up from age 40 to 85 years. Results of sensitivity analyses.

	Breast cancer mortality reduction %	Overdiagnosis of all breast cancer %
	Mean (95% CI)	Mean (95% CI)
Base case	18.1 (17.3, 19.0)	5.6 (5.1, 6.1)
Modified uptake		
100%	24.3 (23.3, 25.2)	7.5 (6.9, 8.1)
15% decreased	12.2 (11.1, 13.3)	3.9 (3.3, 4.5)
Modified sensitivity		
100%	20.0 (19.0, 21.0)	6.2 (5.6, 6.8)
Modified mean sojourn time		
Long sojourn times ^a	18.0 (17.1, 18.9)	9.7 (9.1, 10.3)
Short sojourn time ^b	14.8 (13.9, 15.7)	3.1 (2.6, 3.5)

NEXT STEPS - FEEDBACK

- Overall certainty: very low

How **GRADE** concept and guidance articles are developed and approved



- Proof of concept
- Brainstorming
- Stakeholder feedback
- User testing
- Application in real or hypothetical situations
- Final guidance or concepts

- Rigorous, multi-stage approval process by the GRADE Working Group
- Authorship determined following ICMJE principles
- Development of tools, e.g. GRADEpro

Guideline topics, guideline questions, evidence review questions and recommendations

- A topic/module describes the general area of the guideline (the scope)
 - E.g., screening for tuberculosis (as opposed to all topics including treatment)
- Guideline question – “should” question - Population, interventions, comparison
 - E.g., should A or B be used for people with X
- Evidence (systematic) reviews – Population, interventions, comparison, outcomes (PICO) questions
 - In people with X, what is the accuracy of test A compared with test B
 - In people with X, what is the impact of A compared with B on outcomes 1, 2, 3, ...
 - What value do people with X place on outcomes 1, 2, 3, ...
 - In people with X, how cost effective is intervention A compared with B
 - In people with X, when compared to intervention B is intervention A feasible to implement
 - ...
- Recommendation - provide the answers to the “should” questions
 - In people with X, the guideline panel recommends/suggests using A rather than B