DEVELOPMENT AND USE OF HEALTH OUTCOME DESCRIPTORS: A GUIDELINE DEVELOPMENT CASE STUDY

TEJAN BALDEH

DEVELOPMENT AND USE OF HEALTH OUTCOME DESCRIPTORS: A GUIDELINE DEVELOPMENT CASE STUDY

By TEJAN BALDEH, B.SC.

A Thesis Submitted to the School of Graduate Studies in Partial Fulfillment of the Requirements for the Degree Master of Public Health

McMaster University MASTER OF PUBLIC HEALTH (2018)

Hamilton, Ontario (Health Research Methodology, Evidence, & Impact)

TITLE: Development of Health Outcome Descriptors for Outcome Importance and Utility Rating A Guideline Development Case Study

AUTHOR: Tejan Baldeh, B.SC. (Queen's University)

SUPERVISOR: Professor H.J. Schünemann

NUMBER OF PAGES: xii, 83

ABSTRACT:

OBJECTIVES: During health guideline development, panel members often have implicit, different definitions of health outcomes that can lead to variability in evidence synthesis and recommendations. McMaster GRADE Centre researchers developed a standardized description of health outcomes using the health marker state format. We aimed to determine which aspects of the development, content, and use of marker states were valuable to guideline developers.

STUDY DESIGN & SETTING: We conducted a case study of marker state development with the European Commission Initiative on Breast Cancer (ECIBC) Guidelines Development Group (GDG). Eighteen GDG members provided written and interview feedback on the process. Using the health marker states, 2 health utility rating surveys were conducted near the beginning and end of development respectively.

RESULTS: We developed 24 marker states for outcomes related to breast cancer screening and diagnosis. Feedback from GDG members revealed that marker states could be useful for developing recommendations and improving transparency of guideline methods. Comparison of the two health utility surveys showed a decrease in standard deviation in the second survey across 21 (88%) of the outcomes.

CONCLUSIONS: Health marker states are a promising method, satisfying the pre-requisite of being feasible, acceptable, and with some initial result on reduction of variance of health utility scores.

PREFACE

This thesis has been conducted as a "sandwich thesis" and consists of an individual manuscript submitted to a journal for publication. The format is as follows:

CHAPTER1: Introduction

CHAPTER 2: Manuscript 1: "Development of Health Outcome Descriptors for Outcome Importance and Utility Scores: A Guideline Development Case Study"

CHAPTER 3: Conclusion

At the time of submission my manuscript has been sent to members of the ECIBC (who are authors on this papers) for approval prior to journal submission.

ACKNOWLEDGEMENTS:

I would like to express my sincere gratitude to my advisor Professor Holger Schünemann for his continuous generosity and mentorship. He gave me the opportunity to work with him many years ago, and I would not be who I am, and where I am today without his support.

I extend sincere thanks to the members of my supervisory committee: Dr. Paola Muti, and Dr. Nancy Santesso, whose input on my project not only facilitated the writing of my thesis but improved the quality of my other research and coursework.

Next, I extend my gratitude to Dr. Zuleika Saz-Parkinson, the staff at the ECIBC and the members of the GDG. Without all their help, patience, and cooperation this study would not have been possible.

I would also like to thank my colleagues for engaging in stimulating discussions with me, taking the time to teach me the foundations of research, and most of all being fun people.

Finally, I thank my family and friends for their unequivocal support in guiding me through life, particularly as I wrote my thesis.

TABLE OF CONTENTS

CHAPTER 1: Introduction 1
1.1 Evidence Synthesis During Development of Healthcare Guidelines 1
1.2 Importance Rating Using the GRADE approach 2
1.3 Health Utility Assessment During Guideline Development 4
1.4 Defining Outcomes During Evidence Synthesis 6
1.5 Standardized Methods for Defining Health Outcomes 8
1.5.1 Guidance on Management of Attributes in Outcome Definitions 9
1.5.2 Guidance on Management of Evaluation Techniques 9
1.5.3 Guidance on Format of Outcome Definitions 10
1.6 Summary of background for thesis 10
1.7 Thesis Objectives & Rationale 11
APPENDIX 13
1.8 References 15
CHAPTER 2: Development of Health Outcome Descriptors for Outcome Importance and Utility Scores: A Guideline Development Case Study 17
Abstract ..... **Error! Bookmark not defined.**
2.1 Introduction 22
2.2 Methods 25
2.2.1 General Methods 25
2.2.2 Participants 26
2.2.3 Template of Health Marker State Descriptors 27
2.2.4 Development of Draft Health Marker State Descriptors 27
2.2.5 Refinement of Marker State Content and Structure 28
2.2.6 Online Utility Rating Surveys 30
2.2.7 Data Analysis 31
2.3 Results 32
2.3.1 Health Marker State Descriptors 32
2.3.2 ECIBC GDG Interview Feedback 34

2.3.3 Utility Rating Survey Scores 40
2.4 Discussion 42
2.4.1 Key Findings 42
2.4.2 Limitations and Strengths 44
2.4.3 Implications for practice 45
2.4.4 Implications for research 46
2.5 Conclusion 47
Acknowledgments 48
What is new? 49
What this adds to what was known? 49
What is the implication and what should change now? 49
Appendix – ECIBC Health Marker State Descriptors 50
1) **Accessibility to Information** 50
2) Awareness of Information 51
3) **Participation in Screening** 52
4) **Informed Decision Making** 53
5) **Satisfaction with Decision Making** 54
6) **Confidence with Decision Making** 55
7) **Abnormal Screening Results** 56
8) **Recall for Assessment** 57
9) **False-Positive Screening Result** 58
10) Asymmetric Density in Mammogram ..... **Error! Bookmark not defined.**
11) **Mammography Magnification Views** ..... **Error! Bookmark not defined.**
12) **Mass Lesion** ..... **Error! Bookmark not defined.**
13) **Parenchymal Architectural Distortions** ..... **Error! Bookmark not defined.**
14) **Spot Compression** ..... **Error! Bookmark not defined.**
15) **Suspicious Indeterminate Calcifications in Mammography** ..... **Error! Bookmark not defined.**
16) Needle Core Biopsy ..... **Error! Bookmark not defined.**
17) False-Positive Biopsy Result 60

18) Fine Needle Aspiration Cytology.....**Error! Bookmark not defined.**
19) Vacuum Assisted Biopsy.....**Error! Bookmark not defined.**
20) Breast Cancer Detection..... 61
21) Breast Cancer Stage..... 62
22) Determination of Tumour Biomarker Status in Biopsy..... 63
23) Interval Breast Cancer..... 64
24) **Over-diagnosis and Over-treatment**..... 65
25) **False-Negative Screening Result**..... 66
26) Radiation Exposure from Mammograms & Other Assessments Using Radiation .. 67
27) Provision of Surgical Therapy..... 68
28) Mastectomy..... 69
29) Provision of Medical Therapy .....**Error! Bookmark not defined.**
30) Provision of Radiotherapy.....**Error! Bookmark not defined.**
31) Provision of Chemotherapy .....**Error! Bookmark not defined.**
32) Radiation Exposure as Therapy.....**Error! Bookmark not defined.**
33) Other-Cause Mortality 73
2.6. References 74
CHAPTER 3: Conclusion..... 76
3.1 Summary of Findings..... 76
3.2 Implications..... 78
3.3 Limitations and Strengths of Work 79
3.4 Further Research..... 80
3.5 Final remarks..... 82
3.6 References 83

LISTS OF FIGURES AND TABLES

Figures	Page
Chapter 1	
Figure 1: GRADE Importance Scale and Hypothetical Outcome Ratings	13
Figure 2: Visual Analogue Scale with anchor outcomes used during GRADE Outcome Health Utility Assessment.	14
Chapter 2	
Figure 1: Draft Template for Development of Health Marker State Descriptors	27
Figure 2: Health Marker State Descriptor Development Process	30
Figure 3: List of Health Marker State Descriptors Developed for ECIBC	33
Figure 4: Example Health Marker State Descriptor Developed for ECIBC	33
Appendix	
Set of ECIBC Health Marker State Descriptors	50-73
Tables	**Page**
Chapter 2	
Table 1: Mean Health Utility Scores	41

LIST OF ALL ABBREVIATIONS AND SYMBOLS

ECIBC: European Commission Initiative for Breast Cancer

GDG: Guidelines Development Group

TTO: Time Trade Off

SG: Standard Gamble

VAS: Visual Analogue Scale

GRADE: Grading of Recommendations, Assessment, Development and Evaluation

JRC: Joint Research Centre

COMET: Core Outcome Measures in Effectiveness Trials

P.I.C.O.: Population, Intervention, Comparator, Outcome(s)

DECLARATION OF ACADEMIC ACHIEVEMENT

I, Tejan Baldeh, declare this thesis to be my own work. Part of this work may be submitted for publication later.

To the best of my knowledge, the content of this document does not infringe on anyone's copyright.

My supervisor, Dr. Holger Schünemann, and the members of my supervisory committee, Dr. Paola Muti and Dr. Nancy Santesso, have provided guidance and support at all stages of this project. Using their feedback, I drafted protocols, designed study materials, and collected data. Analysis of the qualitative data was completed in duplicate with Mr. Gian Paolo Morgano.

CHAPTER 1: Introduction

1.1 Evidence Synthesis During Development of Healthcare Guidelines

Healthcare guidelines aim to support healthcare professionals, recipients of care and policy makers in making best decisions for care. The primary benefit of guidelines is to generally increase the quality of care, improve consistency of care, and improve patient health outcomes. There is risk of bias at each step of guideline development [1-3]. To minimize bias in the guidelines and maximize trustworthiness of the recommendations, six principles should be followed during guideline development [1, 3-7]:

1. Involvement of multidisciplinary stakeholders
2. Recommendations supported by systematic reviews of the evidence
3. Description and consideration of important subgroups and peoples' values and preferences
4. Management of conflict of interest
5. Ratings of the certainty or quality of evidence and transparency in moving from evidence to recommendations.
6. Update and revise the guideline

Before the fifth principle can be followed, guideline developers must synthesize all available evidence. Clear descriptions of all evidence are required for its synthesis so that guideline developers can decide which evidence to extract

for consideration, identify which information is important to healthcare decision-makers, and balance the relative benefits and harms when developing recommendations [8, 9]. If the evidence is not described clearly, bias can be introduced into the guideline, thereby resulting in inaccurate assessments of the quality of evidence or strength of the recommendation(s) in question [10].

1.2 Importance Rating Using the GRADE approach

To facilitate transparency and consistency of the guideline development process, particularly as it relates to the fifth principle, the Grading of Recommendations, Assessment, Development, and Evaluation (GRADE) approach is widely used by guideline developers and agencies to systematically evaluate the quality of evidence and strength of healthcare recommendations [11]. To achieve transparency and consistency using the GRADE approach, evidence synthesis begins by defining the question and respective outcome set using the Participants, Intervention, Comparator, and Outcomes (PICO) framework [8, 12]. For each question, guideline developers decide upon an outcome set by generating a list of all relevant outcomes and rating them based upon their importance to those who would be affected by the recommendation. Guideline developers base the selection of outcome sets on importance ratings because they assume that healthcare recipients base their preference for a health intervention on the relative importance of all the outcomes incurred from that intervention compared to others [13].

Raters determine the importance of relevant health outcomes in the GRADE approach by placing the outcome on a 1 to 9 scale (1-3 = low importance for decision making, 4-6 = important, but not critical for decision making, 7-9 = critical for decision making) [8]. The scale for rating outcome importance in the GRADE approach is presented in Figure 1. Outcomes with the highest average scores across all raters (indicating that they are "important to stakeholders for decision-making) are selected for the question-specific outcome set. Later in the guideline development process, outcome sets are included in GRADE evidence tables that summarize the key information of a systematic review [14-16]. The tables support panel decision making during the formulating of recommendations by presenting relevant information in the context of the outcome set.

Overall, the GRADE importance rating exercise mitigates several challenges to guideline development. Firstly, it orients panel members to the task of considering outcomes that are important to stakeholders. Secondly, it reduces the number of outcomes deemed to be stakeholder-important, thereby increasing the efficiency of decision-making. Thirdly, importance ratings are indicative of the panel's agreement regarding the outcome-specific balance of benefits and harms. Furthermore, importance ratings identify the relative importance of balance of benefits and harms for each outcome (e.g. within the "critically important" category an outcome rated as 9 will be more important than an outcome rated as 7). Collectively, panels can use this information to inform discussion during development of a recommendation.

1.3 Health Utility Rating During Guideline Development

Health utility ratings are a separate measurement, developed from economic theory, which are used similarly to importance ratings to inform a panel's benefit-harm analysis of health outcomes [17]. In the context of guidelines, health utility is a measure of the values towards the outcomes of those affected by the health outcome [12]. Therefore, utility ratings are also indicative of the importance of an outcome.

For healthcare guidelines, health utility ratings are often unavailable [18]. When available, published cost-utility analyses can be considered as a source of health utility scores. However, scores taken from these sources are likely to be irrelevant to guideline PICOs, biased, or methodologically flawed [12, 19, 20]. In most cases, this is due to varying study populations and methods for calculating health utility among cost-utility assessments. To more accurately assess the collective views of the panel regarding the relative benefits and harms of each outcome, guideline developers sometimes rate the health utility of outcomes internally [19]. By doing this, panels ensure that the health utilities are directly informing the outcomes of interest.

Generally, there are two approaches for measuring health utility: direct preference techniques and multi-attribute techniques. Using direct preference measures, subjects compare the outcome of interest to another 'anchor' outcome [21, 22]. The von Neumann-Morgenstern Standard Gamble (SG), Time Trade Off

(TTO) and Visual Analogue Scale (VAS) are among the most common validated direct preference scaling methods, each with pros and cons regarding reliability, and bias.

The SG method includes an aspect of uncertainty, which is considered by economists to provide a truer representation of participant values [23, 24]. For any given health outcome, the SG method requires subjects to quantify the probability of experiencing the worst outcome (typically 'death') that they would be willing to accept given a reciprocal probability that they attain the best possible outcome (typically 'perfect health'). The TTO method is like the SG method, but decisions are based on time instead of probabilities, presumably making it easier to understand for participants. Participants determine how many years of full health are equivalent to living in a health state that is not full health. There is no gamble involved as the outcome is described as secure. As a result, uncertainty is not a factor of the TTO method [23, 24].

GRADE guidelines have used VAS which, unlike the SG and TTO, requires subjects to give a direct quantitative estimate of the health utility relative to the theoretical best and worst outcomes which anchor the scale [21, 25]. Therefore, risk is not a factor in judgements made with the VAS. The VAS is anchored by the outcomes "death" and "full health" at 0 and 100 respectively (Figure 2). The VAS is subject to the effects of context bias and end-aversion bias, and so it is generally accepted to be less reliable than other scaling methods [25]. Furthermore, VAS utility ratings are systematically lower than SG utility ratings at the sample level

despite both being anchored on the same health outcomes. Despite its biases and systematic error, the VAS is a relatively quick and easy tool for evaluating health utility with participants who have not been trained in statistical analysis, particularly when multiple health outcomes need to be evaluated for guideline development [25].

Using multi-attribute theory, subjects describe a health outcome based upon a series of variable health attributes, usually having to do with degree of function (e.g. mobility, sensation, cognition, etc.) [26, 27]. The health utility of the health outcome, or combination of attributes, is derived from statistical models that consider the values and preferences of the general population. The preferences of the general population are calculated using the mean scores from direct estimation methods.

1.4 Defining Outcomes During Evidence Synthesis

In our work in guideline development we identified a fundamental problem with consideration of outcomes and calibration of the importance and utility rating scales. That is, panel members often have implicit, different definitions of health outcomes that can lead to differences in importance ratings, utility ratings, and final panel recommendations. In fact, the impetus for this thesis was the result of recent informal exercises that we conducted with the European Commission Initiative on Breast Cancer (ECIBC) Guidelines Development Group (GDG). Due to unusually long panel discussion on the outcome "over-diagnosis of breast cancer", GDG

members were asked to define the outcome independently, which had already been rated as important to decision-making. We revealed that there was stark heterogeneity in the panel`s definition of the outcome. Given that outcomes are not explicitly defined in guidelines and until now also not uniformly in the GRADE approach, it is likely that most guideline developers are leaving definitions of outcomes which may be experienced differently (e.g. relatively long versus short wait times, recovery times, emotional response, etc.) to the assumptions of panel members. Furthermore, it is also likely that heterogeneity exists among participants in external health utility ratings (which can inform healthcare guidelines), given that scaling methods are the same as those used during guideline development.

Logic dictates that the heterogeneity could cause a variety of problems during guideline development. Firstly, the transparency of guideline development methods is reduced because guideline end-users cannot be certain of the rationale for judgements made during evidence syntheses. Secondly, the efficiency of panel discussion is reduced because valuable time may be spent trying to harmonize understanding of the outcomes among panel members. Most importantly, the heterogeneity could lead to variability in importance ratings and utility ratings, thereby creating potential for a panel to arrive at recommendations of a different strength or direction than they would have otherwise. Overall, the issues posed from heterogeneity in outcome definitions might become even more problematic on a systematic level as research groups such as the Core Outcome Measures in

Effectiveness Trials (COMET) develop standardized outcome sets for medical research (which can be used as evidence for guidelines) [28].

Researchers have investigated how best to eliminate the heterogeneity. Standardized outcome definitions deemed "maker states", have been used to calibrate health utility ratings using the VAS and SG method [29-31]. In this work marker states improved measurement properties of the VAS but not the SG. Given that marker states may improve guideline transparency and efficiency, explicitly defining outcomes using a standardized format in between outcome generation and importance rating exercises during guideline development may still be merited.

1.5 Standardized Methods for Defining Health Outcomes

In her guidelines for the development of health outcome descriptions, Llewellyn-Thomas argued that any criteria for describing health outcomes would be highly dependent on the purpose of the description and its target population [32]. She proposed that standardized methods used to describe health outcomes be adapted for their purpose based upon three areas: attributes under consideration (e.g. level of detail, evidence source, etc.), evaluation techniques to be applied (e.g. scaling methods for health utility) and format for presenting the health outcomes. Research on health outcomes has given some insight into which criteria might be best for guidelines. COMET has provided guidance for developing core outcomes but they do not sufficiently include the guideline developer perspectives.

1.5.1 Guidance on Management of Attributes in Outcome Definitions

Given the patient-focused nature of guidelines, it is logical that any criteria for management of attributes in outcome definitions during guideline development would facilitate representation of patient values. Sherbourne *et al*. reported that patients value physical, social, and mental health equally during health care decision-making [33]. This suggests that presentation of these dimensions of health should be balanced in outcome definitions, regardless of format.

There is little guideline-specific evidence that provides guidance on how to edit outcome definitions and manage the level of detail of included information. In research initiatives where marker states have successfully been used for health utility ratings, it is standard practice for expert panels to assess the acceptability of the content before use and suggest changes as necessary [29, 31]. This would likely be an appropriate technique for guideline development, given that an expert panel is readily available. Researchers have confirmed the level of detail described in an outcome can influence health utility ratings, but they were unable to conclude whether a high or low amount of detail was responsible for the bias [29]. More research is needed on these topics to properly inform standardization of outcome descriptions in guideline development.

1.5.2 Guidance on Management of Evaluation Techniques

In her guidelines, Llewellyn-Thomas argues that the VAS is the only scaling method that would be efficient enough to use for health utility rating [32]. The

GRADE Working Group already promotes the VAS for rating health utility, and the VAS seems the best candidate to use with standardized outcome definitions.

1.5.3 Guidance on Format of Outcome Definitions

There has been a debate among researchers regarding whether to use long narratives or point-form table formats to present outcome definitions. Researchers found that patients preferred a short table format over a narrative format because they found it easier to understand [29, 34].

In her guidelines, Llewellyn-Thomas recommends that those seeking to standardize outcome definitions consider how the mode of presentation (computer, written, etc), order of the outcomes, framing of the language might bias health utility ratings [35]. No further research has been done to resolve these issues in the context of guideline development, likely because they are very situational. Guideline developers already rely on computers and survey software is easily accessible [36]. Therefore, online, randomized presentation of outcome descriptions seems plausible and appropriate for importance and utility ratings during guideline development.

1.6 Summary of background for thesis

For guideline development, it is implicitly understood that clear descriptions of evidence are required for its synthesis and appraisal. During evidence synthesis, relevant health outcomes are generated and rated for importance [8]. These ratings

facilitate the weighing of the balance of benefits and harms and inform panel discussion during formation of recommendations. After conducting informal exercises with a guideline panel, we found that there was heterogeneity in the implicit definitions of important health outcomes among panel members. Heterogeneity in how guideline panel members understand outcomes reduces transparency of guideline methodology, efficiency of panel discussions, and reliability of importance and utility ratings (which can cause different recommendations). This suggests that in between outcome generation and importance ratings, outcomes should be explicitly defined, using a standardized format, to calibrate the importance and utility rating scales. To tackle this problem, members of our team from McMaster University made a template for developing standardized definitions of health outcomes that we call health marker state descriptors. Evidence is limited about the use of marker states in guidelines. In fact, we are aware of only two guidelines, but results have not been published. This thesis explores methods for standardizing health outcome definitions and using developed health marker state descriptors for evidence synthesis in the context of guideline development.

1.7 Thesis Objectives & Rationale

The main section of this thesis includes one scientific article (Chapter 2). That article is a case study of the development process of health marker state descriptors in the context of the current ECIBC breast guidelines. A case study design was selected because we thought it would allow us the best understanding

of the process of health marker state descriptor development in the context of guideline development. The objectives of the study include:

1. To determine which aspects of the development, content and use of health marker state descriptors are valuable to guideline developers;
2. To further develop and validate our template for health marker state descriptors;
3. To provide guidance on how best to develop health marker state descriptors for guideline development and use them to facilitate health utility rating exercises, and panel discussion.

Thus, the aim of the work presented here is to standardize definitions of health outcomes, with an emphasis on improving GRADE methods for guideline development.

APPENDIX

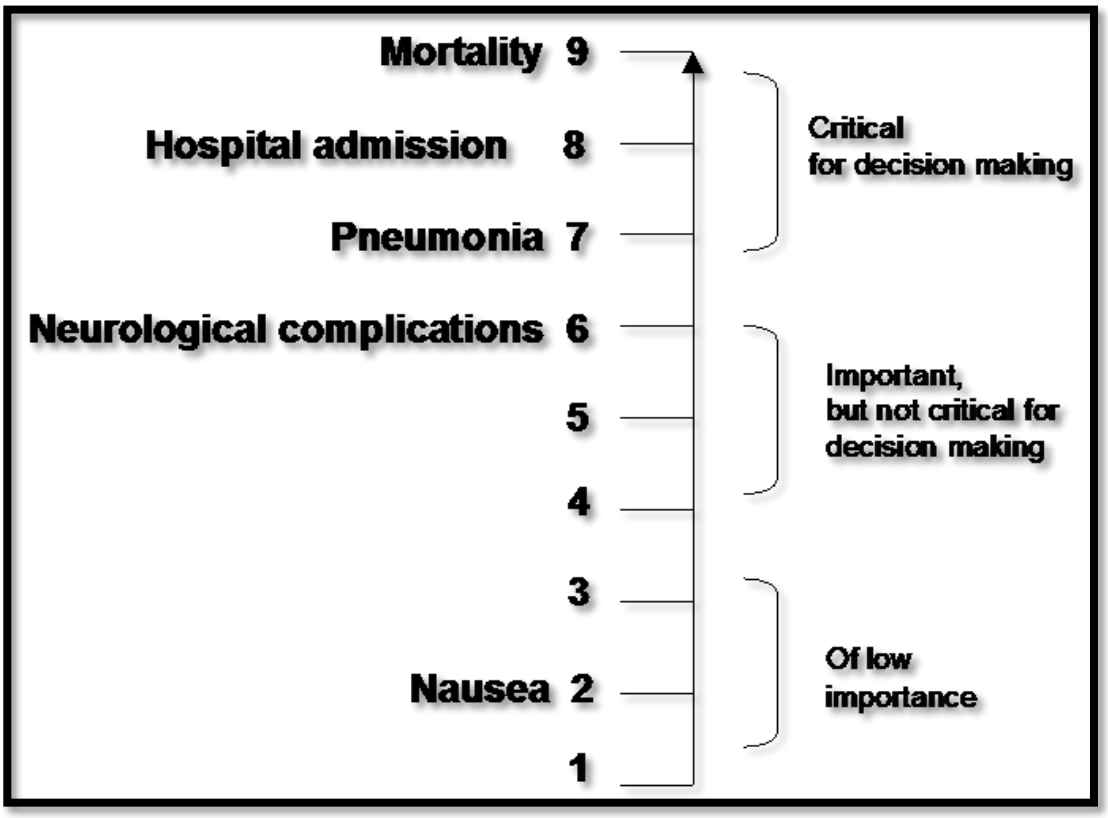

Figure 1: GRADE Importance Scale and Hypothetical Outcome Ratings

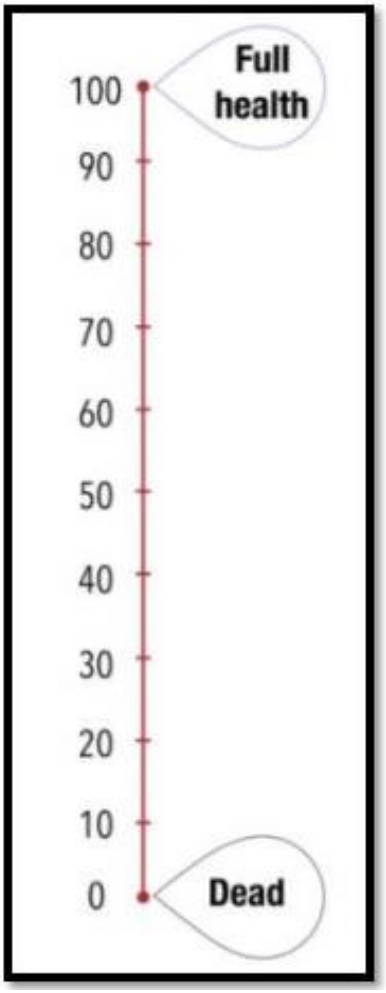

Figure 2: Visual Analogue Scale with anchor outcomes used during Health Utility Rating.

1.8 References

1. Guidelines, I.o.M.C.o.S.f.D.T.C.P., *Clinical Practice Guidelines We Can Trust*. 2011: National Academies Press.
2. Brouwers, M.C., et al., *AGREE II: advancing guideline development, reporting and evaluation in health care.* J Clin Epidemiol, 2010. **63**(12): p. 1308-11.
3. Schunemann, H.J., et al., *Guidelines 2.0: systematic development of a comprehensive checklist for a successful guideline enterprise.* CMAJ, 2014. **186**(3): p. E123-42.
4. Shekelle, P., et al., *Developing clinical practice guidelines: reviewing, reporting, and publishing guidelines; updating guidelines; and the emerging issues of enhancing guideline implementability and accounting for comorbid conditions in guideline development.* Implement Sci, 2012. **7**: p. 62.
5. Eccles, M.P., et al., *Developing clinical practice guidelines: target audiences, identifying topics for guidelines, guideline group composition and functioning and conflicts of interest.* Implement Sci, 2012. **7**: p. 60.
6. Woolf, S., 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**: p. 61.
7. Qaseem, A., et al., *Guidelines International Network: toward international standards for clinical practice guidelines.* Ann Intern Med, 2012. **156**(7): p. 525-31.
8. Guyatt, G.H., et al., *GRADE guidelines: 2. Framing the question and deciding on important outcomes.* J Clin Epidemiol, 2011. **64**(4): p. 395-400.
9. Manchikanti, L., S. Abdi, and L.F. Lucas, *Evidence synthesis and development of guidelines in interventional pain management.* Pain Physician, 2005. **8**(1): p. 73-86.
10. Cook, D.J., C.D. Mulrow, and R.B. Haynes, *Systematic reviews: synthesis of best evidence for clinical decisions.* Ann Intern Med, 1997. **126**(5): p. 376-80.
11. Guyatt, G., et al., *GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables.* J Clin Epidemiol, 2011. **64**(4): p. 383-94.
12. Group, G.W., *Handbook for grading the quality of evidence and the strength of recommendations using the GRADE approach.* 2013.
13. Gordon Guyatt DR, M.O.M., and Deborah J. Cook, *Users' Guides to the Medical Literature: A Manual for Evidence-Based Clinical Practice.* Second Edition ed. 2008: McGraw-Hill Education.
14. Carrasco-Labra, A., et al., *Improving GRADE evidence tables part 1: a randomized trial shows improved understanding of content in summary of findings tables with a new format.* J Clin Epidemiol, 2016. **74**: p. 7-18.
15. Langendam, M., et al., *Improving GRADE evidence tables part 2: a systematic survey of explanatory notes shows more guidance is needed.* J Clin Epidemiol, 2016. **74**: p. 19-27.
16. Santesso, N., et al., *Improving GRADE evidence tables part 3: detailed guidance for explanatory footnotes supports creating and understanding GRADE certainty in the evidence judgments.* J Clin Epidemiol, 2016. **74**: p. 28-39.
17. Guyatt, G.H., D.H. Feeny, and D.L. Patrick, *Measuring health-related quality of life.* Ann Intern Med, 1993. **118**(8): p. 622-9.

18. Torrance, G.W., *Measurement of health state utilities for economic appraisal.* J Health Econ, 1986. **5**(1): p. 1-30.
19. Dolan, P., et al., *Valuing health states: a comparison of methods.* J Health Econ, 1996. **15**(2): p. 209-31.
20. Rutten-van Molken, M.P., et al., *Methodological issues of patient utility measurement. Experience from two clinical trials.* Med Care, 1995. **33**(9): p. 922-37.
21. David L Streiner, G.R.N., *Health Measurement Scales: a practical guide to their development and use* Fifth Edition ed. 1989: Oxford.
22. Brazier, J., et al., *A review of the use of health status measures in economic evaluation.* Health Technol Assess, 1999. **3**(9): p. i-iv, 1-164.
23. Froberg, D.G. and R.L. Kane, *Methodology for measuring health-state preferences--II: Scaling methods.* J Clin Epidemiol, 1989. **42**(5): p. 459-71.
24. Brazier, J., M. Deverill, and C. Green, *A review of the use of health status measures in economic evaluation.* J Health Serv Res Policy, 1999. **4**(3): p. 174-84.
25. Torrance, G.W., D. Feeny, and W. Furlong, *Visual analog scales: do they have a role in the measurement of preferences for health states?* Med Decis Making, 2001. **21**(4): p. 329-34.
26. Torrance, G.W., et al., *Multiattribute utility function for a comprehensive health status classification system. Health Utilities Index Mark 2.* Med Care, 1996. **34**(7): p. 702-22.
27. Torrance, G.W., M.H. Boyle, and S.P. Horwood, *Application of multi-attribute utility theory to measure social preferences for health states.* Oper Res, 1982. **30**(6): p. 1043-69.
28. Williamson, P.R., et al., *The COMET Handbook: version 1.0.* Trials, 2017. **18**(Suppl 3): p. 280.
29. Bennett, K., G. Torrance, and P. Tugwell, *Methodologic challenges in the development of utility measures of health-related quality of life in rheumatoid arthritis.* Control Clin Trials, 1991. **12**(4 Suppl): p. 118S-128S.
30. Schunemann, H.J., et al., *Do clinical marker states improve responsiveness and construct validity of the standard gamble and feeling thermometer: a randomized multi-center trial in patients with chronic respiratory disease.* Qual Life Res, 2006. **15**(1): p. 1-14.
31. Schunemann, H.J., et al., *A clinical trial to evaluate the measurement properties of 2 direct preference instruments administered with and without hypothetical marker states.* Med Decis Making, 2003. **23**(2): p. 140-9.
32. Llewellyn-Thomas, H.A., *Health state descriptions. Purposes, issues, a proposal.* Med Care, 1996. **34**(12 Suppl): p. DS109-18.
33. Sherbourne, C.D., R. Sturm, and K.B. Wells, *What outcomes matter to patients?* J Gen Intern Med, 1999. **14**(6): p. 357-63.
34. Schunemann, H.J., et al., *A comparison of narrative and table formats for presenting hypothetical health states to patients with gastrointestinal or pulmonary disease.* Med Decis Making, 2004. **24**(1): p. 53-60.
35. Llewellyn-Thomas, H., et al., *Describing health states. Methodologic issues in obtaining values for health states.* Med Care, 1984. **22**(6): p. 543-52.
36. Burgers, J.S., et al., *Towards evidence-based clinical practice: an international survey of 18 clinical guideline programs.* Int J Qual Health Care, 2003. **15**(1): p. 31-45.

CHAPTER 2: Development and Use of Health Outcome Descriptors: A Guideline Development Case Study

AUTHORS

Tejan Baldeh[a,b], Zuleika Saz-Parkinson[c], Paola Muti[a,d], Nancy Santesso[a,b], Gian Paolo Morgano[a,b], Wojtek Wiercioch[a,b], Robby Nieuwlaat[a,b], Axel Gräwingholt[e,f], Mireille Broeders[e], Stephen Duffy[e], Solveig Hofvind[e,g,h], Lennarth Nystrom[e], Lydia Ioannidou-Mouzaka[e], Sue Warman[e], Helen McGarrigle[e], Susan Knox[e], Patricia Fitzpatrick[e], Paolo Giorgi Rossi[e], Cecily Quinn[e,i], Bettina Borisch[e], Annette Lebeau[e,j,k], Chris de Wolf[e], Miranda Langendam[e,l,m], Thomas Piggott[a], Livia Giordano[e], Cary Vanlandsveld-Verhoeven[e], Jacques Bernier[e], Peter Rabe[e], Holger J. Schünemann[a,b,e,l,n]

[a]Department of Health Research Methodology, Evidence & Impact, McMaster University, Hamilton, Ontario, Canada.

[b]McMaster GRADE Center, 1280 Main Street West, Hamilton, Ontario L8S 4K1, Canada.

[c]Health and Consumer Protection. European Commission, JRC Ispra

[d]Department of Oncology, McMaster University, Hamilton, Canada.

[e]European Commission Initiative on Breast Cancer Guidelines Development Group. European Commission, JRC Ispra

[f]Private Group Practice for Radiology, Radiologie am Theater, Paderborn, Germany.

[g]Cancer Registry of Norway, PO 5313 Majorstua, Oslo, Norway

[h]Oslo Metropolitan University, Pilestredet 46, 0167 Oslo, Norway

[i]School of Medicine, University College Dublin, BreastCheck, Irish National Breast Screening Programme, St. Vincent's University Hospital, Elm Park, Dublin 4, Ireland

[j]Department of Pathology, University Medical Center Hamburg-Eppendorf, Hamburg

[k]Private Group Practice for Pathology, Lübeck, Germany

[l]Cochrane Applicability and Recommendations Methods Group, 1280 Main Street West, Hamilton, Ontario L8S 4K1, Canada.

[m]Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Amsterdam UMC, University of Amsterdam, Meibergdreef 9, Amsterdam, The Netherlands

[n]Department of Medicine, McMaster University, 1280 Main Street West, Hamilton, Ontario L8S 4K1, Canada

Emails:

Tejan Baldeh – baldeht@mcmaster.ca

Zuleika Saz-Parkinson - Zuleika.SAZ-PARKINSON@ec.europa.eu

Paola Muti - muti@mcmaster.ca

Nancy Santesso - santesna@mcmaster.ca `

Gian Paolo Morgano - morganog@mcmaster.ca

Wojtek Wiercicoch - wierciww@mcmaster.ca

Robby Nieuwlaat - nieuwlr@mcmaster.ca

Axel Gräwingholt - axel.graewingholt@t-online.de

Stephen Duffy - s.w.duffy@qmul.ac.uk

Solveig Hofvind - solveig.hofvindreftregisteret.no

Lennarth Nystrom - Lennarth.Nystrom@umu.se

Lydia Ioannidou-Mouzaka - lydiamouzaka@yahoo.gr

Sue Warman - suegeoff@btinternet.com

Helen McGarrigle - helen.mcgarrigle@wales.nhs.uk

Susan Knox - susan.knox@europadonna.org

Patricia Fitzpatrick – patricia.fitzpatrick@ucd.ie

Paolo Giogi Rossi – paolo.giorgirossi@ausl.re.it

Cecily Quinn - c.quinn@svuh.ie

Bettina Borisch - bettina.borisch@unige.ch

Annette Lebeau - a.lebeau@uke.de

Chris de Wolf - chris@iquat.org

Miranda Langendam - m.w.langendam@amc.uva.nl

Thomas Piggott - thomas.piggott@gmail.com

Livia Giordano - livia.giordano@cpo.it

Cary Vanlandsveld-Verhoeven - c.vanlandsveld-verhoeven@LRCB.NL

Jacques Bernier - jbernier@genolier.net

Peter Rabe - prabe@koop-mammo.de

Holger J. Schünemann – schuneh@mcmaster.ca

405 **CORRESPONDENCE**

406 Holger J. Schünemann
407 Department of Health Research Methodology, Evidence & Impact
408 McMaster University Health Sciences Centre
409 1280 Main Street West. Hamilton, ON, L8N 4K1, Canada
410 Email: schuneh@mcmaster.ca
411 **Tel: +1 905 525 9140 x 24931**
412

413

414

415

416

417

418

419

420

421

422

423

424

425

426

427

428

429

430

431

ABSTRACT:

OBJECTIVES: During health guideline development, panel members often have implicit, different definitions of health outcomes that can lead to variability in evidence synthesis and recommendations. McMaster GRADE Centre researchers developed a standardized description of health outcomes using the health marker state format. We aimed to determine which aspects of the development, content, and use of marker states were valuable to guideline developers.

STUDY DESIGN & SETTING: We conducted a case study of marker state development with the European Commission Initiative on Breast Cancer (ECIBC) Guidelines Development Group (GDG). Eighteen GDG members provided written and interview feedback on the process. Using the health marker states, 2 health utility rating surveys were conducted near the beginning and end of development respectively.

RESULTS: We developed 24 marker states for outcomes related to breast cancer screening and diagnosis. Feedback from GDG members revealed that marker states could be useful for developing recommendations and improving transparency of guideline methods. Comparison of the two health utility surveys showed a decrease in standard deviation in the second survey across 21 (88%) of the outcomes.

CONCLUSIONS: Health marker states are a promising method, satisfying the pre-requisite of being feasible, acceptable, and with some initial result on reduction of variance of health utility scores.

2.1 Introduction

Healthcare guidelines aim to support healthcare professionals, recipients of care and policy makers in making best decisions for care. Guidelines are not without risk of bias [1-3]. For guidelines to be trustworthy they generally should be developed according to 6 principles [1, 3-7]:

1. Involvement of multidisciplinary stakeholders
2. Recommendations supported by systematic reviews of the evidence
3. Description and consideration of important subgroups and peoples' values and preferences
4. Management of conflict of interest
5. Ratings of the certainty or quality of evidence and transparency in moving from evidence to recommendations.
6. Updating and revisions

The Grading of Recommendations, Assessment, Development, and Evaluation (GRADE) approach is widely used by guideline developers and agencies to systematically evaluate the quality of evidence and strength of healthcare recommendations for the fifth principle [8]. Transparency of the guideline development process is a key goal. GRADE accomplishes this by directing guideline developers to consider health outcomes that are deemed to be "critical" or "important" to stakeholders for decision-making [9]. Those deciding which

outcomes to include in the decision-making, ideally by focusing on what matters to those affected by the recommendation, determine the importance of relevant health outcomes by placing the outcome on a 1 to 9 scale (1-3 = low importance for decision making, 4-6 = important, but not critical for decision making, 7-9 = critical for decision making) in GRADE [10]. The highest-rated outcomes (rated at most "important") are included in GRADE evidence tables that summarize the key information of a systematic review [11-14]. These tables support decision making, including the formulation of recommendations by guideline panels. The importance rating exercise intends to mitigate several challenges in guideline development. It orients panel members to the task of focusing on outcomes that matter, reduces the number of outcomes deemed to be patient-important, identifies the level of agreement for the outcome of interest, and indicates the relative importance of the beneficial and harmful outcomes (e.g. within the "critically Important" category an outcome rated as 9 will be more important than an outcome rated as 7).

Health utility ratings are used similarly in a guideline panel's harm-benefit analysis of health outcomes [15]. Health utility is a measure of the values attached to the outcomes [16]. Outcome-specific health utility ratings are often not available or are not applicable to certain target populations [17]. Therefore, panels sometimes rate the health utility of outcomes internally to most accurately measure their collective views on the relative benefits and harms of each outcome. For instance, guideline panel members may rate the outcome on the validated Visual

Analogue Scale (VAS) which is anchored by the outcomes "death" and "full health" at 0 and 100 respectively.

However, in the McMaster GRADE team's work with guideline developers, a fundamental problem with consideration of outcomes and calibration of the importance and utility rating scales was identified. That is, panel members often have implicit different definitions of health outcomes that can lead to differences in importance ratings, utility ratings, and final panel recommendations. In fact, the impetus for this study was a recent observation with the European Commission Initiative on Breast Cancer (ECIBC) Guidelines Development Group (GDG). We revealed that there was considerable variation between GDG members' definition of the outcome "over-diagnosis of breast cancer". However, clear agreement by a guideline panel on what constitutes an outcome is required to balance benefits and harms, to communicate with the public, and to conduct research. Furthermore, to promote transparency of guideline development methods, guideline end-users require clear explanations of what constitutes each important outcome.

To tackle this problem in the ECIBC, we utilized a template developed by researchers at McMaster GRADE Centre to standardize descriptors of health outcomes that are akin to health marker state descriptors [18, 19]. Health marker state descriptors are primarily intended to support the generation of recommendations by guideline developers and promote understanding of development methods by guideline end-users secondarily. Here, we describe the development and use of these health marker state descriptors in the context of the

European guidelines for breast cancer screening and diagnosis. The purpose of this case study was to determine which aspects of the development, content and use of health marker state descriptors are valuable to guideline developers broadly. We describe lessons learned to improve the structure of the tool and provide guidance for the future development and use of health marker state descriptors.

2.2 Methods

2.2.1 General Methods

We conducted a case study of the development of health marker state descriptors in the context of the European guidelines for breast cancer screening and diagnosis. We selected a case study design to elicit high quality feedback from guideline developers involved in the process of health marker state development. The case study began and ended during development of the guidelines, but it was separate from guideline development. The design of the health marker state descriptor development methods were based upon proposed guidelines for their development [20]. We developed first drafts of the health marker state descriptors using a template (Figure 1). Throughout development, GDG members provided feedback on the drafts and development process. This was done through three rounds of semi-structured interviews and online written feedback. Iterative changes were made to the content and format of the health marker state descriptors based upon the observations of McMaster University researchers and GDG feedback. In between rounds of feedback, GDG members also completed two online health

554 utility assessments. Unique to this study, marker states were used to facilitate the
555 exercises. We analyzed the utility scores to quantitively assess whether the
556 development process had an impact on harmonization of outcome definitions as
557 well as values and preferences towards the health outcomes.

558 ***2.2.2 Participants***

559 We formed a steering committee to coordinate the development of the
560 health marker state descriptors for the European guidelines for breast cancer
561 screening and diagnosis consisting of five researchers: four health methods
562 researchers (HS, NS, PM, ZSP) and one graduate student with training in health
563 sciences (TB).

564 Members of the guidelines development group (GDG) participated in the
565 development of the health marker state descriptors. These members were
566 clinicians, epidemiologists, cancer scientists, methodologists, economists, and
567 patients. Each GDG member declared their interests to the ECIBC as part of their
568 agreement to participate in the guideline development.

569 All GDG members, including those participating in this study, were invited
570 by the ECIBC to develop the European guidelines for breast cancer screening and
571 diagnosis. Participation in this study was voluntary and signed consent was
572 obtained from all those providing feedback. The methods for this study were
573 approved by the Hamilton Integrated Research Ethics Board (HiREB).

2.2.3 Template of Health Marker State Descriptors

We utilized a draft template (Figure 1) for health marker state descriptors [18-20]. The format was purposefully designed to be concise; written at a Grade 8 reading level (as indicated by the Flesch–Kincaid readability tests) from the perspective of the healthcare recipient, who is the primary beneficiary of any healthcare guideline. The template included 4 bulleted domains: "Symptoms", "Time Horizon", "Treatment and Testing", and "Consequences".

[Name of Health Outcome] – importance rating

Symptoms: [List most common symptoms]

Time Horizon: [Describe how long symptoms will persist for and how they might change over time. Also describe approximate timing of relevant healthcare]

Testing and Treatment: [Describe relevant healthcare or interventions].

Consequences: [Describe relevant consequences resulting from the health outcome or relevant healthcare]

Figure 1: Draft Template for Development of Health Marker State Descriptors

2.2.4 Development of Draft Health Marker State Descriptors

The methods for development of the 1st draft health marker state descriptors are summarized in Figure 2 (steps 1 - 3). Realizing the need to harmonize understanding of the outcomes, but after the ECIBC guidelines were initiated, the

steering committee used the draft template (Figure 1) to write 24 draft health marker state descriptors relevant to breast cancer screening, diagnosis, and treatment. For this study, outcomes were selected for marker state development when it had been determined that they should be included in GRADE evidence tables from discussion with guideline developers and use of the GRADE importance rating exercise (indicating that they had been deemed important or critical to decision-making by the GDG). To populate the draft template, we utilized information from quality of life instruments, scientific literature, and collective subject experience [21-30].

2.2.5 Refinement of Marker State Content and Structure

Figure 2 summarizes our methods for reviewing the content of the health marker state descriptors (steps 4-10). After we completed internal development of the drafts, the content refinement process included comments from the ECIBC GDG members on the development methods, content, and structure of the health marker state descriptors. Ten of 30 GDG members volunteered to participate in individual semi-structured interviews at the JRC-Ispra location and the subsequent online comment, nine provided written comments only. All interviews were conducted at quarterly GDG meetings, by the same interviewer (TB), using the same list of prompting questions with transcription for analyses. Each GDG member had different time commitments at the meetings and so their availability to participate in interviews varied. Whenever possible, we repeated interviews with available panel members at different meetings to get their feedback throughout

development. During the written online feedback, GDG members could actively discuss content issues with other ECIBC GDG members. We developed 2nd drafts of all health marker state descriptors after reviewing the GDG's feedback and making the relevant changes to the health marker state descriptors when there were factual errors or important omissions in our content. When we were unsure whether to make changes based upon GDG feedback we looked for supporting literature before approving the changes. We then held two additional rounds of GDG feedback (each having an interview and online component) and made edits using the same approach to develop a 3rd and 4th draft, respectively. Throughout the development process, we ensured that all health marker state descriptors were reviewed by at least one member of the GDG. After each round of feedback with the GDG, the drafts were presented to the ECIBC guideline developers (including the GDG) for review or approval.

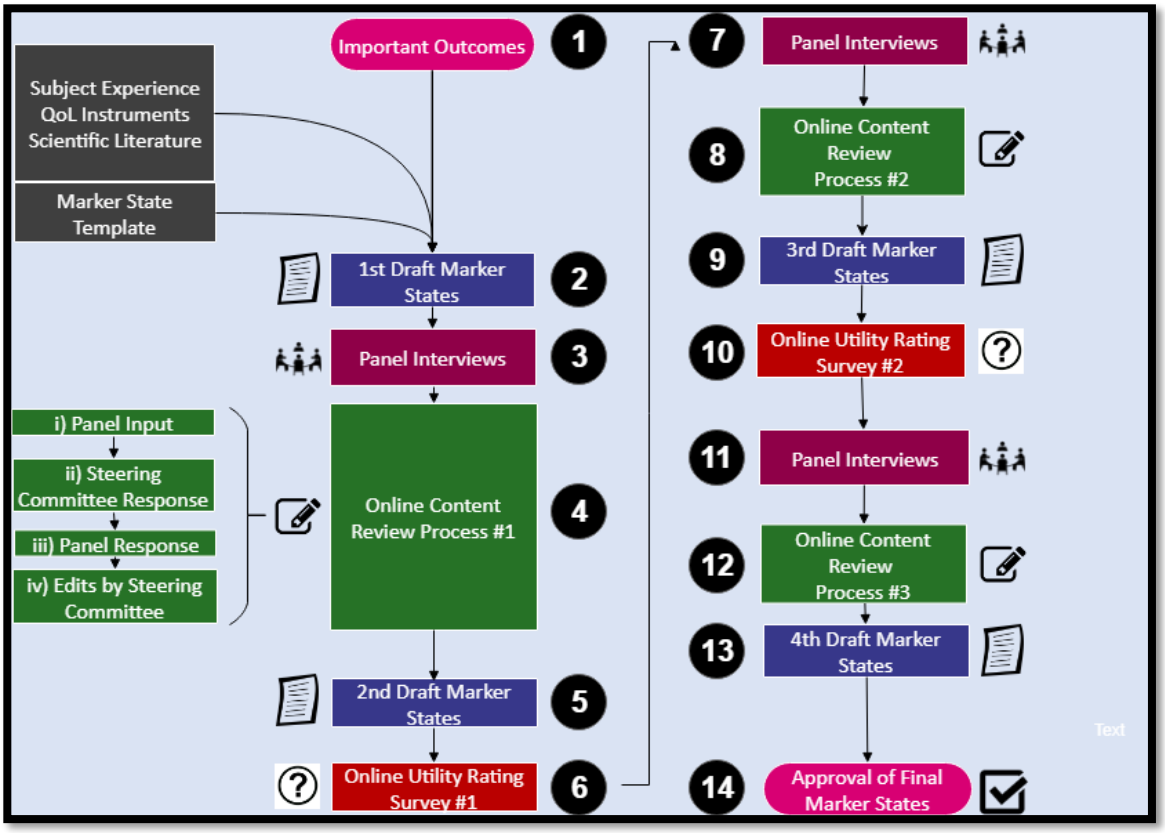

Figure 2: Health Marker State Descriptor Development Process. McMaster researchers *developed first drafts of the health marker state descriptors using a template and relevant source material. GDG members provided feedback on the drafts in semi-structured interviews with McMaster researchers and online. This was done through three rounds of semi-structured interviews and online written feedback. Iterative changes were made to the content and format of the health marker state descriptors based upon the observations of McMaster University researchers and GDG feedback. In between rounds of feedback, GDG members also completed two online health utility assessments.*

2.2.6 Online Utility Rating Surveys

Separate from health marker state descriptor development, we conducted online surveys to elicit health utilities from the GDG for the 24 health outcomes using a VAS and further examine the uses for health marker state descriptors. We

636 did this to validate our work on the marker states. On our 0 to 100 VAS, 0 is
637 anchored at "dead" and 100 at "full health" [18, 19]. We administered the surveys
638 to the entire GDG immediately after development of the 2nd and 3rd marker state
639 drafts respectively. Thus, by design, the GDG members that participated rated the
640 health utility of health outcomes twice (once per survey). The most current versions
641 of the health marker state descriptors were used to describe all health outcomes
642 in the surveys, including the VAS anchors. The steering committee made iterative
643 changes to the survey instructions based upon thematic analysis of the GDG's
644 interview feedback.

645 ***2.2.7 Data Analysis***

646 We conducted thematic analysis of the transcribed GDG interviews and
647 utility surveys in six steps [31] using NVIVO version 11 software. First, two
648 McMaster GRADE Centre researchers (TB, GPM) reviewed the interview
649 transcripts and survey feedback. Second, each reviewer independently coded the
650 material. Third, coding was reviewed to identify themes. Care was taken to note
651 the respective timing of the themes in development, and how they changed over
652 time. Fourth, the reviewers met to pool the themes and ensure that the codes were
653 appropriate for each theme. Fifth, the reviewers discussed and agreed upon
654 refinement of the themes. Finally, the first author applied the themes during
655 manuscript drafting for review by the steering committee.

We conducted all quantitative analyses of the health utility ratings using IBM SPSS version 20. For the descriptive analysis, we calculated the outcome-specific mean utility ratings per survey, and corresponding standard deviation for each health marker state descriptor. If our health marker state descriptors were effective for harmonizing understanding of outcomes, we expected to observe a reduction in variance of mean health utility scores across outcomes. For each outcome we performed Levene's F-tests to assess whether the variance in mean utility ratings for both surveys were equivalent to one another. The raters and outcomes were the same for both surveys so we hypothesized that there would be less variance over time if through the iterative process the content of the marker states improved.

2.3 Results

2.3.1 Health Marker State Descriptors

We developed 24 health marker state descriptors (Figure 3); each was approved by ECIBC guideline developers (including the GDG). An example health marker state descriptor is provided in Figure 4 and the full ECIBC health marker state descriptors is presented in the Appendix and the GRADE health outcome descriptor or marker state database. This database already houses health outcome descriptors for nearly one hundred outcomes for several conditions and developers are invited to submit their work to enhance the database.

1. Accessibility to Information
2. Awareness of Information
3. Participation in Screening
4. Informed Decision Making
5. Satisfaction with Decision Making
6. Confidence with Decision Making
7. Abnormal Screening Result
8. Recall for Assessment
9. False Positive Screening Result
10. Suspicious Indeterminate Calcification
11. False Positive Biopsy Result
12. Breast Cancer Detection
13. Breast Cancer Stage
14. Determination of Biomarker Status
15. Interval Breast Cancer
16. Over-Diagnosis & Over-Treatment of Breast Cancer
17. False Negative Screening Result
18. Radiation Exposure from Mammogram & Assessments Using Radiation
19. Provision of Surgical Therapy
20. Mastectomy
21. Provision of Medical Therapy
22. Provision of Radiotherapy
23. Provision of Chemotherapy
24. Other Cause Mortality

Figure 3: List of Health Marker State Descriptors Developed for ECIBC

False-Negative Screening Result

This marker state refers to receiving a negative screening result (no breast cancer) when you actually have a breast cancer. This is called a false negative screening result. Not all women become aware that they had a false negative screening result. This marker state describes when they do become aware after subsequent diagnosis.

False Negative Screening Result – importance and utility rating

What you feel or experience: When you find out that you did have breast cancer and it was missed, you are likely to feel anger, fear, and anxiety.
Time Horizon: It may take months to years before you find out that you did have breast cancer when you were told you did not.
Testing and Treatment: Following the discovery of your breast cancer later on, you may have to undergo treatment that is more intense than if the cancer had been detected right away, as the cancer may have developed to a more advanced stage.
Consequences: The consequences of late detection of a slow growing breast cancer will probably be not substantial with respect to treatment and prognosis. However, if the breast cancer has grown, your predicted outcome is likely worse than if it had been diagnosed at the screen. Survival from breast cancer that has a false-negative diagnosis may be worse compared to women with screen-detected breast cancer, but comparable to women who do not attend screening.

Figure 4: Example Health Marker State Descriptor Developed for ECIBC

675

2.3.2 ECIBC GDG Interview Feedback

We conducted fourteen semi-structured interviews with ten GDG members to collect feedback on the development methods, content, use, and implementation plans for health marker state descriptors. Six interviews, four interviews, and four interviews were conducted after the development of the first, second, and third health marker state descriptor drafts, respectively. The thematic analysis of the interview transcripts revealed six themes.

Theme 1: Marker State Development Process

Overall, GDG members felt that the methods used for marker state development in this study were appropriate. In each round of interviews, the online refinement process was described as acceptable, quick, and effective for improving the quality of the content to an acceptable level. However, the process of participation was considered as a challenge at the beginning of development. GDG members had been invited to participate in marker state development prior to this study. Yet, no GDG member had enough initial interest or availability to take on the task and the steering committee took sole responsibility for early development. GDG members who eventually became involved in marker state development did so only after realizing the importance of health marker state descriptors and offering serious concerns about the content of the first drafts:

"It is so important that you get the content [of the first drafts] at least 80 to 90% right. There were so many things in there that were so far off the mark that it coloured my view."

Despite repeated presentations at GDG meetings, participants felt that the methods and purpose of marker state development in the context of this study was not made clear to them. Therefore, GDG members described insufficient training on the development process and aims of health marker state descriptors as initial barriers to their participation in development.

Theme 2: Comprehensibility of Health Marker State Descriptors

Most members of the GDG felt that the wording of the health marker state descriptors became clear and consistent by the end of the review process. Reading level and emotional sensitivity emerged as important factors for facilitating the use of health marker state descriptors by guideline end-users. Some GDG members felt that the reading level should be relatively high because end-users might feel intellectually insulted by a low reading level:

"The reading level should be increased. We cannot offend women."

Other members suggested that the content should be at a lower reading to facilitate use of health marker state descriptors by less educated members of the public:

"If [health marker state descriptors] are to be used by the broad public I think they need re-wording for someone of a lower literacy level."

The panel was split regarding whether harsh language and mention of negative health effects should be avoided to improve emotional sensitivity of the health marker state descriptors. There was mixed feedback about whether multiple versions of health marker state descriptors (e.g. for healthcare recipients, panel members, healthcare professionals, etc.) should be developed for a single guideline based upon the appropriateness of wording and emotional sensitivity for specific end-user populations.

Theme 3: Data Presentation

Throughout development, the GDG members tended to prefer inclusion of generic attributes in the health marker state descriptors. They were concerned that the information in the health marker state descriptors was only relevant to a small population of those experiencing a health outcome. The use of descriptive statistics emerged as an important factor in improving the generalizability of health marker state descriptors. GDG members felt that use of the averages did not represent the variety of possibilities that an individual could experience for any health outcome:

"Whether it be weeks, days or months; there can be a lot of variation [in timing of symptoms]. So, it seems a bit artificial to state a specific time"

The health marker state descriptors were described as more representative when the minimum and maximum feasible data values were listed in the form of time periods and ranges.

Theme 4: Marker State Structure & Content

GDG members deemed the format of health marker state descriptors to be acceptable. All participants thought that the domains were comprehensive, presented in a logical order, and easily identifiable. However, they explained that the wording of some of the "Symptoms" domain should be changed to make them more intuitive.

Several GDG members acknowledged that the "Consequences" domain was necessary for describing any outcomes. However, some felt that there was little variation and a considerable amount of repetitive content in the domains among the ECIBC health marker state descriptors. This suggest that the scope of the content in the domain should be narrowed and better explained to panel members to ensure that there is little overlapping content. However, it is likely that outcomes for a specific problem or disease and narrow interventions will incur similar consequences.

One GDG member mentioned that the "Testing and Treatment" domain was not appropriate for outcomes for screening programs and preventive efforts because healthcare recipients might not receive treatment:

"Most women that go for screening will not enter any kind of diagnostic efforts, let alone be treated. So, I find it very artificial to be reading up on health marker state descriptors that are directly related to the screening process, and then being pushed [to consider] the treatment area"

That GDG member recommended separating "Testing" and "Treatment" into two domains and explicitly stating when the domains are not relevant.

Theme 5: Using Health Marker State Descriptors

During early development, very few GDG members were able to identify possible uses for health marker state descriptors. However, as GDG members became more familiar with health marker state descriptors they thought that they could be useful for consolidating understanding of outcomes among guideline developers, facilitating panel discussion, and improving the transparency of guideline methods. One GDG member reflected upon the development process in the following:

"I think [health marker stat descriptors] have been very valuable to the [GDG] because it has made us discuss with you, and the rest of the [GDG], what we really mean."

There was agreement among GDG members that health marker state descriptors would need to be referenced and enforced by guideline panel chairs to

be useful for guideline development. For external use, the GDG felt that attaching the health marker state descriptors to the recommendations or publishing them online was best for making them available to end-users.

Theme 6: Utility Rating Survey

Most GDG members indicated that the first online survey was problematic and difficult to complete. Much of the difficulty they described referenced the inappropriateness of the VAS anchors ("dead" and "full health") for rating the health utility of outcomes which had emotional and psychological implications as opposed to physical (e.g. the health marker state descriptor 'Awareness to Information'):

> *"The survey was problematic for me. I tried to complete it honestly but some of the [outcomes], did not lend themselves to the scale of death and full health."*

After the first survey, it emerged that some participants were inappropriately making attribute-based comparisons (e.g. considering only physical or mental or emotional symptoms) or comparing the total number of implications described in each health marker state descriptor. Some did not realize that they were intended to use a holistic strategy to rate how the physical, emotional, and mental implications might affect overall health relative to the anchors. Therefore, we modified the instructions in the second survey to better direct GDG members through the health utility rating process. Other comments from GDG members

suggested that difficulties with the VAS may have manifested from problems with outcome generation:

> *"Some of [the outcomes]…why on earth are there health marker state descriptors for that? It becomes hard to rate if you don't see [the outcome] as important"*

2.3.3 Utility Rating Survey Scores

Twelve of the thirty GDG members participated in each of the utility rating surveys, respectively. Six of those GDG members participated in both surveys. The mean utility ratings for each survey, the results of the pairwise comparison, and variability comparison are presented in Table 1. We attempted to evaluate if the health marker state descriptor revisions had important impact on the health utility ratings. Between the first and second surveys, we observed an increase to the mean scores in 14 outcomes and a decrease in ten outcomes. The variability, that is the magnitude of the standard deviation, of the ratings improved in 21 pairs and it remained similar in 2 pairs. In one health marker state descriptor the standard deviation increased slightly by as much as two percent.

812 *Table 1: Mean Health Utility Ratings using a VAS (0 = 'Dead', 100 = 'Full Health')*

Health Marker State Descriptor	1st Survey Mean Score (SD)	2nd Survey Mean Score (SD)	Levene's F-Test
Accessibility to Information	78 (18)	88 (9)	0.106
Awareness of Information	73 (17)	86 (14)	0.045
Participation in Screening	79 (15)	84 (15)	0.505
Informed Decision Making	82 (16)	89 (11)	0.239
Satisfaction with Decision-Making	80 (12)	89 (12)	0.084
Confidence with Decision-Making	78 (18)	88 (14)	0.162
Abnormal Screening Result	62 (24)	78 (15)	0.044
Recall for Assessment	64 (27)	74 (12)	0.208
False Positive Screening Result	68 (24)	69 (17)	0.861
Suspicious Indeterminate Calcification	64 (21)	68 (18)	0.622
False Positive Biopsy Result	67 (26)	56 (19)	0.252
Breast Cancer Detection	60 (31)	54 (19)	0.573
Breast Cancer Stage	60 (29)	52 (8)	0.386
Determination of Biomarker Status	68 (20)	66 (19)	0.795
Interval Breast Cancer	42 (28)	40 (15)	0.872
Over-Diagnosis & Over-Treatment	54 (23)	62 (18)	0.357
False Negative Screening Result	41 (29)	43 (18)	0.861
Radiation Exposure from Mammogram & Assessments Using Radiation	69 (26)	80 (19)	0.270
Provision of Surgical Therapy	62 (28)	54 (15)	0.395
Mastectomy	49 (26)	43 (16)	0.520
Provision of Medical Therapy	59 (28)	47 (11)	0.160
Provision of Radiotherapy	57 (26)	51 (13)	0.473
Other Cause Mortality	10 (20)	11 (22)	0.869

813

814

815

2.4 Discussion

2.4.1 Key Findings

This case study assessed the development of 21 health marker state descriptors in the context of the European guidelines for breast cancer screening and diagnosis. Thematic analysis of GDG interview feedback revealed that our novel and succinct format was useful and flexible for describing health outcomes. This finding builds upon prior research that identified short narratives as the preferred marker state format by healthcare recipients [19].

Strengthening GDG understanding of outcomes and improving the transparency of guideline methods were identified as the most impactful uses for health marker state descriptors. Changes made to the descriptors after the second round of GDG feedback resulted in reduction in variance of the mean health utility scores rated with the VAS. This suggests that the process of marker state development helped consolidate the values and preferences of the guideline panel, which is crucial for decision-making during the development of recommendations.

GDG members described lack of sufficient training on health marker state descriptor development methods as a barrier to their participation. In this study, most GDG members had only been introduced to the GRADE approach in the context of the ECIBC guidelines, and so general lack of exposure to methods for outcome generation and importance rating as well as other core guideline methods

in an ever-expanding field may have contributed to the confusion regarding marker state descriptors.

Online feedback was an effective and easy method for refining outcome-specific content. The GDG's serious concerns with the content of the first drafts suggest that a multi-disciplinary group, involving representatives from the guideline panel, should be involved in all stages of health marker state descriptor development. Opinions on the appropriate balance of wording, reading level, and emotional sensitivity for end-users was varied. More research must be done on the needs of specific end-user populations to definitively say whether multiple tailored versions of health marker state descriptors are necessary.

Participants also described having significant difficulty with the VAS for health utility rating because they felt that the outcomes anchoring the scale were inappropriate for rating some of the health marker state descriptors. This was particularly true of the information and decision-making outcomes, where the desired and undesired effects may have been perceived as independent from physical health status. Difficulties with the anchor outcomes are further supported by the health marker state descriptor for "Other-Cause Mortality" being valued a mean health utility score of 10. Given that the health marker state descriptor had similar content to the anchor outcome "Dead" (which was visible during the rating exercise), it was expected to be valued at 0. While the rating of 10 suggests some difficulties of completing the exercise it may be explained by simple error. Relevant literature on the VAS describes it as being more acceptable and practical than

other validated scaling methods [32]. Furthermore, the outcomes "dead" and "full health" are widely-used as anchors for scaling methods [33]. Given this, it is most likely that the difficulty with the survey was due to poor instructions, or context bias resulting from rating the health utility of all health outcomes in the same survey. This was our reasoning for changing the instructions between surveys.

Although one participant provided feedback that the testing and treatment domain was inappropriate for outcomes related to preventive interventions, we did not make changes to the format. We believe that testing and treatment should both be considered and connected to healthcare interventions on a pathway that follow from a health state, even if no testing or treatment follows which in itself is important information.

2.4.2 Limitations and Strengths

A limitation of this study was that development of health marker state descriptors for most of the outcomes occurred after they had already been rated for importance and included in GRADE summary of findings tables. The timing of development may have caused confusion about the need and purpose of health marker state descriptors in the guideline development process, although the development need resulted from disagreement during that rating exercise. The timing of the surveys also resulted in there being no control depicting the variance in mean health utility scores without exposure to marker states. Thus, it is difficult to distinguish between the effects of health marker state descriptor development

and growing awareness among GDG members on the observed changes in variance.

Furthermore, marker state development occurred in the context of only one breast cancer screening guideline, which limits our generalization to other panels and healthcare topics. Finally, this study had a small sample size all together, and the response rate of the online utility ratings surveys in this study was poor. The relatively small number of pairwise comparisons for each health outcome reduced the statistical power of our variance analysis.

A strength of this study is that all data was collected from a real-life guideline panel, which is rare among published literature on outcome descriptors. By conducting this case study in the context of a real guideline panel, our results can be used to inform outcome descriptor standardization efforts for guideline development, where we originally identified the problem of heterogeneity. We also carefully planned marker state development methods and interaction with GDG members to capture reliable feedback at each stage of marker state development. Collectively, our planning and analysis ensure that the results from this study can be used to inform all stages of health marker state descriptor development.

2.4.3 Implications for practice

This study's findings highlight the attitudes towards health marker state descriptor development and use among guideline panel members. Results suggest that guideline developers using health marker state descriptors should work with a

multidisciplinary subgroup of panel members in a few rounds with online or in person feedback, to develop first drafts and final versions of the health marker state descriptors respectively. Prior to development, guideline panel members should be trained on development methods accordingly. Our findings may help inform and guide future development of health marker state descriptors for guideline development. We will use the ECIBC health marker state descriptors to better inform users of the outcomes that were considered in each of the questions and publish them on the ECIBC web platform and use them in decision support tools.

2.4.4 Implications for research

Further research will show if multiple versions of the marker state descriptors for different target audiences are necessary, and how the reading level and wording of each version might be tailored for various end-user populations. Our preference is that simple descriptors, that provide a common language for those providing health care and those receiving that care, should be used. There seems to be no logical reasons for a different language for different users. Using a common language will reduce the probability that misunderstandings will occur.

For the use of health marker state descriptors to become more common in guideline development, there is a need to determine how guideline end-users make use of them, so instructions can be altered accordingly. Most importantly, researchers should investigate whether health marker state descriptors do improve transparency and understanding of guideline methods for end-users, as the GDG

members in this study suggested. Additional research efforts can build upon the present study by examining attitudes towards health marker state descriptor use by end-users, particularly healthcare recipients who may not have medical knowledge or experience with illness [34]. Other research efforts might focus on how health marker state descriptors might be adapted for use for other purposes including but not limited to research and education.

Researchers should also concentrate efforts on determining the reliability of the VAS when health marker state descriptors are used, because we were unable to draw meaningful conclusions about this due to the limited statistical power.

2.5 Conclusion

This study described the experiences of health marker state descriptor development for a health care guideline and provided guidance for future efforts. Our standardized marker state descriptor format was useful for facilitating development of recommendations and improving transparency of guideline methods. GDG members used health marker state descriptors with the VAS to improve precision of health utility ratings, but more research must be done to validate this method and reduce measurement error.

Acknowledgments

Eight of the authors of this study are members of the GRADE Working Group and have contributed to the development of the GRADE approach in some capacity (TB, HJS, NS, WW, RN, GPM, TP, ML).

What is new?

Key findings:

Health marker state descriptor development reduces variance during health outcome utility assessment. Guideline panel members believe that health marker state descriptors are effective for harmonizing understanding of outcomes among panel members and improving the transparency of guideline methods.

What this adds to what was known?

This article provides guideline developers with guidance on: (1) developing tools to harmonize understanding of health outcomes among guideline panel members (2) using the newly developed tools to improve the validity of health utility assessments and better inform panel discussion.

What is the implication and what should change now?

Marker state descriptors improve guideline methods by consolidating panel understanding of outcomes and improving transparency. Guideline developers should consider developing, using, and publishing health marker state descriptors with their guidelines.

Appendix – ECIBC Health Marker State Descriptors

1) *Accessibility to Information*

This marker state refers to being able to access information about any breast cancer topic easily if you have been invited to participate in screening. It only considers the period for which you are receiving breast related healthcare.

Accessibility to Information – importance and utility rating

- **What you experience or feel:** You may need to invest effort to seek out information from different sources, including but not limited to your healthcare provider, personal contacts and the internet. You may feel satisfied if you obtained all the information you needed easily.
- **Time Horizon:** You may seek out information on breast cancer screening or on breast cancer a few weeks before you begin regular screening, or a few days after a test result has been communicated to you (or indeed at any other time). You may identify relevant information within minutes to hours depending on the accessibility of what you search for, and how you search for it.
- **Testing and Treatment:** The information which you access may affect your diagnostic and treatment experience in the context of shared decision making. Easy access to information may influence the type and frequency of diagnostic tests, but not screening tests, you may undergo. Depending on the quality of the information you obtain, your screening frequencies, and, if appropriate, diagnostic tests and treatment for your potential breast cancer may be positively or negatively influenced as well.
- **Consequences:** You may find screening and other clinical experiences enhanced by greater knowledge as a result of access to information. On the other hand, you may experience anxiety due to having only a partial understanding of screening, breast cancer, or the risk of suffering from it. Although accessible, the information you find may be inaccurate and in that case, you may make uninformed decisions.

2) Awareness of Information

This marker state refers to being knowledgeable about any breast cancer topic during the period of time for which you are receiving any breast related healthcare for potential/confirmed breast cancer. You may receive information from your healthcare professional, health authorities, the internet, and other sources.

Awareness of Information – Importance & Utility Rating

- **What you experience or feel:** If you are aware of information, you may feel satisfied with your breast healthcare.
- **Time Horizon:** You may start researching breast cancer and screening/diagnostic testing information a few weeks before your first screening/diagnostic test or immediately after a possible diagnosis of breast cancer or recall invitation. Your level of awareness about screening, breast cancer and diagnostic tests for breast cancer may increase over time.
- **Testing and Treatment:** Having a high level of awareness may impact the type and frequency of any diagnostic tests, but not screening tests, you may undergo. Depending on the quality of the information you obtain, your screening frequencies, and, if appropriate, diagnostic tests and treatment for your potential breast cancer may be positively or negatively influenced as well.
- **Consequences:** You may experience anxiety due to a partial understanding of screening, breast cancer, or the risk of suffering from it. Alternatively, you may feel more satisfied given that you are aware of the consequences of testing and treatment for early breast cancer.

3) *Participation in Screening*

This marker state refers to participating in breast cancer screening or testing. In all situations, you will have an opportunity to express the value you place on the benefits and harms to health care professionals.

Participation in Screening or Testing – importance and utility rating

- **What you experience or feel:** You may receive a verbal or written invitation for mammography from a screening programme or a healthcare professional. The invitation will give you the details for having the mammography and information about the expected benefits and harms that you can obtain by participating in screening. Before or at the screening appointment, you can ask questions about this information and decide if you will participate in the screening programme. If you feel fully informed (described in a separate marker state) you might feel satisfied with the decision-making process.
- **Time Horizon:** Once you decide to participate in a screening programme, it may take a few days, weeks, or months before you undergo the test. If you receive an invitation for screening, it will usually take some weeks.
- **Testing and Treatment:** Depending on the results of the tests, additional testing and, if breast cancer is diagnosed, subsequent treatment may be required, or you may not require additional testing until the next time you are invited or decide to participate. You may receive tests or treatments that you and your doctor have decided are appropriate for you.
- **Consequences:** If you undergo a recommended test and your decision is based on the information you received, you may be satisfied (what satisfaction may mean to you is addressed in a separate marker state). If you are recalled for further assessment you may visit your healthcare professional again. If you are recalled for a further assessment, you will eventually be found to have or not have breast cancer. The clinical outcome may or may not extend your lifetime as a result of early detection of cancer.

4) *Informed Decision Making*

This marker state refers to you and your healthcare professional, together making healthcare decisions based on as much relevant information as possible.

Informed Decision – importance and utility rating

- **What you experience or feel**: You might feel empowered, confident, and satisfied with the decision-making process and the decision itself.
- **Time Horizon:** You may become more informed on the subject of breast cancer, breast cancer screening, diagnosis and treatment during the period for which you are receiving breast cancer healthcare. The amount of external influence on your decisions may vary over time.
- **Testing and Treatment:** The amount of knowledge you have before making a decision may affect the type and frequency of testing and treatment you may undergo.
- **Consequences:** You may ignore or be unaware about breast cancer information outside your current knowledge. You make the decision that is right for you, based on all available evidence and bearing in mind your values, priorities and lifestyle. However, you and your loved ones may occasionally feel uncomfortable, because of differences between your personal understanding and the advice from your healthcare professional, or because the new information overturns opinions you held previously.

5) *Satisfaction with Decision Making*

This marker state refers to the level of satisfaction you feel about the decision-making process and any decision that you and your healthcare provider have made about your breast cancer testing and/or treatment.

Satisfaction with Decision Making – importance and utility rating

- **What you experience or feel:** You may have the opportunity to provide input in your breast-related healthcare decisions. You may feel content with the process and the actual decision.
- **Time Horizon:** You may be content both immediately after information is presented to you and within a few days of making any decision related to testing and/or treatment. This feeling could disappear or change over time.
- **Testing and Treatment:** You may receive tests or treatments that are based on your informed decisions. Your satisfaction with the decisions made by you and your healthcare provider may affect the type and frequency of tests and/or treatments you undergo.
- **Consequences:** You may be satisfied with your breast healthcare. You may have less anxiety about your care and have a positive relationship with your healthcare provider.

6) *Confidence with Decision Making*

This marker state refers to making a decision (with consultation from your doctor) about your breast cancer-related healthcare with high confidence.

Confidence in Making Decisions – importance and utility rating

- **What you experience or feel:** You may have the opportunity to provide input in your breast cancer-related healthcare decisions. With high confidence in your decisions, you may feel satisfied in the decision-making process. With little confidence, you may feel dissatisfied.
- **Time Horizon:** You may start making breast cancer testing decisions weeks before your first regular screening or diagnostic test. You may be confident from that point onward.
- **Testing and Treatment:** Your confidence in the decisions made by you (and your healthcare professional) may affect the type and frequency of any screening or diagnostic tests you may undergo.
- **Consequences:** Additionally, you may ignore or be unaware about breast cancer information outside your current knowledge. Despite being confident, your decision may be right or wrong for you. However, it is more likely to be right for you if you have confidence in your decision.

7) *Abnormal Screening Results*

This marker state refers to any abnormal screening mammography result that requires you to be recalled for further diagnostic assessment. Your healthcare provider will organise this follow up (recall).

Abnormal Screening Result – Importance & Utility Rating

- **What you experience or feel:** When you are informed (in person, by phone or by letter) that a suspicious abnormality has been identified on the screening mammogram you may be concerned and anxious.
- **Time Horizon:** You will receive the results of your test and/or be recalled for further assessment within 1-2 weeks of your screening mammogram being performed.
- **Testing and Treatment:** Further assessment may include additional imaging, and eventual biopsy, and/or other testing; all of which may be performed by a specialist healthcare professional in an assessment centre or hospital. If cancer is diagnosed, you will be referred for treatment based upon the stage of your breast cancer, tumour biomarker status, age, and your general health. You may also be treated for anxiety arising from the disease.
- **Consequences:** You and your loved ones may experience periods of stress and anxiety because of uncertainty associated with being recalled and going through the experience of additional assessment. Going to additional assessments may necessitate taking time off work or other inconvenience. If the results suggest the possible presence of breast cancer you will be advised to have additional testing, biopsy, and, if breast cancer is diagnosed, treatment. If you have a biopsy, this may have physical side effects (see marker states 16, 18 and 19). You may feel relief if the assessment shows that the suspicious lesion turns out not to be cancer.

8) *Recall for Assessment*

This marker state refers to being recalled for further assessment due to abnormal mammographic findings (or technically inadequate mages) at the screening examination. Further assessment is needed to rule out or confirm breast cancer.

Recall for assessment – Importance & Utility Rating

What you experience or feel: When you are informed (by phone and/or letter) that a suspicious abnormality has been identified on the screening mammogram you may be concerned and anxious.

Time Horizon: You will receive the results of your test and/or be recalled for further assessment within 1-2 weeks of your screening mammogram being performed.

Testing and Treatment: Further assessment may include additional imaging, and eventual biopsy, and/or other testing; all of which may be performed by a specialist healthcare professional in an assessment centre or hospital. If cancer is diagnosed, you will be referred for treatment based upon the stage of your breast cancer, tumour biomarker status, age, and your general health. You may also be treated for anxiety arising from the disease.

Consequences: You and your loved ones may experience periods of stress and anxiety because of uncertainty associated with being recalled and going through the experience of additional assessment. Going to additional assessments may necessitate taking time off work or other inconvenience. If the results suggest the possible presence of breast cancer you will be advised to have additional testing, biopsy, and, if breast cancer is diagnosed, treatment. If you have a biopsy, this may have physical side effects (see marker states 16, 18 and 19). You may feel relief if the assessment shows that the suspicious lesion turns out not to be cancer.

9) *False-Positive Screening Result*

This marker state refers to the effects associated with having a screening mammogram that caused a recall for further assessment and therefore led you to believe you might have breast cancer when you do not.

False-Positive Screening Result– importance and utility rating

- **What you experience or feel:** When you are informed (by phone and/or letter) that a suspicious abnormality has been identified on the screening mammogram you may be concerned and anxious.
- **Time Horizon:** You will receive the results of your test and/or be recalled for further assessment within 1-2 weeks of your screening mammogram being performed.
- **Testing and Treatment:** Further assessment may include additional imaging, and eventual biopsy, and/or other testing; all of which may be performed by a specialist healthcare professional in an assessment centre or hospital. If you have a biopsy, this may have physical side effects (see marker states 16, 18 and 19).
- **Consequences:** You and your loved ones may experience anxiety and resource use. When you receive the result that there is no breast cancer on assessment, you may feel relief.

10) ***Suspicious Indeterminate Calcifications in Mammography***

This marker state refers to the state of having a diagnostic mammography result that identifies calcifications, which might be suggestive of breast cancer.

Suspicious Indeterminate Calcifications in Mammography – Importance & Utility Rating

- **What you experience or feel:** On your mammogram, a radiologist may detect calcifications suspicious of breast cancer. These radiological findings typically do not give symptoms. You may experience anxiety about the uncertainty of your diagnosis.
- **Time Horizon:** You will receive the results of your test and/or be recalled for further assessment within 1-2 weeks of your screening mammogram being performed.
- **Testing and Treatment:** Further assessment may include additional imaging, and eventual biopsy, and/or other testing; all of which may be performed by a specialist healthcare professional in an assessment centre or hospital. If you have a biopsy, this may have physical side effects (see marker states 16, 18 and 19). Depending on whether breast cancer is diagnosed, you may be advised to have treatment for breast cancer.
- **Consequences**: You and your loved ones may experience anxiety after you have been recalled for further assessment and during the time until the diagnosis is concluded and the decision about whether or not to have treatment is agreed upon.

11)False-Positive Biopsy Result

This marker state refers to the effects associated with having a biopsy result that led you to believe you might have breast cancer when you do not.

False-Positive Biopsy Result – importance and utility rating

- **What you experience or feel:** You think that you have breast cancer when in reality you do not. You may experience intense anxiety, and consequent physical symptoms such as sleeping problems, as a result of having to undergo a biopsy for a possible breast cancer. After you realize that you were given a false positive diagnosis you may experience relief and anger.
- **Time Horizon:** Times for identifying a false positive diagnosis vary according to the type of lesion and the procedures at your breast cancer assessment centre or hospital. A false positive diagnosis is likely to be identified within a few weeks of the biopsy. You may experience anxiety (among other symptoms) during the time you believe you have breast cancer. You may also continue to worry after being told that the result was inaccurate and that you do not have breast cancer.
- **Testing and Treatment:** The biopsy may take place in a breast assessment centre or hospital by a healthcare professional. Generally, false positive breast biopsies are very rare. As a result of the false positive biopsy, you may undergo surgery and removal of breast tissue. In very rare circumstances, your entire breast may be removed.
- **Consequences:** If you are having surgery, you may experience swelling, soreness of the skin or infection in the area of the tissue sample collection. You may experience unnecessary cosmetic damage to your breast and/or loss of your breast as a result of any surgery. You and your loved ones may experience anxiety and may feel frustrated due to unnecessary resource use.

12)Breast Cancer Detection

This marker state refers to the correct diagnosis of breast cancer after a positive mammogram followed by further diagnostic assessment and tests.

Breast Cancer Detection – Importance & Utility Rating

- **What you experience or feel:** When you are told you have breast cancer, you may experience considerable anxiety, which in turn may cause physical symptoms such as sleeping problems. However, you may feel relieved if your breast cancer was detected in an early stage. You may experience considerable uncertainty about whether your cancer is likely to develop and requires treatment.
- **Time Horizon:** The diagnosis of breast cancer is confirmed at the end of the assessment process. This includes full histopathological assessment of the tissue that has been removed from your breast. The whole process may take 1 to 4 weeks from obtaining the results of your screening mammogram. You may begin to experience emotional symptoms after receiving your screening result, indicating the possibility that you may have breast cancer.
- **Testing and Treatment:** After confirmation of breast cancer, your diagnosis and treatment options may be discussed by a multidisciplinary team. You may be referred for further diagnostic testing to determine the extent of the cancer in your body. The multidisciplinary team may propose a targeted treatment which may vary according to the stage of your breast cancer, tumour biomarker status, age and your general health.
- **Consequences:** During the time that your treatment plan is being formulated by the multidisciplinary team you may feel additional stress and anxiety.

13)Breast Cancer Stage

This marker state refers to the state of having any stage of breast cancer. An early stage indicates that the breast tumour is relatively small and has not spread to other parts of the body. This means that you may be offered less aggressive treatment and may have a better prognosis. A later stage indicates that the breast cancer has reached a greater size and/or has spread to regional lymph nodes or to other parts of the body. This usually requires more aggressive treatment and is associated with a worse prognosis. In addition to tumour size and extent, prognosis and treatment will also depend on the characteristics of the tumour including the histological grade and the biomarker status.

Breast Cancer Stage – Importance & Utility Rating

- **What you experience or feel:** When you are told you have breast cancer, you may experience considerable anxiety, which in turn may cause physical symptoms such as sleeping problems. Due to presence of a breast cancer, you may also experience symptoms such as breast skin thickening, changes to breast size, shape or appearance or nipple discharge. If the cancer has spread to other parts of the body you may feel a lump under your arm or symptoms referable to body sites involved by tumour. These symptoms may not be present at all and if present may vary in intensity. If you have early stage breast cancer you may experience relief that it is been detected early.
- **Time Horizon:** The amount of time it takes for a cancer to go from an early to a late stage varies from months to years.
- **Testing and Treatment:** A sample of your breast tissue may be removed with a needle to make a diagnosis of your breast cancer (please see marker states 16, 18 and 19). Further testing such as ultrasound, bone scan, computerised tomography, MRI and/or a PET scan (positron emission tomography) may be performed to assess the stage of your breast cancer. You will be referred for treatment based upon the results of the tests. Treatment will vary according to stage of your breast cancer, tumour biomarker status, age, and your general health.
- **Consequences:** Your breast cancer may shorten your life. Breast cancer detected at an early stage will be more likely to be cured than breast cancer detected at a late stage. You and your loved ones may experience anxiety.

14)Determination of Tumour Biomarker Status in Biopsy

The biomarker status of a tumour refers to the expression or otherwise of certain proteins by the the tumour. Expression of these features by a breast tumour predicts how the tumour may behave and more specifically how it might respond to specific treatment. The most important tumour biomarkers are expression of estrogen/progesterone hormone receptors and the HER2 (human epidermal growth factor receptor 2) oncogene. Some centres also assess the Ki67 index of the tumour to see how fast it is growing and to assist decision making regarding the need for chemotherapy.

Determination of Tumour Biomarker Status – importance and utility rating

- **What you experience or feel:** You do not feel the expression of a tumour biomarker. You may experience relief if your biomarker status suggests a relatively good prognosis or if the biomarker status allows a targeted therapy directed against the tumour. However, you might be concerned if the biomarker suggests a possibly worse outcome.
- **Time Horizon:** You will receive results of testing for the tumour biomarker within approximately 10 days of the biopsy procedure.
- **Testing and Treatment:** Your biomarker status will be determined using immunohistochemical and in situ hybridization techniques. The tests are performed in a histopathology laboratory. A multidisciplinary team will discuss your treatment options. The presence of certain biomarkers in a breast cancer will have an impact on the type of treatment that you will be offered. Expression of estrogen/progesterone receptors suggests you may benefit from endocrine therapy. Expression of HER2 suggests you may benefit from anti-HER2 therapy. If none of the biomarkers is expressed you may benefit from an alternative type of chemotherapy.
- **Consequences:** You may experience anxiety in the time between having a biopsy performed and receiving results of your biomarker status. The results will have an impact on the type of treatment you receive. They also influence your chances of being cured of breast cancer.

15)Interval Breast Cancer

This marker state refers to having a diagnostic test correctly identify a cancer after you have had a screening test, with or without further assessment, which was negative for malignancy, either: before the next invitation to screening; or within a time period equal to the screening interval after you have reached the upper age limit for screening.

Interval Cancer – Importance & Utility Rating

- **What you experience or feel:** When you are told you have breast cancer, you may experience considerable anxiety, which in turn may cause physical symptoms such as sleeping problems. You may feel relieved if your breast cancer was detected in an early stage. Due to the presence of breast cancer you may experience symptoms such as a breast lump, nipple discharge, skin thickening or a change in the size, shape or appearance of your breast. You may also feel concern that your tumour may have been present at the time of screening and was not detected.
- **Time Horizon:** This tumor may have become symptomatic in the period of time since your prior screening examination. The methods of assessment used to identify the tumor and confirm the diagnosis, including the time taken, are outlined in marker states 16, 18, 19, 20, 21 and 22 above.
- **Testing and Treatment:** Following the mammogram, additional mammographic views, ultrasound, MRI and/or contrast enhanced mammography (CESM) may be performed for further assessment of your breast. This will be carried out in a hospital or in a breast centre. Treatment will vary according to the stage of your breast cancer, tumour biomarker status, age, and your general health.
- **Consequences:** Since the tumor was not visible at prior screening it might be fast growing and biologically more likely to spread. However, it is possible that your tumour is still at an early stage. Your breast cancer may shorten your life. Breast cancer detected at an early stage will be more likely to be cured than breast cancer detected at a late stage. You and your loved ones may experience anxiety.

16) *Over-diagnosis and Over-treatment*

In screening, it is possible to diagnose a breast cancer which is so slow-growing that it would never have been diagnosed in a person's lifetime if the person had not been screened. The scientific term for breast cancer that would have not been diagnosed without screening is "over-diagnosis" of cancer. We cannot tell which cancers are of this type, however. Because it is unknown which cancers are over-diagnosed, treatment is the same as if it was not over-diagnosed. This is referred to as over-treatment. An over-diagnosed cancer is likely to be detected at an early stage.

Over-diagnosis and over-treatment – Importance & Utility Rating

- **What you experience or feel:** When you are told you have breast cancer, you may experience considerable anxiety, which in turn may cause physical symptoms such as sleeping problems. However, you may feel relieved if your breast cancer was detected in an early stage. You may experience considerable uncertainty about whether your cancer is likely to develop and requires treatment.
- **Time Horizon:** The time between receiving the diagnosis due to a recall from screening and receiving treatment is the same whether or not the cancer is over-diagnosed. If treatment is confined to local therapy, it is completed in 6-8 weeks. Other therapy, such as hormone therapy can last several years. If you had not participated in screening, you would have remained unaware of the cancer and free of symptoms throughout your normal lifetime.
- **Testing and Treatment:** The screening mammography is performed in a breast screening centre by a healthcare professional. Due to suspicious findings on your mammogram, you will be called for further assessment at a breast cancer assessment centre or a hospital. Detection of the cancer will not be beneficial to your health because your tumour is of no clinical importance. You will be referred for treatment based upon the results of the assessment. Treatment will vary according to stage of your breast cancer, tumour biomarker status, age, and your general health.
- **Consequences:** Any treatment you receive may have side effects (described in other marker states). You will have to return to your healthcare professional for additional diagnostic testing and treatment. You and your loved ones may experience anxiety and costs compared to if the breast cancer had never been diagnosed.

17) *False-Negative Screening Result*

This marker state refers to receiving a negative screening result (no breast cancer) when you actually have a breast cancer. This is called a false negative screening result. Not all women become aware that they had a false negative screening result. This marker state describes when they do become aware after subsequent diagnosis.

False Negative Screening Result – importance and utility rating

- **What you feel or experience:** When you find out that you did have breast cancer and it was missed, you are likely to feel anger, fear, and anxiety.
- **Time Horizon:** It may take months to years before you find out that you did have breast cancer when you were told you did not.
- **Testing and Treatment:** Following the discovery of your breast cancer later on, you may have to undergo treatment that is more intense than if the cancer had been detected right away, as the cancer may have developed to a more advanced stage.
- **Consequences:** The consequences of late detection of a slow growing breast cancer will probably be not substantial with respect to treatment and prognosis. However, if the breast cancer has grown, your predicted outcome is likely worse than if it had been diagnosed at the screen. Survival from breast cancer that has a false-negative diagnosis may be worse compared to women with screen-detected breast cancer, but comparable to women who do not attend screening.

18)Radiation Exposure from Mammograms & Other Assessments Using Radiation

This marker state refers to being exposed to any dose of radiation from undergoing a mammographic examination and any other related assessments only. It does not refer to therapeutic radiation.

Radiation Exposure from Mammograms & Other Assessments Using Radiation – Importance & Utility Rating

- **What you experience or feel**: You do not feel the radiation itself. However, you may be anxious if you are not aware that the radiation dose is low or if you feel concerned at the prospect of any radiation dose associated with the examination.
- **Time Horizon:** Considering the low doses of radiation, no short-acting effects occur. In extremely rare cases, exposure to radiation may induce cancer in your breast. This may take many years.
- **Testing and Treatment:** You will be brought to a mammography device so images of your breast can be taken. Your breast will be placed on a plate and compressed to have a mammogram. Compression is needed to flatten the breast which will keep the radiation dose as low as is reasonably achievable.
- **Consequences:** Exposing your breast to radiation may induce cancer in the breast tissue. The scale of the harm is extremely small and difficult to quantify. It will increase with the number of mammograms over a lifetime.

19)Provision of Surgical Therapy

This marker state refers to the state of undergoing surgery to the breast or axilla. This includes breast conserving surgery (removal of a breast lump with a rim of surrounding tissue), mastectomy (complete removal of your breast), open biopsy (removal of a small piece of tissue from your breast for diagnosis) and axillary surgery (removal of one or more lymph nodes, including the sentinel lymph node). It does not refer to any combination therapy.

Provision of Surgical Therapy – Importance & Utility Rating

- **What you experience or feel:** You may experience anxiety and fear because of the procedure that will be performed. If breast conserving surgery (lumpectomy or quadrantectomy) or mastectomy is performed, you may experience loss of part or all of your breast and that may have an influence on your physical and psychological well-being. Preparation for surgery may involve other examinations and tests.
- **Time Horizon:** Surgery will be planned and scheduled. It may take weeks (or months if you receive chemotherapy prior to surgery) before the surgery is performed. The time taken for the operation will vary depending on the type of surgery and will be longer if you undergo reconstructive surgery at the same time.
- **Testing and Treatment:** All surgeries will be performed in an operating room. For breast conserving surgery or a mastectomy, you will be given general anesthesia, so you will be asleep. During the surgery, 1-2 incisions may be made in your breast. Some of your breast tissue (or entire breast) and, lymph nodes, and/or chest muscle may be removed depending on the type and stage of your cancer. This will be discussed with you by your surgeon before surgery. Following surgery, a histopathologist will examine the breast and axillary tissue that has been removed to analyze the tumour with regard to size, grade, type etc. The histopathologist will also examine the lymph nodes to see if the tumour has spread to these.
- **Consequences:** After the procedure, you may experience bruising, infection, haematoma, and/or tenderness of the breast. In rare cases, you may experience collapse of the lung. Additionally, you may have discomfort, inconvenience, embarrassment, and reduced self-esteem because of the loss of all or part of your breast, although this may be mitigated by reconstructive surgery.

20)Mastectomy

This marker state refers to having any type of mastectomy performed. This is usually accompanied by removal of one or more axillary lymph nodes.

Mastectomy – Importance & Utility Rating

- **What you experience or feel:** Before surgery you may be anxious and afraid. After surgery, you may experience pain. You may be concerned about the loss of your breast and how it will appear to other people.
- **Time Horizon:** The procedure takes approximately 2 - 3 hours. It may take longer if reconstruction of your breast is included as part of the surgical procedure. You will be admitted to a hospital and stay for approximately 1- 3 days if there are no complications. The remainder of your recovery may take place in your home. Your discomfort will disappear over the next weeks.
- **Testing and Treatment:** Your mastectomy will be performed by a breast surgeon or senologist at a hospital. You will be put under general anesthesia, so you will be asleep. During the surgery, a cut will be made into your breast and armpit (axilla), according to your pre-surgical discussion with your breast surgeon or senologist. Axillary lymph nodes will likely be removed in addition to your breast.
- **Consequences:** The planning of the procedure may make you feel anxious. After the procedure, you may experience pain related to the wound, bruising and breast tenderness. Occasionally you may experience infection, haematoma, and rarely lung collapse. You will not be able to conduct physical exercise or heavy lifting for a few weeks after the surgery. Additionally, you may have long-term discomfort, inconvenience, embarrassment, expenses, and reduced self-esteem for cosmetic reasons, although this may be mitigated by reconstructive surgery.

21)Provision of Medical Therapy

This marker state refers to the state of receiving medical therapy for breast cancer treatment. This includes, but is not limited to chemotherapy or hormonal therapy. Counselling and psychological evaluation may be provided to support the psychological burden of breast cancer.

Provision of Medical Therapy – Importance & Utility Rating

- **What you experience or feel:** During the course of the treatment you may experience anxiety, fear, or a feeling or sense of confusion.
- **Time Horizon:** You may begin treatment as early as within one week of diagnosis. The duration of your treatment will vary according to the type of treatment you are receiving.
- **Testing and Treatment:** Medical treatments may include pills, injections and infusions. More invasive or aggressive treatments will take place in your breast cancer centre or hospital. You may be referred to a psychiatrist for evaluation or psychotherapy in combination with your medical therapy.
- **Consequences:** During the course of treatment, you may have to visit your healthcare professional frequently. Medications and various forms of treatment may cause side effects (described in other health marker states).

22)Provision of Radiotherapy

This marker state refers to the state of receiving radiotherapy after surgery to reduce the risk of local breast cancer recurrence. This includes, but is not limited to external beam breast radiation, internal breast radiation, or brachytherapy. It does not refer to any combination therapy.

Provision of Radiotherapy – Importance & Utility Rating

- **What you experience or feel:** You may experience feelings of anxiety when you undergo radiotherapy. Additionally, you may experience fatigue, or skin irritation at the site of radiotherapy.
- **Time Horizon:** You may experience symptoms within hours of exposure. However, generally the amount of time between radiation and the onset of radiation exposure symptoms is dependent upon how much radiation you have been exposed to. Symptoms may occur months or even years after the treatment.
- **Testing and Treatment:** You will visit a radiotherapy clinic for your radiotherapy. During each session of treatment, you will lie under a machine that applies radiation to your breast to kill cancerous cells, potentially still present after surgery.
- **Consequences:** From hours to years after receiving radiotherapy at your breast, you may experience infections, itchiness, bone weakening, skin cancer, and low blood pressure after radiation exposure. Additionally, very few women may develop lung symptoms such as breathlessness, cardiovascular disease as a result of cumulative radiation exposure to the left breast or have a small risk of other cancers.

23)Provision of Chemotherapy

This marker state refers to the state of receiving chemotherapy alone.

Provision of Chemotherapy – Importance & Utility Rating

- **What you experience or feel:** During the course of the treatment you may experience fatigue, pain, hair loss, mouth and throat sores, diarrhea, nausea, vomiting, constipation, bleeding, infections and nervous system effects such as numbness or tingling. The severity of your symptoms may vary from very little to severe.
- **Time Horizon:** Each individual chemotherapy treatment may last up to 3 or 4 hours. You may experience nausea and vomiting within a few hours of every chemotherapy treatment. Other symptoms may occur within days to months.
- **Testing and Treatment**: For oral chemotherapy, you can take the medication yourself at home. If you are receiving intravenous therapy you will be given the drug through a needle inserted into one of your veins. This type of chemotherapy is normally performed in your healthcare professional's clinic. You will have physical examinations and blood samples taken. You may also have further radiological tests to assess response to treatment. If you suffer a complication, e.g. an infection, you will receive treatment for it.
- **Consequences:** During the course of treatment, you may have to visit your healthcare professional frequently and your quality of life may decrease. You may experience anxiety. Rarely you may suffer permanent impairment from a complication of treatment.

24)Other-Cause Mortality

This marker state refers to the state of being dead due to factors unrelated to your breast cancer. It does not refer to the process of dying or outcomes that precede it (e.g. the breathlessness related to it or pain).

Other Cause Mortality – Importance & Utility Rating

- **What you experience or feel:** You are dead and feel no pain. You may experience symptoms prior to dying from causes other than breast cancer but you do not feel those when you are dead.
- **Time Horizon:** Before you die, you experience other states of disease of varying duration.
- **Testing and Treatment:** Tests and treatment will have ceased.
- **Consequences:** You lose your vital bodily and mental functions, ending your life.

2.6. References

1. Guidelines, I.o.M.C.o.S.f.D.T.C.P., *Clinical Practice Guidelines We Can Trust*. 2011: National Academies Press.
2. Brouwers, M.C., et al., *AGREE II: advancing guideline development, reporting and evaluation in health care.* J Clin Epidemiol, 2010. **63**(12): p. 1308-11.
3. Schunemann, H.J., et al., *Guidelines 2.0: systematic development of a comprehensive checklist for a successful guideline enterprise.* CMAJ, 2014. **186**(3): p. E123-42.
4. Shekelle, P., et al., *Developing clinical practice guidelines: reviewing, reporting, and publishing guidelines; updating guidelines; and the emerging issues of enhancing guideline implementability and accounting for comorbid conditions in guideline development.* Implement Sci, 2012. **7**: p. 62.
5. Eccles, M.P., et al., *Developing clinical practice guidelines: target audiences, identifying topics for guidelines, guideline group composition and functioning and conflicts of interest.* Implement Sci, 2012. **7**: p. 60.
6. Woolf, S., 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**: p. 61.
7. Qaseem, A., et al., *Guidelines International Network: toward international standards for clinical practice guidelines.* Ann Intern Med, 2012. **156**(7): p. 525-31.
8. Langer, G., et al., *[GRADE guidelines: 1. Introduction - GRADE evidence profiles and summary of findings tables].* Z Evid Fortbild Qual Gesundhwes, 2012. **106**(5): p. 357-68.
9. Guyatt, G.H., et al., *GRADE guidelines 6. Rating the quality of evidence--imprecision.* J Clin Epidemiol, 2011. **64**(12): p. 1283-93.
10. Guyatt, G.H., et al., *GRADE guidelines: 2. Framing the question and deciding on important outcomes.* J Clin Epidemiol, 2011. **64**(4): p. 395-400.
11. Guyatt, G., et al., *GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables.* J Clin Epidemiol, 2011. **64**(4): p. 383-94.
12. Carrasco-Labra, A., et al., *Improving GRADE evidence tables part 1: a randomized trial shows improved understanding of content in summary of findings tables with a new format.* J Clin Epidemiol, 2016. **74**: p. 7-18.
13. Langendam, M., et al., *Improving GRADE evidence tables part 2: a systematic survey of explanatory notes shows more guidance is needed.* J Clin Epidemiol, 2016. **74**: p. 19-27.
14. Santesso, N., et al., *Improving GRADE evidence tables part 3: detailed guidance for explanatory footnotes supports creating and understanding GRADE certainty in the evidence judgments.* J Clin Epidemiol, 2016. **74**: p. 28-39.
15. Guyatt, G.H., D.H. Feeny, and D.L. Patrick, *Measuring health-related quality of life.* Ann Intern Med, 1993. **118**(8): p. 622-9.
16. Group, G.W., *Handbook for grading the quality of evidence and the strength of recommendations using the GRADE approach.* 2013.
17. Torrance, G.W., *Measurement of health state utilities for economic appraisal.* J Health Econ, 1986. **5**(1): p. 1-30.

18. Schunemann, H.J., et al., *A clinical trial to evaluate the measurement properties of 2 direct preference instruments administered with and without hypothetical marker states.* Med Decis Making, 2003. **23**(2): p. 140-9.
19. Schunemann, H.J., et al., *A comparison of narrative and table formats for presenting hypothetical health states to patients with gastrointestinal or pulmonary disease.* Med Decis Making, 2004. **24**(1): p. 53-60.
20. Llewellyn-Thomas, H.A., *Health state descriptions. Purposes, issues, a proposal.* Med Care, 1996. **34**(12 Suppl): p. DS109-18.
21. Berglund, G., et al., *Late effects of adjuvant chemotherapy and postoperative radiotherapy on quality of life among breast cancer patients.* Eur J Cancer, 1991. **27**(9): p. 1075-81.
22. Bilodeau, B.A. and L.F. Degner, *Information needs, sources of information, and decisional roles in women with breast cancer.* Oncol Nurs Forum, 1996. **23**(4): p. 691-6.
23. Dow, K.H., et al., *An evaluation of the quality of life among long-term survivors of breast cancer.* Breast Cancer Res Treat, 1996. **39**(3): p. 261-73.
24. Ferrell, B.R., et al., *Quality of life in breast cancer. Part I: Physical and social well-being.* Cancer Nurs, 1997. **20**(6): p. 398-408.
25. Ferrell, B.R., et al., *Quality of life in breast cancer. Part II: Psychological and spiritual well-being.* Cancer Nurs, 1998. **21**(1): p. 1-9.
26. Gram, I.T., E. Lund, and S.E. Slenker, *Quality of life following a false positive mammogram.* Br J Cancer, 1990. **62**(6): p. 1018-22.
27. Hack, T.F., et al., *Do patients benefit from participating in medical decision making? Longitudinal follow-up of women with breast cancer.* Psychooncology, 2006. **15**(1): p. 9-19.
28. Lerman, C., et al., *Psychological and behavioral implications of abnormal mammograms.* Ann Intern Med, 1991. **114**(8): p. 657-61.
29. Lerman, C., et al., *Psychological side effects of breast cancer screening.* Health Psychol, 1991. **10**(4): p. 259-67.
30. Sutton, S., et al., *Does routine screening for breast cancer raise anxiety? Results from a three wave prospective study in England.* J Epidemiol Community Health, 1995. **49**(4): p. 413-8.
31. Braun, V. and V. Clarke, *What can "thematic analysis" offer health and wellbeing researchers?* Int J Qual Stud Health Well-being, 2014. **9**: p. 26152.
32. Green, C., J. Brazier, and M. Deverill, *Valuing health-related quality of life. A review of health state valuation techniques.* Pharmacoeconomics, 2000. **17**(2): p. 151-65.
33. Torrance, G.W., D. Feeny, and W. Furlong, *Visual analog scales: do they have a role in the measurement of preferences for health states?* Med Decis Making, 2001. **21**(4): p. 329-34.
34. Froberg, D.G. and R.L. Kane, *Methodology for measuring health-state preferences--III: Population and context effects.* J Clin Epidemiol, 1989. **42**(6): p. 585-92.

CHAPTER 3: Conclusion

In this work we developed health marker state descriptors with a real guideline panel, used them for health utility rating with the panel, and analyzed panel feedback on the entire process. We also analyzed the health utility ratings resulting from use of the health marker state descriptors.

This work is part of an ongoing effort to further develop the GRADE approach, thereby improving guideline development methods [1]. The results from this study will inform methods used by guideline developers to synthesize evidence and improve transparency of guideline development methods.

3.1 Summary of Findings

We used our template to develop 21 health marker state descriptors in the context of guideline development. Each health marker state was successfully used in combination with a VAS to conduct a health utility assessment with guideline panel members.

Lack of sufficient training on health marker state descriptor development methods and the GRADE approach was a barrier to the panel's participation in development, which was initially low. Once participation increased, online feedback and in-person feedback were effective and easy methods for refining outcome-specific content during development. This is consistent with findings that

online collaboration tools are useful for facilitating groupwork and have become common in guideline development [2, 3].

The panel experienced challenges rating health utility using a VAS, and error was identified in the health utility rating exercise. This is inconsistent with research on utility scaling methods that deem it to be an easy and acceptable technique [4-6]. Therefore, we attributed the difficulties to contextual bias, poor survey instructions, or methodological issues during outcome generation (which occurred prior to the study).

Overall, panel members thought our presentation for describing health outcomes was most useful for harmonizing understanding of the outcomes among panel members and improving transparency of guideline methods. Most panel members supported Llewellyn-Thomas' proposal that outcome descriptors should be tailored to facilitate use by end-users [7]. Interestingly, opinions on the descriptor attributes, such as appropriate balance of wording, reading level, and emotional sensitivity for end-users, were varied among panel members. Our preference is that simple descriptions should be used that provide a common language for those providing health care and those receiving that care. Currently, there are no logical reasons to use a different language for different people. Using a common language will reduce the probability that misunderstandings will occur.

3.2 Implications

Guidance on best methods for standardizing outcome descriptors has been lacking, particularly as it relates to guideline development. Previous research on health marker state descriptors revealed that short bulleted or table formats were best for presenting health outcomes to patients [8, 9]. This study builds upon prior work by further developing the short-bulleted format and informing best practice for development and use of health marker state descriptors during guideline development.

Our results suggest that health marker state development is most efficient when developers work with a multidisciplinary subgroup of guideline panel members at each stage. Panel input should be collected through a few rounds of online or in person feedback. To prepare panel members for the feedback process, they should be trained on health marker state descriptor development methods prior to development. Most guideline panels are trained on guideline development methods by the guideline organization and so we expect implementation of health marker state descriptor training to be relatively easy [3]. Our findings also suggest that changes should be made to GRADE training to facilitate better understanding and execution of outcome generation and importance rating exercises, which are crucial for health marker state descriptor development and overall guideline development [10].

In addition to the internal guideline development uses for our health marker state descriptors, the ECIBC breast guideline health marker state descriptors will be published on the ECIBC web platform to improve guideline transparency for guideline end-users. There may also be clinical and research applications for published health marker state descriptors. We believe that health care providers might be able to use health marker state descriptors to inform shared decision-making with healthcare recipients. Health marker state descriptors might also be used to present health outcomes in the context of research. This may of interest to research groups such as COMET, who may wish to use the results of this study to improve development and presentation of outcome sets [11].

3.3 Limitations and Strengths of Work

One of the major challenges in this study was the timing of health marker state descriptors development relative to the progress of guideline development. We identified heterogeneity in outcome definitions after the GDG had already rated outcome importance. The timing of development may have caused confusion about the need and purpose of health marker state descriptors in the guideline development process, although the development need resulted from disagreement during that rating exercise. Furthermore, health marker state descriptor development occurred in the context of only the European breast cancer screening and diagnosis guideline, which limits our generalization to other panels and healthcare topics. Finally, this study had a small sample size all together, and the response rate of the online utility ratings surveys in this study was poor. The

relatively small number of pairwise comparisons for each health outcome reduced the statistical power of our analyses.

A strength of this study is that all data was collected from a real-life guideline panel, which is rare among published literature on outcome descriptions. By conducting this case study in the context of a real guideline panel, our results can be used to inform outcome descriptor standardization efforts for guideline development, where we originally identified the problem of heterogeneity. We also carefully planned health marker state descriptor development methods and interaction with GDG members to capture reliable panel feedback at each stage of health marker state descriptor development. Collectively, our planning and analysis ensure that the results from this study can be used to inform all stages of health marker state descriptor development.

3.4 Further Research

The primary goal of future research efforts should be to further develop the format of health marker state descriptors to maximize usefulness to guideline developers and end-users.

In this study we were unable to draw conclusions regarding health marker state descriptor attributes, such as appropriate balance of wording, reading level, and emotional sensitivity. Collectively, it is likely that these attributes will influence usefulness of health marker state descriptors, as proposed by Llewellyn-Thomas [7]. Therefore, these attributes are issues that by be investigated further by

researchers. Special emphasis should be put on investigating these issues in the context of healthcare guidelines.

To maximize usefulness of health marker state descriptors for guideline developers and end-users, researchers first must assess how those populations might use them. This study examined the internal use of health marker state descriptors by guideline developers. Therefore, future research efforts should expand upon our work and investigate how healthcare professionals and healthcare recipients might use health marker state descriptors as end-users.

Given that health outcomes are used in fields of work other than guidelines (e.g. research, policy, etc.), we suspect that health marker state descriptors can be used for more than developing guidelines [11, 12]. Other research efforts might focus on how health marker state descriptors might be adapted for use for other purposes and populations including but not limited to research, and education.

Researchers should also concentrate efforts on determining the reliability of the VAS when health marker state descriptors are used, because we were unable draw meaningful conclusions about this due to limited statistical power. Such research should include healthcare recipients who may not have medical knowledge or experience with illness, since health status has been shown to influence outcome utility ratings [13].

3.5 Final remarks

The work in this thesis further developed methods for standardizing health outcome descriptors. It provides guidance on how to develop health marker state descriptors and use them for outcome health utility assessment in the context of guideline development.

3.6 References

1. Guyatt, G., et al., *GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables.* J Clin Epidemiol, 2011. **64**(4): p. 383-94.
2. CHU , S. *MEDIAWIKI AND GOOGLE DOCS AS ONLINE COLLABORATION TOOLS FOR GROUP PROJECT CO-CONSTRUCTION*. in *Proceedings of the 2009 International Conference on Knowledge Management*. 2009.
3. Burgers, J.S., et al., *Towards evidence-based clinical practice: an international survey of 18 clinical guideline programs.* Int J Qual Health Care, 2003. **15**(1): p. 31-45.
4. Green, C., J. Brazier, and M. Deverill, *Valuing health-related quality of life. A review of health state valuation techniques.* Pharmacoeconomics, 2000. **17**(2): p. 151-65.
5. Torrance, G.W., *Measurement of health state utilities for economic appraisal.* J Health Econ, 1986. **5**(1): p. 1-30.
6. Torrance, G.W., D. Feeny, and W. Furlong, *Visual analog scales: do they have a role in the measurement of preferences for health states?* Med Decis Making, 2001. **21**(4): p. 329-34.
7. Llewellyn-Thomas, H.A., *Health state descriptions. Purposes, issues, a proposal.* Med Care, 1996. **34**(12 Suppl): p. DS109-18.
8. Schunemann, H.J., et al., *A clinical trial to evaluate the measurement properties of 2 direct preference instruments administered with and without hypothetical marker states.* Med Decis Making, 2003. **23**(2): p. 140-9.
9. Schunemann, H.J., et al., *A comparison of narrative and table formats for presenting hypothetical health states to patients with gastrointestinal or pulmonary disease.* Med Decis Making, 2004. **24**(1): p. 53-60.
10. Guyatt, G.H., et al., *GRADE guidelines: 2. Framing the question and deciding on important outcomes.* J Clin Epidemiol, 2011. **64**(4): p. 395-400.
11. Williamson, P.R., et al., *The COMET Handbook: version 1.0.* Trials, 2017. **18**(Suppl 3): p. 280.
12. Navarro, V., et al., *Politics and health outcomes.* Lancet, 2006. **368**(9540): p. 1033-7.
13. Froberg, D.G. and R.L. Kane, *Methodology for measuring health-state preferences--III: Population and context effects.* J Clin Epidemiol, 1989. **42**(6): p. 585-92.