

METHODS FOR OPTIMIZING EVIDENCE SYNTHESSES OF
COMPLEX INTERVENTIONS: CASE STUDY OF A SYSTEMATIC
REVIEW AND META-ANALYSIS OF DIABETES QUALITY
IMPROVEMENT TRIALS

KRISTIN JULIANNA DANKO

Thesis submitted to the University of Ottawa
in partial fulfillment of the requirements for the
degree of Doctor of Philosophy in Epidemiology

School of Epidemiology and Public Health
Faculty of Medicine
University of Ottawa

ABSTRACT

Healthcare decision-makers need high quality evidence to inform policy and practice decisions. Systematic reviews of randomized controlled trials (RCTs), including meta-analyses of study effects, are considered one of the highest forms of evidence to inform such decisions. Most applications of systematic reviews and meta-analyses are based on a standardized cannon of methods that seek to collect, abstract, assess, and synthesize evidence from primary studies to produce a comprehensive and unbiased summary of the evidence. While useful, standard synthesis methods tend to assume simple data structures (e.g., two-arm comparison of a single intervention vs. a similar control evaluated in a parallel individual randomized design) and some practices (e.g., author contact) may not always be supported by empirical evidence.

Complex interventions are of increasing focus in healthcare and public health and pose challenges to the standard methods of systematic review and meta-analysis. While different definitions of complex interventions have been proposed, most definitions assume: i) multiple intervention ‘components’ that may or may not interact with each other to increase or decrease observed intervention effects and ii) effect modification by study-specific characteristics (e.g., healthcare setting, patient population). At least three challenges may result from this complexity. First, reviewers will likely have to contact authors for additional information about intervention components and contextual factors that may operate as effect modifiers. Unfortunately, evidence supporting optimal strategies for achieving response from author contact is lacking. Second, complex interventions are often evaluated using a cluster randomized trial (CRT) design that randomize units of patients to different healthcare/health policy interventions. Analyses from CRTs that are not adjusted for the clustering effect are said to have unit of analysis errors, which if incorporated in meta-analyses could lead to biased summary estimates and overly precise confidence intervals (CIs). Methods for reviewers to appropriately appraise abstract evidence from CRTs are limited. Thirdly, standard meta-analyses estimate an overall effect of a singular ‘complex intervention’. Such analyses answer the question *“Do complex interventions as a whole lead to a difference in observed*

outcomes?” and tend to exhibit high statistical heterogeneity since variation in intervention components and effect modifiers are not accounted for. Hierarchical multivariate meta-regression models have been proposed as an alternative synthesis approach for complex interventions to better account for observed heterogeneity and answer the question decision-makers are really interested in; that is “*What component(s) (or combination of components) work and under what conditions?*”. Hierarchical multivariate meta-regression models however have yet to be applied in the review of complex healthcare interventions. The overall aim of my doctoral research was to explore the utility of three methodological approaches to address these challenges and optimize the synthesis of complex interventions using a large systematic review of diabetes quality improvement interventions as a case study.

The first objective of this thesis was to do an RCT evaluation of the effect of telephone call versus repeated email contact of non-responding authors for additional study information on response rates and research costs. We found authors contacted by telephone call were more likely to complete requests for additional information (response rate 36.7% vs. 20.2%; adjusted odds ratio 2.26 [95% CI 1.10-4.76]) but the intervention took more time to deliver in total (20 vs. 10 hours over several months vs. one month) and was more expensive overall (approximately \$505 vs. \$253).

The second objective of this thesis was to better account for evidence from CRTs and involved a descriptive study and a methodological study. The descriptive study described the proportion of studies with unit of analysis errors and the nature of the error (inappropriate analysis versus unclear or incomplete reporting). The methodological study investigated the utility of building a database of intracluster correlation coefficients (ICCs) and use of an ICC posterior predictive distribution model to correct unit of analysis errors identified in the descriptive study. We found that although trials often adjusted for the cluster effect (67% across outcomes; range 25%-81%), most did not report enough information to extract adjusted effect estimates required for meta-analysis (an average of 77% of studies with remaining unit of analysis errors across outcomes; range 42%-100%). We were able to construct a posterior predictive distribution of the ICC for most outcomes in our review using estimates of the ICC

obtained from the descriptive study combined with external estimates and use these distributions to impute missing ICCs to correct unit of analysis errors.

Finally, the third objective of this thesis was to illustrate the use of hierarchical multivariate meta-regression for quantitative synthesis when estimating the effects of complex interventions and exploring effect heterogeneity. Using an arm-based analysis of post-treatment means of one continuous outcome, we demonstrated that hierarchical multivariate meta-regression models can be used to estimate a 'response surface' that accounts for complex intervention multiple components and study characteristics, and these models can be used to infer estimates of component effects, interactions among components, and effect modification by study covariates.

Collectively the results from this thesis suggest three methodological approaches (contacting authors by telephone, imputing missing ICCs using a predictive distribution, estimating complex intervention effects using a hierarchical multivariate meta-regression) can be used to optimize the processes of synthesizing complex interventions. Further work is needed to evaluate the impact of additional study-covariates on explaining residual heterogeneity and testing these methods in other reviews of complex interventions.

ACKNOWLEDGEMENTS

I could not have completed my doctoral training and research if it were not for the support of many individuals to whom I am sincerely grateful. Foremost, I thank my PhD Supervisor, Dr. Jeremy Grimshaw, who, since coming to work for him as a Research Assistant ten years ago, has never stopped supporting my academic, professional, and personal growth. It was Dr. Grimshaw's passion for evidence synthesis and implementation science that led me to pursue my PhD studies and I am grateful for his financial support, methodological guidance, and professional advice that led me through this process. I am grateful to have worked with such a wonderful human being who reminded me of the important things in life (family), and who welcomed my son Elliot to not one, but two, Cochrane conferences!

I am also grateful to have had four wonderful committee members, Drs. David Moher, Issa Dahabreh, Noah Ivers, and Monica Taljaard, who provided exceptional methodological guidance, answered my countless email queries, and helped me meet all my deadlines. I would also like to thank Dr. Thomas Trikalinos, who with Dr. Dahabreh, supported my training of these advanced statistical methods and hosted me twice during my doctoral studies at the Center for Evidence Synthesis in Health at Brown University for further training and collaboration.

In addition to my mentors above, I must thank several other individuals who believed in my academic abilities and encouraged me to pursue a PhD, including Drs. John Last, Dawn Stacey, Deshayne Fell, Heather Colquhoun, and Francina Webster.

I am grateful for the training environment provided by the School of Epidemiology and Public Health at the University of Ottawa, whose individualized study plan approach allowed me deepen my understanding of epidemiology methods while also learn from experts in implementation science and evidence synthesis. I thank my cohort of PhD students, particularly Kylie Tingley, Susan Rogers Van Katwyk, and Brianne Wood whose fierce academic minds were a constant source of inspiration and insight.

To my friends who fed me, took me for runs, watched my kids, and generally just kept cheering ‘Go! Go! Go!’ – thank you! I could not have completed this marathon without you. In particular, I would like to thank Drs. Deshayne Fell and Andrea Patey who provided so much encouragement and support in the final thesis stages.

To my parents, thank you for fostering my interest in learning and modeling such a strong work ethic. In particular, I am grateful for the years of dance training you supported that made me the person and researcher I am today.

Lastly, I am grateful to my husband, Graeme Danko, and children Julianna, Elliot, and Myles Danko, for their patience, confidence, and love. Your belief in me kept me strong, and your selfless support made me want to make you proud. I was a better PhD student because I was your wife and mom.

CONTRIBUTION OF AUTHORS

Manuscript 1: Danko KJ, Dahabreh IJ, Ivers NM, Moher D, Grimshaw J. Contacting non-responding authors of studies included in systematic reviews by telephone increases response rates compared to repeat emailing: results of a randomized controlled trial. To be submitted to the *Journal of Clinical Epidemiology*.

I conceived the idea for this randomized controlled trial with Dr. Grimshaw. I developed the original protocol for the study with feedback Dr. Dahabreh, Dr. Ivers, Dr. Grimshaw. I performed author randomization, delivered interventions, and collected outcome data. I conducted all data analyses, interpreted the results, and drafted the manuscript. Dr. Dahabreh was involved in providing statistical advice, interpreting the results, and editing the manuscript. Dr. Ivers, Dr. Moher, and Dr. Grimshaw contributed to the interpretation of study results and edited the manuscript. All authors critically reviewed the article for intellectual content.

Manuscript 2: Danko KJ, Sullivan KJ, Taljaard M, Ivers NM, Moher D, Grimshaw JM. Descriptive study of the proportion of studies with unit of analysis errors in a systematic review of diabetes quality improvement trials. To be submitted to *Systematic Reviews Journal*.

I conceived the idea for this study, with input from Dr. Taljaard, Dr. Grimshaw, and Dr. Ivers. Dr. Taljaard's feedback was particularly helpful in guiding the methods for data abstraction based on a sample of studies we extracted together. Katrina Sullivan, a Research Associate at the Ottawa Hospital Research Institute, was my second reviewer for the data abstraction. Dr. Taljaard acted as a third reviewer to resolve conflicts or clarify statistical interpretations. I conducted all analyses, interpreted the results, and drafted the manuscript. Ms. Sullivan, Dr. Taljaard, Dr. Ivers, Dr. Moher, and Dr. Grimshaw contributed to the interpretation of the study results and edited the manuscript. All authors critically reviewed the article for intellectual content.

Manuscript 3: Danko KJ, Taljaard M, Ivers NM, Moher D, Grimshaw J. Development of an ICC database and imputation model addressed unit of analysis errors in a systematic review of diabetes quality improvement trials. To be submitted to the *Journal of Clinical Epidemiology*.

I conceived the idea for this study, and developed its protocol, in collaboration with Dr. Dahabreh. I performed all data collection and extraction. I conducted all analyses, interpreted the results, and drafted the manuscript. Dr. Dahabreh provided statistical advice and contributed to the interpretation of results. Dr. Taljaard, Dr. Ivers, and Dr. Grimshaw contributed to the interpretation of the study results and edited the manuscript. All authors critically reviewed the article for intellectual content.

Manuscript 4: Danko KJ, Dahabreh IJ, Trikalinos T, Ivers NM, Moher D, Grimshaw JM. Evidence synthesis for complex interventions using hierarchical regression models. To be submitted to *Statistics in Medicine*.

The original idea for this methodological approach was conceived by Dr. Trikalinos and Dr. Dahabreh as part of the update of the diabetes quality improvement review (Ivers et al, *Syst Rev* 2014;3:88). I developed the detailed protocol for the study in collaboration with Dr. Dahabreh. I led the re-abstraction (and/or confirmation) of data from the original review. Samir Nazarali, a medical student volunteer, was my second reviewer for the data abstraction. I conducted all analyses, interpreted the results, and drafted the manuscript. Throughout, Dr. Dahabreh provided critical feedback on statistical theory and programming. Dr. Trikalinos, Dr. Ivers, Dr. Grimshaw, Dr. Moher, and Dr. Dahabreh, contributed to the interpretation of the study results and edited the manuscript. All authors critically reviewed the article for intellectual content.

STATEMENT OF ORIGINATLITY

I declare that the studies presented in this thesis represent original scientific work carried out during my PhD program in the School of Epidemiology and Public Health at the University of Ottawa. While I received guidance from my PhD supervisor and committee members on methodological aspects of my studies and feedback on the manuscripts, I confirm that I conceived and executed all studies.

STATEMENT OF FINANCIAL SUPPORT

During the first year of my thesis I was financially supported by a stipend from Dr. Grimshaw. The last three years of my thesis were supported by a Frederick Banting and Charles Best Canada Graduate Scholarship Doctoral Research Award from the Canadian Institutes of Health Research (CIHR), supplemented by a stipend from Dr. Grimshaw. Funds for the author incentive described in Manuscript 1 came a CIHR grant supporting the update of a systematic review of diabetes quality improvement interventions (FRN-123345). The same CIHR grant also supported early collaborations with Dr. Dahabreh and Dr. Trikalinos from the Center for Evidence Synthesis in Health (CESH) at Brown University that underscored the methods developed in Manuscript 4. Finally, just prior to my final thesis year I received a CIHR Michael Smith Foreign Study Supplement to study with Dr. Dahabreh at the CESH. A substantial amount of the work that resulted in Manuscripts 3 and 4 was performed during the three-month training intensive supported by Michael Smith award.

TABLE OF CONTENTS

ABSTRACT	II
ACKNOWLEDGEMENTS	V
CONTRIBUTION OF AUTHORS.....	VII
STATEMENT OF ORIGINATLITY	IX
STATEMENT OF FINANCIAL SUPPORT	X
TABLE OF CONTENTS.....	XI
LIST OF TABLES	XVI
LIST OF FIGURES	XVII
LIST OF BOXES	XVIII
LIST OF ACRONYMNS AND ABBREVIATIONS	XIX
CHAPTER 1 INTRODUCTION	1
1.0 INTRODUCTION	2
1.1 BACKGROUND	2
1.2 RESEARCH OBJECTIVES	3
1.3 ORGANIZATION OF DISSERTATION	4
CHAPTER 2 LITERATURE REVIEW	5
2.0 LITERATURE REVIEW	6
2.1 EVIDENCE SYNTHESIS FOR DECISION-MAKING	6
2.2 SYNTHESIS OF COMPLEX INTERVENTIONS.....	7
2.2.1 <i>Characteristics of complex interventions.....</i>	7
2.2.2 <i>Case study: diabetes quality improvement interventions</i>	10
2.2.3 <i>Challenge 1: Obtaining data on intervention components and effect modifiers</i>	11
2.2.3.1 Missing data on components and effect modifiers.....	11
2.2.3.2 Contacting authors to obtain missing data	12
2.2.3.3 Case study: Missing data in diabetes QI review and attempts to contact authors	12
2.2.4 <i>Challenge 2: Abstracting adjusted estimates, or correcting unadjusted estimates, from cluster randomized trials.....</i>	13
2.2.4.1 Unit of analysis errors in cluster randomized trials	13
2.2.4.2 Identifying and correcting unit of analysis errors in CRTs.....	14
2.2.4.3 Case study: Adjusting unit of analysis errors in the diabetes QI review	16
2.2.5 <i>Challenge 3: Synthesizing evidence from complex interventions</i>	17
2.2.5.1 Standard meta-analytic approaches for synthesizing intervention effect estimates	17
2.2.5.2 Analytic approaches for synthesizing complex interventions.....	17
2.2.5.3 Case study: Synthesis of diabetes QI interventions	18

2.2.6	<i>Overview of thesis studies to address methodological challenges.....</i>	19
-------	---	----

CHAPTER 3 [MANUSCRIPT 1] CONTACTING NON-RESPONDING AUTHORS OF STUDIES INCLUDED IN SYSTEMATIC REVIEWS BY TELEPHONE INCREASES RESPONSE RATES COMPARED TO REPEAT EMAILING: RESULTS OF A RANDOMIZED CONTROLLED TRIAL 20

3.0	[MANUSCRIPT 1] CONTACTING NON-RESPONDING AUTHORS OF STUDIES INCLUDED IN SYSTEMATIC REVIEWS BY TELEPHONE INCREASES RESPONSE RATES COMPARED TO REPEAT EMAILING: RESULTS OF A RANDOMIZED CONTROLLED TRIAL	21
3.1	PREFACE TO MANUSCRIPT 1.....	21
3.2	TITLE PAGE.....	23
3.3	ABSTRACT	24
3.4	INTRODUCTION	26
3.5	METHODS.....	27
3.5.1	<i>Design</i>	27
3.5.2	<i>Sample</i>	28
3.5.3	<i>Randomization.....</i>	28
3.5.4	<i>Telephone contact</i>	29
3.5.5	<i>Email contact</i>	29
3.5.6	<i>Delivery of interventions.....</i>	30
3.5.7	<i>Ethics review</i>	30
3.5.8	<i>Outcomes.....</i>	31
3.5.9	<i>Statistical analysis.....</i>	32
3.6	RESULTS	32
3.6.1	<i>Author contact flow diagram.....</i>	32
3.6.2	<i>Telephone contact</i>	32
3.6.3	<i>Email contact</i>	35
3.6.4	<i>Comparison of telephone and email contact.....</i>	37
3.7	DISCUSSION	38
3.8	CONCLUSION	41

CHAPTER 4 [MANUSCRIPT 2] DESCRIPTIVE STUDY OF THE PROPORTION OF STUDIES WITH UNIT OF ANALYSIS ERRORS IN A SYSTEMATIC REVIEW OF DIABETES QUALITY IMPROVEMENT TRIALS 44

4.0	[MANUSCRIPT 2] DESCRIPTIVE STUDY OF THE PROPORTION OF STUDIES WITH UNIT OF ANALYSIS ERRORS IN A SYSTEMATIC REVIEW OF DIABETES QUALITY IMPROVEMENT TRIALS	45
4.1	PREFACE TO MANUSCRIPT 2	45
4.2	TITLE PAGE.....	46
4.3	ABSTRACT	47
4.4	BACKGROUND	49

4.4.1	<i>Case example: Diabetes quality improvement review.....</i>	51
4.5	METHODS.....	52
4.5.1	<i>Did the study report the use of appropriate methods to adjust for clustering?.....</i>	52
4.5.2	<i>Did study report effect estimate required by the review?.....</i>	53
4.5.3	<i>Did the study report an estimate of the ICC?.....</i>	53
4.6	RESULTS	54
4.6.1	<i>Did the study report the use of appropriate methods to adjust for clustering?.....</i>	54
4.6.2	<i>Did study report effect estimate required by the review?.....</i>	55
4.6.3	<i>Did the study report an estimate of the ICC?.....</i>	56
4.7	DISCUSSION	56
4.8	CONCLUSION	59
 CHAPTER 5 [MANUSCRIPT 3] DEVELOPMENT OF AN ICC DATABASE AND IMPUTATION MODEL ADDRESSED UNIT OF ANALYSIS ERRORS IN A SYSTEMATIC REVIEW OF DIABETES QUALITY IMPROVEMENT TRIALS		
5.0	[MANUSCRIPT 3] DEVELOPMENT OF AN ICC DATABASE AND IMPUTATION MODEL ADDRESSED UNIT OF ANALYSIS ERRORS IN A SYSTEMATIC REVIEW OF DIABETES QUALITY IMPROVEMENT TRIALS.....	61
5.1	PREFACE TO MANUSCRIPT 3	62
5.2	TITLE PAGE.....	63
5.3	ABSTRACT	64
5.4	BACKGROUND	65
5.4.1	<i>Case example: Diabetes quality improvement review.....</i>	66
5.5	METHODS.....	68
5.5.1	<i>Defining the target population, cluster and outcomes.....</i>	68
5.5.2	<i>Searching and abstracting relevant ICCs to build an ICC database.....</i>	69
5.5.3	<i>Specifying a model of ICC estimates</i>	69
5.5.4	<i>Calculating the variance of ICC estimates</i>	70
5.5.5	<i>Constructing outcome-specific posterior distribution of the ICC.....</i>	70
5.5.6	<i>Constructing outcome-specific posterior predictive distribution of the ICC</i>	71
5.5.7	<i>Comparison of correcting unit of analysis errors using a single ICC vs. ICCs drawn from the posterior predictive distribution.....</i>	71
5.6	RESULTS	72
5.6.1	<i>Building the ICC database</i>	72
5.6.2	<i>Constructing posterior and posterior predictive distributions.....</i>	75
5.6.3	<i>Comparison of approaches to correct unit of analysis errors.....</i>	75
5.7	DISCUSSION	76
5.8	CONCLUSION	78

CHAPTER 6	[MANUSCRIPT 4] EVIDENCE SYNTHESIS FOR COMPLEX INTERVENTIONS USING HIERARCHICAL REGRESSION MODELS.....	80
6.0	[MANUSCRIPT 4] EVIDENCE SYNTHESIS FOR COMPLEX INTERVENTIONS USING HIERARCHICAL REGRESSION MODELS	81
6.1	PREFACE TO MANUSCRIPT 4.....	81
6.2	TITLE PAGE.....	82
6.3	ABSTRACT	83
6.4	BACKGROUND	84
6.5	MOTIVATING EXAMPLE: QUALITY IMPROVEMENT INTERVENTIONS TO IMPROVE DIABETES CARE.....	85
6.6	CHALLENGES IN META-ANALYSES FOR COMPLEX INTERVENTIONS	87
6.7	SPECIFICATION OF THE BASIC HIERARCHICAL REGRESSION MODEL, PRIOR SPECIFICATION, AND INFERENCE	90
6.7.1	<i>Specification of the basic model</i>	90
6.7.2	<i>Prior specification</i>	91
6.7.3	<i>Inference</i>	92
6.8	MODEL EXTENSIONS	93
6.8.1	<i>Interactions among intervention components</i>	93
6.8.2	<i>Effect modification by study covariates</i>	94
6.8.3	<i>Discrete outcomes</i>	95
6.8.4	<i>Missing data</i>	96
6.9	APPLICATION TO THE MOTIVATING EXAMPLE.....	96
6.9.1	<i>Comparison of hierarchical meta-regression and pairwise synthesis models</i>	97
6.9.2	<i>Ranking of intervention components</i>	99
6.9.3	<i>Assessing interactions among components</i>	100
6.9.4	<i>Assessing effect modification by covariates</i>	101
6.9.5	<i>Predicting the effect of a new study</i>	103
6.9.6	<i>Assessment of convergence, model evaluation, robustness</i>	104
6.10	DISCUSSION	106
6.11	CONCLUSION	108
CHAPTER 7	DISCUSSION.....	110
7.0	DISCUSSION	111
7.1	SUMMARY OF RESEARCH FINDINGS.....	112
7.2	LIMITATIONS	116
7.3	IMPLICATIONS FOR FUTURE RESEARCH	117
7.3.1	<i>Implications for the diabetes QI review</i>	118
7.3.2	<i>General implications for review methodology of complex interventions</i>	119
7.3.3	<i>Implications for planning and reporting trials of complex interventions</i>	120

7.4	FINAL CONCLUSIONS	120
	REFERENCE LIST	121
	APPENDIX A: CERTIFICATE OF ETHICS APPROVAL.....	148
	APPENDIX B: SUPPLEMENTAL INFORMATION FOR MANUSCRIPT 1.....	149
	APPENDIX C: SUPPLEMENTAL INFORMATION FOR MANUSCRIPT 2.....	167
	APPENDIX D: SUPPLEMENTAL INFORMATION FOR MANUSCRIPT 3	177
	APPENDIX E: SUPPLEMENTAL INFORMATION FOR MANUSCRIPT 4.....	183

LIST OF TABLES

TABLE 1 COMPARISON OF COMPLEXITY DOMAINS ACROSS GUIDANCE ON THE METHODS OF SYSTEMATIC REVIEWS OF COMPLEX INTERVENTIONS	9
TABLE 2 AUTHOR CHARACTERISTICS.....	34
TABLE 3 RESULTS OF CONTACT BY TELEPHONE VS. EMAIL ON SURVEY COMPLETION.....	36
TABLE 4 DETAILS ON TIME (MINUTES) TO DELIVER TELEPHONE AND EMAIL INTERVENTIONS	37
TABLE 5 OUTCOMES REPORTED IN CRTs INCLUDED IN THE DIABETES QI SYSTEMATIC REVIEW.....	53
TABLE 6 ANALYSIS AND REPORTING OF CONTINUOUS OUTCOMES IN CLUSTER RANDOMIZED TRIALS IN A DIABETES QUALITY IMPROVEMENT REVIEW	55
TABLE 7 OUTCOMES ASSESSED IN DIABETES QI REVIEW	68
TABLE 8 DATABASE OF ICC ESTIMATES MEETING TARGET CRITERIA OBTAINED FROM A SYSTEMATIC REVIEW OF DIABETES QUALITY IMPROVEMENT TRIALS AND ADDITIONAL SOURCES	73
TABLE 9 ESTIMATED DISTRIBUTIONS OF THE ICC AND PREDICTED ICC.....	75
TABLE 10 IMPACT OF CORRECTION OF UNIT OF ANALYSIS ERRORS WITH SINGLE VS. POSTERIOR PREDICTIVE ICC DISTRIBUTION IN THE META-ANALYSIS OF A CONTINUOUS OUTCOME (% HbA1c).....	76
TABLE 11 TAXONOMY OF QUALITY IMPROVEMENT (QI) COMPONENTS	86
TABLE 12 SUMMARY OF RESULTS COMPARING STANDARD AND HIERARCHICAL REGRESSION MODEL	99
TABLE 13 EFFECT OF QI STRATEGY IN ARMS WITH CONTROLLED VS. UNCONTROLLED HbA1c AT BASELINE	102
TABLE 14 ESTIMATED POSTER AND PREDICTIVE DISTRIBUTIONS OF PREDICTED EFFECTS OF UNTESTED COMBINATIONS	104

LIST OF FIGURES

FIGURE 1 GENERAL PROCESS OF A SYSTEMATIC REVIEW INCLUDING A META-ANALYSIS.....	6
FIGURE 2 PROCESS FOR IDENTIFYING AND CORRECTING UNIT OF ANALYSIS ERRORS OF DATA TO BE SYNTHESIZED IN META-ANALYSIS	15
FIGURE 3 OVERVIEW PROPOSED METHODOLOGICAL IMPROVEMENTS EVALUATED IN THESIS	19
FIGURE 4 OVERVIEW OF EMAIL AND CALL INTERVENTIONS.....	30
FIGURE 5 FLOW DIAGRAM OF THE RANDOMIZED CONTROLLED TRIAL	33
FIGURE 6 CUMULATIVE RESPONSE PROPORTION OF TELEPHONE AND EMAIL CONTACT	37
FIGURE 7 PROCESS FOR IDENTIFYING AND CORRECTING UNIT OF ANALYSIS ERRORS OF DATA TO BE SYNTHESIZED IN META-ANALYSIS	50
FIGURE 8 NUMBER OF CRTs WITH APPROPRIATE ANALYSIS AND REPORTING ACROSS OUTCOMES	54
FIGURE 9 FREQUENCY OF EVALUATIONS OF QI INTERVENTIONS COMPRISED OF COMPONENT QI STRATEGIES	89
FIGURE 10 RANKING OF QI STRATEGY COMPONENTS	100
FIGURE 11 COEFFICIENT ESTIMATES OF PAIRWISE INTERACTIONS BETWEEN QI STRATEGY COMPONENTS..	101
FIGURE 12. EFFECT MODIFICATION BY BASELINE HbA1c.....	103
FIGURE 13 POSTERIOR PREDICTIVE CHECKS	105

LIST OF BOXES

Box 1 TAXONOMY OF QI STRATEGIES	10
Box 2 RECOMMENDATIONS FOR THE CONDUCT AND REPORTING OF FUTURE CRTs AND EVIDENCE SYNTHESIS OF CRTs	59

LIST OF ACRONYMS AND ABBREVIATIONS

ICC	Intraclass correlation coefficient
CI	Confidence interval
CrI	Credible interval
CRT	Cluster randomized trials
QI	Quality improvement
SWAR	Study within a review

CHAPTER 1 Introduction

1.0 Introduction

1.1 Background

Healthcare decision-makers need high quality evidence to inform policy and practice decisions [1,2]. Systematic reviews of randomized controlled trials (RCTs), including meta-analyses of study effects, are considered one of the highest forms of evidence to inform such decisions [3,4]. Most applications of systematic reviews and meta-analyses are based on a standardized cannon of methods that seek to collect, abstract, assess, and synthesize evidence from primary studies to produce a comprehensive and unbiased summary of the evidence [5]. While useful, standard synthesis methods tend to assume simple data structures (e.g., two-arm comparison of a single intervention vs. a similar control evaluated in a parallel individual randomized design) and some practices (e.g., author contact) may not always be supported by empirical evidence [6,7].

Complex interventions are of increasing focus in healthcare and public health and pose challenges to the standard methods of systematic review and meta-analysis [8–10]. While different definitions of complex interventions have been proposed, most definitions assume: i) multiple intervention ‘components’ that may or may not interact with each other to increase or decrease observed intervention effects and ii) effect modification by study-specific characteristics (e.g., healthcare setting, patient population) [11]. At least three challenges may result from this complexity. **First**, reviewers will likely have to contact authors for additional information about intervention components and contextual factors that may operate as effect modifiers [12]. Unfortunately, evidence supporting optimal strategies for achieving response from author contact is lacking [13].

Second, complex interventions are often evaluated using a cluster randomized trial (CRT) design that randomize healthcare organisations (clusters) to different healthcare/health policy interventions and collect individual patient level data [14,15]. Responses of individuals within a cluster tend to be correlated, therefore violating the assumption of independence required for most statistical tests to be valid [16]. The degree of correlation among individual responses needs to be accounted for at both the design

and analysis stage of a CRT [17,18]. Failing to account for clustering during design can lead to underpowered studies, whereas failing to account for clustering during analyses can lead to overly precise estimates of effects [16]. Analyses from CRTs that are not adjusted for the clustering effect are said to have ‘unit of analysis’ [19] errors, which if incorporated in meta-analyses could lead to biased summary estimates and overly precise confidence intervals (CIs) [16]. Methods for reviewers to appropriately appraise, abstract, and synthesize evidence from CRTs are limited [20].

Finally, standard meta-analyses estimate an overall effect of a singular ‘complex intervention’. Such analyses answer the question “*Do complex interventions as a whole lead to a difference in observed outcomes?*” and tend to exhibit high statistical heterogeneity since variation in intervention components and effect modifiers are not accounted for [21]. Hierarchical multivariate meta-regression models performed in a Bayesian framework have been proposed as an alternative synthesis approach for complex interventions to better account for observed heterogeneity and answer the questions of real interest to decision-makers: *What component(s) (or combination of components) work and under what conditions?* [22,23]. However, such models have yet to be applied in the synthesis of complex healthcare interventions.

The overall aim of my doctoral research was to explore the utility of three methodological approaches to address these challenges and optimize the synthesis of complex interventions using a large systematic review of diabetes quality improvement (QI) interventions as a case study.

1.2 Research objectives

The three specific objectives my thesis research were:

- i. To investigate the effectiveness of contacting non-responding authors of studies in our systematic review by telephone call as compared to email to request additional information on intervention components and potential effect modifiers (study population, setting, and context).
- ii. For cluster randomized trials in our review:

- a. To determine the proportion of studies with unit of analysis errors and the reasons for unit of analysis errors (incorrect analysis vs. unclear or incomplete reporting), and;
 - b. To investigate the utility of building a database of intracluster correlation coefficients (ICCs) and imputation model to correct unit of analysis errors.
- iii. To illustrate the use of hierarchical multivariate meta-regression for quantitative synthesis when estimating the effects of complex interventions and exploring effect heterogeneity.

1.3 Organization of dissertation

The following manuscript-based thesis is organized around four manuscripts representing four original studies. A detailed review of the literature pertinent to the dissertation is provided in Chapter 2. Chapters 3 to 6 present the findings of the four thesis studies: an RCT of telephone calling vs. emailing non-responding trial authors for additional information (Chapter 3), a descriptive study of the proportion of studies with unit of analysis errors (Chapter 4), a methodological study examining the use of a posterior predictive distribution of the ICC to correct unit of analysis errors (Chapter 5), and a methodological study examining the use of hierarchical multivariate meta-regression models for synthesizing the effects of complex interventions and exploring effect heterogeneity (Chapter 6). Each manuscript chapter is introduced by a brief preface to describe the specific objective addressed by the study, the contribution of coauthors, and any related appendices and approvals. Finally, the dissertation concludes with Chapter 7, which presents a discussion of the key findings, limitations and conclusions of the thesis.

CHAPTER 2 Literature Review

2.0 Literature Review

2.1 Evidence synthesis for decision-making

The rationale for evidence synthesis for decision-making has long been established [24]: decision-makers (including clinicians, patients, policy-makers) are faced with vast amounts of evidence from healthcare research and the comprehensive synthesis of this evidence, using robust and systematic methods, is needed to efficiently inform evidence-based decisions [24,25]. Early synthesis pioneers developed the foundations of systematic review and meta-analysis to address decision-makers needs [25–27]. While several decades of research have added to, and refined, the canon of evidence synthesis methods, the basic goal and processes of the approach remains the same (Figure 1)[5,28]. That is, to obtain an unbiased summary of available evidence pertaining to a focused question of interest using standardized methods of collection, abstraction, assessment, and synthesis [5,6]. Robust applications of evidence synthesis have led to important changes in healthcare practice [26,29] and are considered the cornerstone to the production of clinical practice guidelines [30] and justification for future studies [31–33].

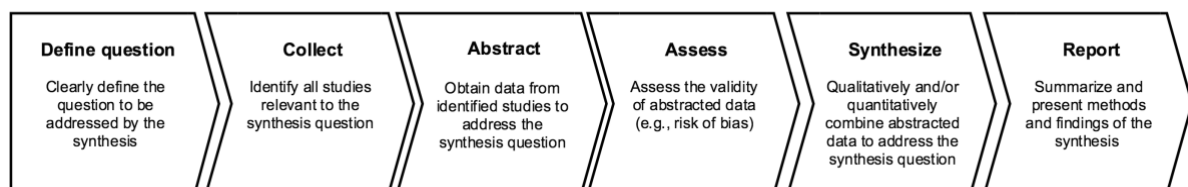


Figure 1 General process of a systematic review including a meta-analysis

Yet despite the benefits of evidence synthesis for decision-making, important methodological challenges remain. First, the rigorous conduct [5,34] and reporting [35,36] of a high-quality evidence synthesis is an increasingly time consuming and costly enterprise [37], particularly when the evidence to be synthesized is large (e.g., broad synthesis question, rapidly expanding evidence base). Second, although the basic tenets of evidence synthesis are well established, many of the recommended methods that underscore the synthesis process lack empirical justification [6]. For example, standard texts of evidence synthesis often recommend contacting authors to obtain additional information not available in published reports [38–40]. However the best approach to

contact authors is unknown; we have sparse evidence on the relative effect of different contact strategies on obtaining author response, the time and costs required to contact authors using different strategies, and the ultimate impact of each of these methods on obtained data and synthesized findings [13,41,42]. Finally, standard synthesis methods are most readily applied to data obtained from pairwise comparisons of a ‘simple’ intervention versus control and face challenges, both in the application of methods and interpretation of results, when applied to more complex data. The remainder of this chapter explores the methodological challenges of evidence synthesis in the context of complex interventions. First however, the concept of complex interventions is introduced.

2.2 Synthesis of complex interventions

2.2.1 Characteristics of complex interventions

Interventions to change aspects of healthcare and public health are inherently complex. For example, implementation science (a scientific field focused on the “study of methods to promote the systematic uptake of evidence-based interventions into practice and policy” [43]) commonly evaluate complex interventions designed to target barriers to change to improve practice, policy and patient outcomes [15,44,45]. Despite the increasing focus on the evaluation of complex interventions, and synthesis of their findings [46], definitions over the specific characteristics of a complex intervention continue to be debated [11,46,47]. Definitions have included features of the interventions, the context and methods of their implementation, and various relationships between these factors [11]. The Medical Research Council (MRC) of United Kingdom proposed one of the first definitions of a complex intervention based on five attributes of complexity, including the:

- Number of interacting components with the experimental and control conditions
- Number and difficulty of behaviors required by those delivering or receiving the intervention
- Number of groups or organizational levels targeted by the intervention
- Number and variability of the outcomes
- Degree of flexibility or tailoring of the intervention permitted [48]

Subsequently, in a series of articles on the systematic review and meta-analysis of complex interventions published in the *Journal of Clinical Epidemiology* (JCE), Anderson et al. proposed a definition of a complex intervention informed by complex adaptive systems [47]. The JCE definition extends the definition of the MRC to consider characteristics of complexity inherent to the system in which the intervention is implemented to be capable of causally influencing observed outcomes. It includes four dimensions of complexity that, may or may not interact, to lead to observed variations in effects. The four dimensions include:

- *Intervention complexity*: variant properties or characteristics of the intervention;
- *Implementation complexity*: variant characteristics of the implementation process;
- *Context complexity*: variant properties or characteristics of the setting or context in which an intervention is implemented;
- *Complexity in participant response*: variant characteristic of participants receiving the intervention [47].

Most recently, in another series of papers published in the JCE, the US Agency for Healthcare Research & Quality (AHRQ) proposed a further definition of complex interventions that built on the Anderson definition but distinguished between aspects of complexity that may be present in some complex interventions but not in others. The AHRQ definition of complex interventions includes:

Primary attributes:

- *Intervention complexity*: multiple components;
- *Pathway complexity*: complex causal pathways due to mediators or moderators of effect and non-linear relationships (e.g., feedback loops, synergies);

Secondary (optional) attributes:

- *Population complexity*: target multiple participants, groups, or organizational levels;
- *Implementation complexity*: require multifaceted adoption, uptake, or integration strategies;
- *Contextual complexity*; work in a dynamic multidimensional environment [46].

Table 1 presents an overview of similarities and differences across the MRC, Anderson, and AHRQ definitions. The table identifies up to five domains of complexity believed to causally impact observed outcomes; two related to the intervention and three related to the context or process in which the intervention is implemented. Looking across definitions, most assume: i) multiple interventions ‘components’ that may or may not interact with each other to increase or decrease observed intervention effects and ii) effect modification by study-specific characteristics (e.g., healthcare setting, patient population). Kuhne et al. performed a literature review of the concepts of ‘complex intervention’ and ‘components’ and came to a similar assessment of the literature. That is, components are “those active, content-related ingredients of an intervention that have the potential to causally influence outcomes” and complex interventions are “interventions consisting of multiple interacting components characterized by variability, adaptively, and interactions with context”[11].

Table 1 Comparison of complexity domains across guidance on the methods of systematic reviews of complex interventions

Guidance		Domains of complexity				
		Internal to intervention		External to intervention		
		Intervention	Non-standardization	Population	Setting/Context	Implementation
MRC (2008)* [48]	Within-domain complexity	- multiple interacting components - number and difficulty of behaviours required by those delivering or receiving intervention	degree of flexibility or tailoring permitted	the number of groups or organizational levels targeted		
Anderson (2013) [47]		<i>intervention complexity</i> (variant factors of intervention)		<i>complexity in participant response</i> (variant characteristics of participants)	<i>context complexity</i> (variant factors of the setting or context)	<i>implementation complexity</i> (variant factors of the implementation process)
AHRQ (2017) [46]		<i>intervention complexity</i> (multiple components)		<i>population complexity</i> (target multiple participants, groups, or organizational levels)	<i>contextual complexity</i> (work in a dynamic multidimensional environment)	<i>implementation complexity</i> (require multifaceted adoption, uptake, or integration strategies)

MRC=Medical Research Council; AHRQ=Agency for Healthcare Research and Quality

*MRC report ‘number and variability of outcomes’ as part of their definition of complex interventions that was not included in this table

2.2.2 Case study: diabetes quality improvement interventions

Diabetes is an increasingly prevalent chronic condition with multiple risk factors and potential complications that warrant attention in clinical practice [49]. Despite a strong-evidence base supporting a range of clinical behaviours for the effective management of diabetes (e.g., appropriate adjustment of medications to ensure glycemic control, regular performance of foot scans to prevent microvascular complications) [50–52], many patients continue to receive suboptimal diabetes care [53,54]. An increasing body of evidence has emerged from trials evaluating complex interventions to address diabetes evidence-to-practice gaps. For example, a 2012 systematic review of 142 trials assessed the effect of diverse quality improvement (QI)¹ interventions (125 of which included multiple components) across a range of process (e.g., foot screening) and intermediate patient (e.g., glycemic control) outcomes [55]. QI interventions were coded using a taxonomy of 12 QI strategies (i.e. components) adapted from Cochrane’s Effective Practice and Organization of Care taxonomy (defined in Box 1) [55–57]. Standard random effects meta-analyses of observed outcomes found improvements across most outcomes, although results were heterogeneous [55]. Further, reviewers faced challenges in coding intervention components and potential effect modifiers to better explore heterogeneity, and abstracting adjusted data from CRTs included in the sample. A recent update of the review (total sample n=272 RCTs)[22] offered a unique opportunity to assess the utility of three methodological approaches to address identified challenges in the synthesis of complex interventions in a large fixed review sample.

Box 1 Taxonomy of QI strategies

QI strategies targeting health systems	<i>Clinician education</i>
<i>Case management</i>	Interventions designed to promote increased understanding of principles guiding clinical care or awareness of specific recommendations for a target disorder or population of patients
Any system for coordinating diagnosis, treatment, or routine management of patients by a person or multidisciplinary team in collaboration with, or supplementary to, the primary care clinician	Clinician reminders Paper-based or electronic systems intended to

¹ Terminology of complex healthcare interventions and their component parts remains a challenge. Based on previous literature denoting the shared scientific goals and lack of reliable or meaningful distinction between various terminologies aimed at implementation at this time [44,240], for my thesis I consider the term quality improvement to be synonymous with implementation, knowledge translation and other related terms.

<i>Team changes</i> Changes to the structure or organization of the primary health care team including adding a team member or shared care, use of multidisciplinary teams, or expansion or revision of professional roles <i>Electronic patient registry</i> General electronic medical record system or electronic tracking system for patients with condition <i>Facilitated relay of clinical information</i> Clinical information collected from patients and transmitted to clinicians by means other than the existing medical record <i>Continuous QI</i> Interventions explicitly identified as involving the techniques of continuous QI, total quality management, or plan-do-study-act, or any iterative process for assessing quality problems, developing solutions to those problems, testing their effects, and then reassessing the need for further action QI strategies targeting health providers <i>Audit and feedback</i> Summary of clinical performance of health care delivered by an individual clinician or clinic over a specified period, transmitted back to the clinician	prompt a health professional to recall patient-specific information <i>Financial incentives</i> Interventions with positive or negative financial incentives directed at providers, patients or system-wide changes in reimbursement QI strategies targeting patients* Patient education Interventions designed to promote greater understanding of a target disorder or to teach specific prevention or treatment strategies, or specific in-person education <i>Promotion of self-management</i> Provision of equipment or access to resources to promote self-management <i>Patient reminders</i> Any effort to remind patients about upcoming appointments or important aspects of self-care *To be included in the review, any intervention including a QI strategy that targeted patients had to also include a QI strategy that targeted the health system and/or a health provider
---	--

2.2.3 Challenge 1: Obtaining data on intervention components and effect modifiers

2.2.3.1 Missing data on components and effect modifiers

Published reports of the effects of complex interventions often fail to report enough information for reviewers to code the specific intervention components believed to influence observed effects [58–60]. Studies that do report intervention details, often define components in varying and inconsistent ways due to the lack of standardized terminology [59,61]. For example, Lokker et al. identified over 50 classification schemes used to characterize implementation interventions [62]. Despite diversity of terms however, a review of QI interventions noted many similarities in the content of interventions labeled under different names [63]. The lack of standardized terminology combined with incomplete reporting may lead reviewers to misclassify intervention components, creating two problems for subsequent meta-analyses: 1) increased noise

due to the inclusion of studies believed to have evaluated a component of interest that did not; and 2) reduced precision due to the exclusion of studies believed to have that not evaluated a component of interest, but did. In addition to details on intervention components, published reports often fail to report enough information on the context in which the intervention was implemented, which if coded and analyzed may help explain some residual heterogeneity [8,64]. As most definitions of a complex intervention assume the effects of an intervention to be modified by contextual details (e.g., healthcare setting, population characteristics), details on these variables are crucial. Reviewers of complex interventions will likely need to obtain additional information to supplement the published reports of trials to better understand the impact of these variables on variation of effects [12].

2.2.3.2 Contacting authors to obtain missing data

To obtain missing information, standard methods of systematic review often recommend contacting primary study authors [5,38,40]. However, evidence supporting the effectiveness of specific author contact strategies and their respective impact on response rates and review costs, data, and conclusions, remain lacking [13,38,41,42]. For example, one Cochrane review of six studies with high risk of bias observed improved response rates in some strategies over others (e.g., email versus letter), however studies were too heterogeneous to be combined, were mostly observational, and aside from one study, were published in abstract form only [13]. One descriptive study that sought to describe the practice of author contact in published systematic reviews found that while the majority of systematic reviewers contact authors of primary studies for additional information (85% Cochrane reviews, 50% non-Cochrane reviews), most do not report the methods with which they do so, or their associated response rates [42].

2.2.3.3 Case study: Missing data in diabetes QI review and attempts to contact authors

The 2012 diabetes review observed poor reporting of key features of the complex intervention and its context [55,65]. As a result, the update of the review sought to obtain additional data on study's intervention components, population, and setting through a tailored online survey [22].

Of the 279 authors contacted (representing 279 studies²), only 274 could be emailed. After three email attempts, only 76 authors (27%) completed the survey. During the conduct of the survey, we observed that speaking to one author by phone to assist with the web survey led a positive researcher-researcher interaction and subsequent survey completion. The study described in Manuscript 1 tested whether contacting non-responding authors by telephone, compared to continuing email reminders, would improve rates of survey completion using a RCT design.

2.2.4 Challenge 2: Abstracting adjusted estimates, or correcting unadjusted estimates, from cluster randomized trials

2.2.4.1 Unit of analysis errors in cluster randomized trials

Complex interventions are often evaluated using cluster randomized trials (CRTs), in which socially intact groups of individuals ('clusters') are randomized to experimental arms but intervention effects are measured at the level of the individual members of the cluster [15]. For example, clinical units may be randomized to an intervention involving audit and feedback and clinician education (vs. 'usual care') but assessed using a patient outcome (e.g., proportion of patients achieving standard of care on a clinical marker). Such designs present statistical and reporting complications for evidence synthesis. Statistically, standard methods of experimental design assume independence of observed outcomes [66]. The clustered design results in outcomes of individuals within a cluster being correlated, as the responses of individuals within one cluster tend to be more similar than the responses of individuals in a different cluster [16]. Analyses that do not account for the correlated (clustered) nature of CRT data are said to have 'unit of analysis' errors [19] characterized by overly precise estimates of intervention effects and increased likelihood of spurious findings [17,67]. Meta-analysis of study estimates with uncorrected unit of analysis errors can lead to biased summary estimates and overly precise confidence intervals (CI's). From a reporting perspective, published reports of CRTs must present enough information for reviewers to adequately assess the methods of analysis and extract adjusted effect measures (or extract data to correct unadjusted estimates). Evidence from reviews of published primary studies however continues to

² The original sample size of the updated review was 279 and used to contact authors. During data abstraction, seven studies were identified as not meeting eligibility criteria and were subsequently excluded, to leave the final sample size of the review (n=272).

find persistent errors in trialists' adequate conduct or reporting of CRTs [67–70]. For example, in a review of 300 randomly sampled CRTs published 2000-2008, only 70% of trials reported taking clustering into account during analysis and only 16% reported an estimate of the ICC that could be used to correct unadjusted effect estimates [68].

2.2.4.2 Identifying and correcting unit of analysis errors in CRTs

To identify and correct unit of analysis errors, reviewers must determine whether appropriate methods were used to adjust for the clustering design (e.g., multilevel models, random effects, use of generalized estimating equations) and where this is the case, abstract adjusted data [64]. Where it is not, and unit of analysis errors are identified, reviewers must seek to correct unadjusted estimates. Some trials will conduct correct analyses but report results in such a way that adjusted effects cannot be abstracted; in such cases, reviewers will need to correct unadjusted estimates as well. While different methods for correcting unadjusted cluster estimates in a meta-analysis have been proposed [71], we adopt the methods recommend by the Cochrane Handbook [64] as they can be easily applied to continuous and dichotomous data and do not require additional assumptions or information about CRT designs that may also be poorly reported [68]. A summary of the approach proposed by Cochrane is presented in Figure 2.

Where unit of analysis errors are identified, the goal is to approximately account for the 'design effect' [72] of the CRT by reducing the sample size of the study to its 'effective sample size' [73] or inflate the observed variance to account for the between-cluster heterogeneity. The design effect can be calculated using an estimate of the average cluster size (m) and an estimate of the intracluster (intraclass) correlation coefficient (ICC; typically denoted by the Greek letter ρ) as defined by the following: $(1 + [m - 1] * \rho)$ [16]. Once study-level data are corrected, meta-analysis may proceed as usual.

However, there are several challenges with implementing the above approach. First, trials do not always report enough information to determine whether appropriate analyses were conducted and/or reviewers may lack sufficient knowledge to determine their appropriateness [64,68]. The Cochrane Handbook for example, recommends

reviewers seek statistical advice in determining the appropriateness of cluster analysis methods [64]. Thus, whether due to poor reporting of primary studies or reviewer knowledge, reviewers may fail to appropriately identify unit of analysis errors in CRTs.

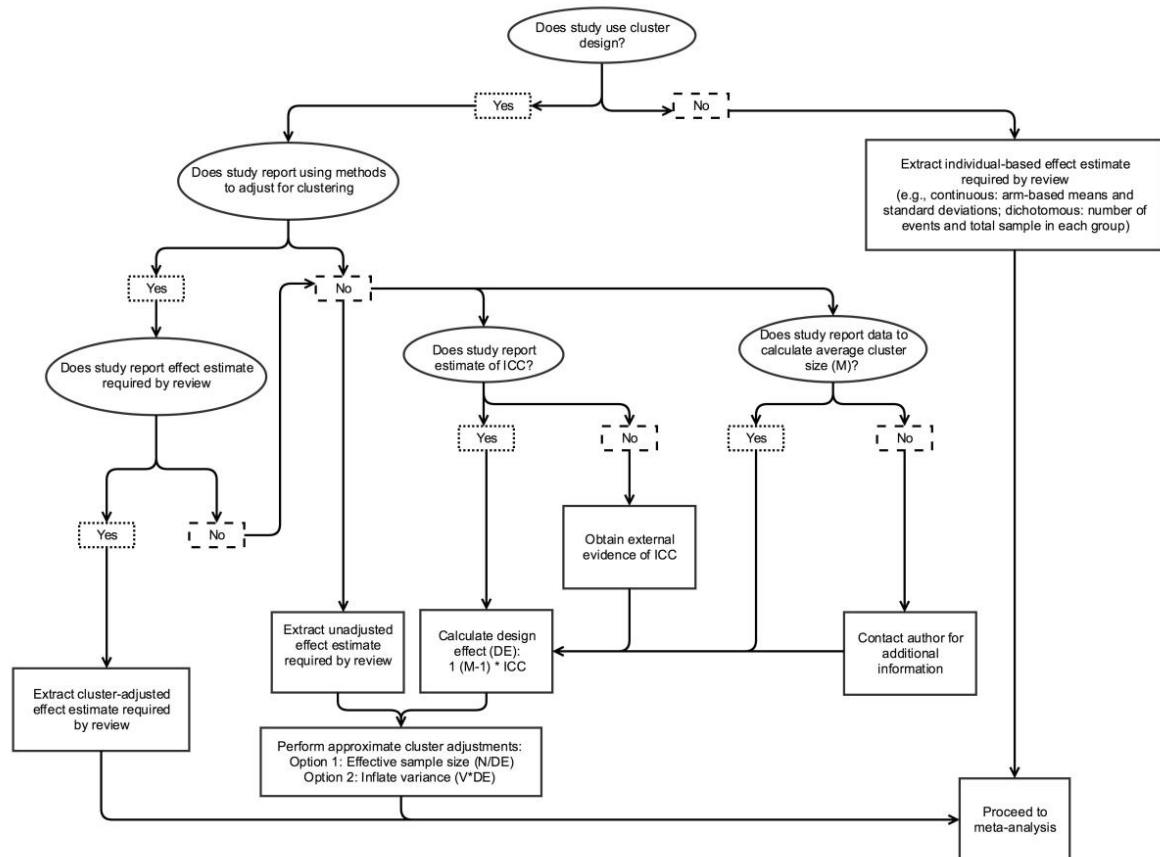


Figure 2 Process for identifying and correcting unit of analysis errors of data to be synthesized in meta-analysis

Second, CRTs may conduct and report appropriate analyses, but report adjusted effect estimates that cannot be readily abstracted and used in meta-analytic models. For example, standard methods for synthesizing continuous effect estimates require a measure of the study mean and its standard error [74]. Although conventional approaches exist to convert reported measures to a standard error for standard pairwise individual RCT data (e.g., 95% confidence interval, p-value for the comparison of two group means), the extent to which these calculations can or should be applied to summary statistics from CRT data is unclear. Third, when unit of analysis errors have been identified and an estimate of the ICC is not reported by the study, common practice is to correct unadjusted data with an external estimate of the ICC [75,76]. Often a single

estimate of the ICC will be used, although sensitivity analyses of the robustness of review findings to different ICCs (where available) is recommended [64]. Unfortunately, the use of any single ICC fails to account for the observed variation in ICC estimates in synthesized estimates [77,78]. Turner et al. proposed a formal approach to incorporate the uncertainty of the ICC into sample size calculations by building a prior distribution from all relevant ICCs in a Bayesian framework [79]. While Turner et al. applied their model to accurately account for ICC uncertainty in the design of CRTs, we believe the approach can be extended to account for ICC uncertainty in correction of unadjusted analyses during synthesis as well.

A review of 50 Cochrane reviews published June to November 2013 demonstrates some of the challenges reviewers' face in assessing, abstracting, and synthesizing data from CRTs. For example, 74% of reviewers failed to state whether effect estimates from CRTs were adjusted, and where RCTs conducted appropriate analyses, only 68% of reviews correctly extracted adjusted data [20]. Of studies that used an external estimate of the ICC to correct unadjusted data, only 50% conducted sensitivity analyses on the magnitude of the ICC. The review authors therefore recommended the need for "consolidated and comprehensive set of guidelines...[to] address the different aspects of including C-RCTs in systematic reviews" [20].

2.2.4.3 Case study: Adjusting unit of analysis errors in the diabetes QI review

The 2012 diabetes review included evidence from 48 CRTs including 84,865 patients. Data on mean difference in glycated hemoglobin (HbA1c%) from 33 CRTs were meta-analyzed. To correct for unit of analysis errors identified, the review used an external estimate of the ICC equal to 0.07, which reduced the total sample size of contributing CRTs from 34,148 to 12,052 [55]. The choice of 0.07 was arbitrary and purposively conservative; other ICCs could have been chosen that would have led to different effective sample sizes (e.g., one study included in the review provided an estimate of 0.027 [80], which if used would have led to an effective sample size of 14,786). In addition, in preparing the update of the review, there was concern that unit of analysis errors may have been misdiagnosed due to poor assessment and abstraction of study

data. Manuscript 2 presents a study that aimed to accurately determine the extent of unit of analysis errors in the diabetes QI literature (due to inappropriate analysis or reporting) and optimize the abstraction of adjusted estimates (including the appropriate calculation of adjusted standard errors from other reported measures) and their ICCs. Manuscript 3 presents a study that assessed the imputation of missing ICCs using the approach recommended by Turner et al. to correct unit of analysis errors identified in Manuscript 2.

2.2.5 Challenge 3: Synthesizing evidence from complex interventions

2.2.5.1 *Standard meta-analytic approaches for synthesizing intervention effect estimates*

Most reviews, even for well-defined narrow questions of ‘simple’ interventions, will have studies that vary in terms study methods, population, setting, etc. that lead to increased variability (i.e., heterogeneity) in observed effects [81]. The conventional method for handling between-study variation is to assume heterogeneity is random and to fit a random effects model [74]. Under such a model, observed study effects are assumed to be realizations of underlying true study effects, which in turn, are assumed to share a common distribution [74]. The goal of synthesis is to estimate the mean $\hat{\mu}$ and variance $\hat{\tau}^2$ (i.e., heterogeneity) of the underlying random effects distribution [81]. In complex interventions, interventions are by definition variable (i.e., have multiple components with potential synergistic/antagonistic effects) and interact with variations in the setting, population and context [11]. Syntheses of complex interventions under a random effects model therefore tend to exhibit high heterogeneity and have limited capacity for exploring the causes of heterogeneity. Further, the average mean estimate, $\hat{\mu}$, does not completely capture the inference of interest to decision-makers who would rather know the average mean effect of specific components of complex interventions (and their combinations) and how they vary when implemented across different populations, settings and healthcare contexts [82].

2.2.5.2 *Analytic approaches for synthesizing complex interventions*

There have been an increasing number of published papers on meta-analytic approaches for complex interventions [8,9,83,84]. Various approaches for synthesizing complex

intervention evidence and exploring their inherent heterogeneity have been proposed, including subgroup analyses [85], meta-regression [8,9], and component network meta-analysis [21,86], among others [8]. The overwhelming limitation of these methods is data sparsity – particularly, in situations where the number of components (and their possible combinations) is extremely large. For example, subgroup analysis and meta-regression can allow for the exploration of hypothesized predictors of heterogeneous effects. However, most reviews of complex interventions will not have enough studies to support such analyses beyond one or two factors of interest [8,87]. Interpretation is therefore hampered by potential confounding by other factors not controlled for in the regression model (or stratified by in subgroup analyses) and the potential correlation of factors [87]. Even if network meta-analyses include studies with suitably comparable populations and settings (and thus, do not need to explore these factors as sources of heterogeneity) there is still likely to be a limit to the number of components a network meta-analysis can compare. For example, a fully saturated network of 4, 6 and 10 components (i.e., including a node for each component and combination of components) would include 16, 64, and 1024 nodes, respectively. Thus as the number of components of interest grows, the more sparse the evidence to support the network is likely to become, to the point where network meta-analysis may not be feasible (as in the diabetes example).

An ideal model for synthesizing the effect of a complex intervention would incorporate data from multiple arms, allow for the assessment of individual and combined intervention components, and allow for the exploration of important setting, population, and contextual effect modifiers on intervention effectiveness. It would also offer opportunities to predict effective combinations of intervention components that have not been studied. Bayesian hierarchical multivariate regression models that specify intervention effects at the level of the arm appear to meet these criteria [23,88,89] but their utility for syntheses of complex healthcare interventions has not been explored [22].

2.2.5.3 Case study: Synthesis of diabetes QI interventions

The 2012 diabetes review performed random effects meta-analyses of the QI interventions (involving one or more components) versus usual care. The overall

summary effects of each outcome (e.g., mean difference in HbA1c) expressed the effectiveness of the QI interventions as a whole and found QI interventions to be generally effective, but highly heterogeneous. To better understand the impact of specific components, subgroup meta-analyses were conducted to express the efficacy of QI interventions containing a specific component of interest (e.g., case management) compared to QI interventions not containing the component of interest (e.g., no case management), regardless of the presence of other components. Although subgroup analyses found improvements associated with most QI components, summarized effects were likely confounded by the presence of co-occurring components that were not controlled for in the univariate analyses. Finally, the meta-analyses could not assess the impact of potential interactions among components or influence of effect modifiers on components or their potential interactions. Thus, despite the large body of evidence, the review was unable to explain observed heterogeneity and effectively inform decision-makers of the likely effect they might observe if a component (or combination of components) were to be implemented in their population, setting, and context.

2.2.6 Overview of thesis studies to address methodological challenges

The aim of my thesis research was to evaluate methods to address the three challenges above to optimize methods for collecting, abstracting, and synthesizing evidence from complex interventions. An overview of the objectives of the thesis described in Chapter 1, as they relate to the overall of process of evidence synthesis, is presented in Figure 3.

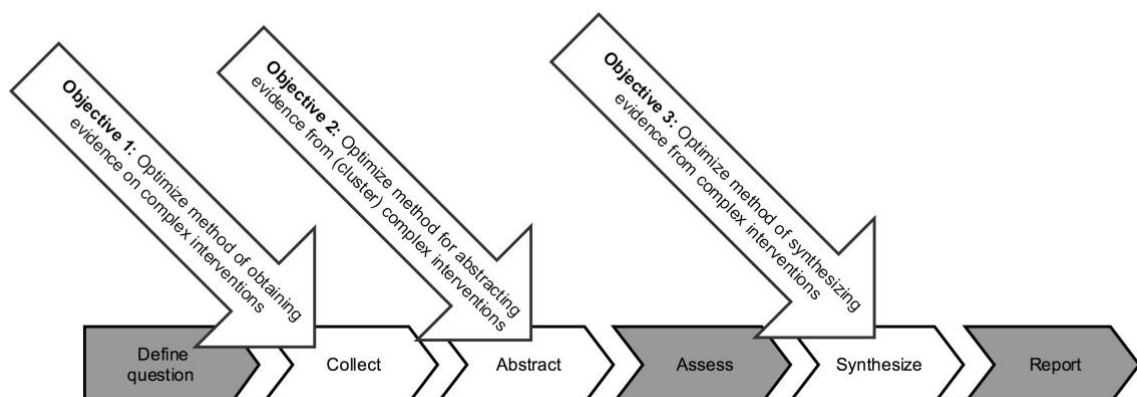


Figure 3 Overview proposed methodological improvements evaluated in thesis

**CHAPTER 3 [Manuscript 1] Contacting non-responding
authors of studies included in systematic reviews by
telephone increases response rates compared to repeat
emailing: results of a randomized controlled trial**

3.0 [Manuscript 1] Contacting Non-Responding Authors Of Studies Included In Systematic Reviews By Telephone Increases Response Rates Compared To Repeat Emailing: Results Of A Randomized Controlled Trial

3.1 Preface to manuscript 1

Study 1 presents a randomized controlled trial that compared the effect of telephone calling versus repeat emailing non-responding authors in a systematic review on response rates and study costs. I conceived the idea for this randomized controlled trial with Dr. Grimshaw after our initial attempts to contact authors by email for additional information resulted in a response rate of only 27%. In conducting the initial email attempts, I observed one instance where speaking to the author by telephone (to address survey troubleshooting issues) led to a positive interaction and subsequent survey completion. We hypothesized that attempting to contact non-responding authors by telephone would improve rates of survey completion compared to continued emailing and we decided to evaluate this hypothesis in a randomized trial.

I developed the original protocol for the study with feedback Dr. Dahabreh, Dr. Ivers, Dr. Grimshaw. I performed author randomization, delivered interventions, and collected outcome data. I conducted all data analyses, interpreted the results, and drafted the manuscript. Dr. Dahabreh was involved in providing statistical advice, interpreting the results, and editing the manuscript. Dr. Ivers, Dr. Moher, and Dr. Grimshaw contributed to the interpretation of study results and edited the manuscript. In the acknowledgements section of the manuscript, I acknowledge Mr. Anton Saarmaki, who provided technical support with developing and managing the online survey platform and Mr. Jordi Pardo Pardo, who provided invaluable feedback on early drafts of the manuscript. I also acknowledge Mr. Samir Nasrali, Ms. Pauline Barbeau, Mr. Mostafa Alabousi who assisted with data abstraction of study characteristics and outcome data for the diabetes quality improvement review, some of which was used for the purpose of this study.

The study was approved by the Ottawa Health Science Network Research Ethics Board (Protocol ID: 20180429-01H) and is registered on the SWAR (Study Within A Review) repository; Registration number SWAR11

(<https://www.qub.ac.uk/sites/TheNorthernIrelandNetworkforTrialsMethodologyResearch/SWATSWARInformation/Repositories/SWARStore/>).

3.2 Title page

Title: Contacting non-responding authors of studies included in systematic reviews by telephone increases response rates compared to repeat emailing: results of a randomized controlled trial

Authors: Danko KJ^{1,2}, Dahabreh IJ^{3,4}, Ivers NM^{5,6,7}, Moher D^{1,2,8}, Grimshaw JM^{1,2,8}

Affiliations:

¹Clinical Epidemiology Program, Ottawa Hospital Research Institute, Ottawa, Canada

²School of Epidemiology and Public Health, University of Ottawa, Ottawa, Canada

³Center for Evidence Synthesis in Health, Brown University, Providence, United States

⁴Departments of Health Services, Policy & Practice and Epidemiology, Brown University, Providence, United States

⁵Family Practice Health Centre, Women's College Research Institute, Toronto, Canada

⁶Institute for Health Systems Solutions and Virtual Care, Women's College Hospital, Toronto, Canada

⁷Department of Family and Community Medicine, Dalla Lana School of Public Health, Institute of Health Policy Management and Evaluation, University of Toronto, Toronto, Canada

⁸Department of Medicine, University of Ottawa, Ottawa, Canada

Corresponding author:

Kristin Julianna Danko

The Ottawa Hospital – General Campus, Box 711

501 Smyth Road

Ottawa, Ontario, Canada

K1H 8L6

Email: kdanko@ohri.ca

3.3 Abstract

Objective: To compare the effects on response rates and research costs of calling vs. continuing to email non-responding authors of studies included in a systematic review.

Study Design and Setting: In a systematic review of diabetes quality improvement interventions, we found that key features of the intervention, population, and setting were poorly reported. In the process of updating the review, we developed a survey to request additional information on the intervention, population, and setting from contact authors of all included studies. After three email contact attempts, only 76 of the 279 authors (27%) had completed the survey. In this study, we randomly assigned non-responding authors to contact by telephone vs. continued emailing to compare the effect of these strategies on response rates and research costs.

Results: Eighty-seven authors were randomized to telephone contact and 89 authors to email contact. Contacting authors by telephone improved the odds of authors completing the survey (36.7% vs. 20.2%; difference of 16.5%; adjusted odds ratio 2.26 [95% confidence interval 1.10-4.76]), but required more time to deliver (20 vs. 10 hours in total; 14 minutes vs. 7 minutes per randomized author; 26 weeks vs. 4 weeks), and was more expensive (total cost \$504 [\$463-\$546 based on salary range] vs. \$252 [\$232-\$273] based on average hourly rate of a research assistant at our institution in Canadian dollars). Because of the higher response rate in the telephone group, the cost per completed survey was approximately the same in both groups (\$16 vs. \$14).

Conclusion: Contacting non-responding authors of included studies for additional information by telephone increases response compared to repeat emailing but is more time consuming and expensive.

Keywords: contacting authors, randomized controlled trial, systematic review methodology, effectiveness, telephone contact, email contact, Study Within A Review

What's new?

- Systematic reviewers may consider telephone calling instead of repeat emailing authors of studies included in systematic reviews who have not responded to initial email requests for information.
- In a randomized controlled trial we compared the effect of telephone calling versus repeat emailing non-responding authors on response rates and study costs.
- We found that contacting non-responding authors for additional information by telephone improves response rates compared to repeat emailing, but requires more investigator time and is more expensive.
- Future studies should compare calling authors upfront vs. the staged approach evaluated in this study.

3.4 Introduction

Publications of primary research studies often do not report enough information about the study methods and results to allow systematic reviewers to assess and summarize the evidence generated by the studies [42,90]. When that is the case, reviewers often attempt to contact the study authors to obtain additional information, for example, about study characteristics [40], research methods [38], or outcomes [5,38,40]. The issue of incomplete reporting, and the need to address it, may be even more acute in reviews of complex interventions where consensus on intervention terminology is lacking [62], descriptions of interventions are typically incomplete [58,59], and variation in components, as well as their interactions with study-specific characteristics, are believed to influence variation in study effects [11,47].

Existing guidance on the conduct of systematic reviews recommends contacting authors to obtain additional information [5,38,40], but the evidence supporting the effectiveness and value of contacting authors is limited [38,41,42]. Five controlled trials identified in a Cochrane review of methods for obtaining unpublished data suggest that authors are more likely to respond to requests for additional information if they pertain to a clarification of methods (vs. missing data about study results), are received by email (vs. letter or fax), and refer to a more recently published study [13]. The quality of the trials identified by the Cochrane review was poor – only two studies randomly assigned authors to different methods of contact and only one of the remaining three non-randomized trials had been published in full text [13]. Furthermore, only two studies considered the time and monetary cost associated with conducting author contact. One non-randomized trial assessed time for author response in days (as compared to investigator time) [91] and another non-randomized trial assessed the cost of intervention delivery, but only for material (i.e., postage and paper)[92].

In a recent update of review of diabetes quality improvement (QI) interventions [55,93], we found that included primary studies often provided incomplete information about intervention components, population, and settings [55,65]. In preparing to update the review again, the review team (including the authors and knowledge users listed in Supplemental Table B-1) developed a tailored survey to request additional information on the intervention, population, and setting from authors of all included studies (n=279) [22]. We sent the web-based survey repeatedly via email to corresponding authors until they responded, requested not to be contacted further, or we completed a maximum of three contact attempts. The emails to the contact authors were sent from the email address of a senior investigator on our research team (JMG). We also provided a financial incentive: authors who completed the survey were entered in a draw for one of five \$100 (CAD) gift certificates. A total of 76 authors (27.2%) completed the survey after three email attempts. We observed that in a single occasion, speaking to a primary study author by telephone (to address survey troubleshooting issues) led to a positive interaction and subsequent survey completion. Based on this observation, we hypothesized that attempting to contact non-responding authors by telephone would improve rates of survey completion compared to continued emailing (i.e., for a fourth time). Given the paucity of evidence about the preferred methods for author contact, we decided to evaluate this hypothesis in a randomized trial.

3.5 Methods

3.5.1 Design

We used a parallel group randomized controlled trial design with study author as the unit of randomization, intervention delivery, and analysis. We followed a pre-specified protocol and registered the study on the Study Within A Review repository [94] (registration number SWAR11). The study is reported according to the recommended reporting items listed in the CONSORT 2010 statement (Supplemental Table B-2) [95].

3.5.2 Sample

We included authors in our trial if they had published a study included in the diabetes QI review, had not completed the survey, and had not asked to be removed from further contact. We linked multiple surveys to the same author (i.e., multiple publications of different studies by the same author) to ensure that we randomized unique authors. Of the 279 authors initially contacted, 76 completed the survey and five asked to be removed from further contact for varying reasons (e.g., retirement, no time, old study). We removed one author because their study did not meet the eligibility criteria of the QI review and we linked 39 surveys to 17 unique authors, leaving a total of 175 unique authors (representing 197 surveys) to be randomized. Post-randomization, but prior to the delivery of the interventions, one author was identified as having not completed the survey, resulting in a final total of 176 authors to be randomized.

3.5.3 Randomization

We randomly assigned authors in 1:1 ratio to the telephone or email intervention group using stratified randomization. Because we expected that year of publication would be an important predictor of author response [13,96], we stratified the randomization by last year of study conduct (or year of study publication, if not reported) using the following strata: 1980-1989 (n=2); 1990-1999 (n=23); 2000-2009 (n=97); 2010-2014 (n=53). For authors with multiple studies, the study with the most recent publication year was used to define the authors' stratum. We generated the random allocation sequence in Stata [97] (Supplemental Text B-1). The investigator responsible for randomization (KJD) masked all author and study details prior to randomization and preserved the masking until allocation had occurred and we were ready to begin the delivery of the interventions. Post-randomization, we excluded two authors that had completed the survey and were randomized in error (one in each group).

3.5.4 Telephone contact

Contacting authors by telephone involved three key phases: searching for authors' phone numbers, calling authors up to three times by telephone to request that they complete the online survey, and following up with consenting authors via email with the survey link and PDF of the authors' study (additional details reported in Supplemental Text B-2). The objective of contacting authors by telephone was to speak directly to non-respondents to promote their completion of the survey and offer assistance with the survey platform, if needed.

3.5.5 Email contact

We sent authors in the email group up to three additional email requests to complete the survey. Emails included the survey link and PDF of the authors' study; authors with multiple studies were sent the unique survey links and PDFs for all their respective studies. The email noted the deadline to complete the survey (three weeks from the email sent date). An overview of the interventions for the telephone and email groups is presented in Figure 4. We compare the interventions in Supplemental Table B-3.

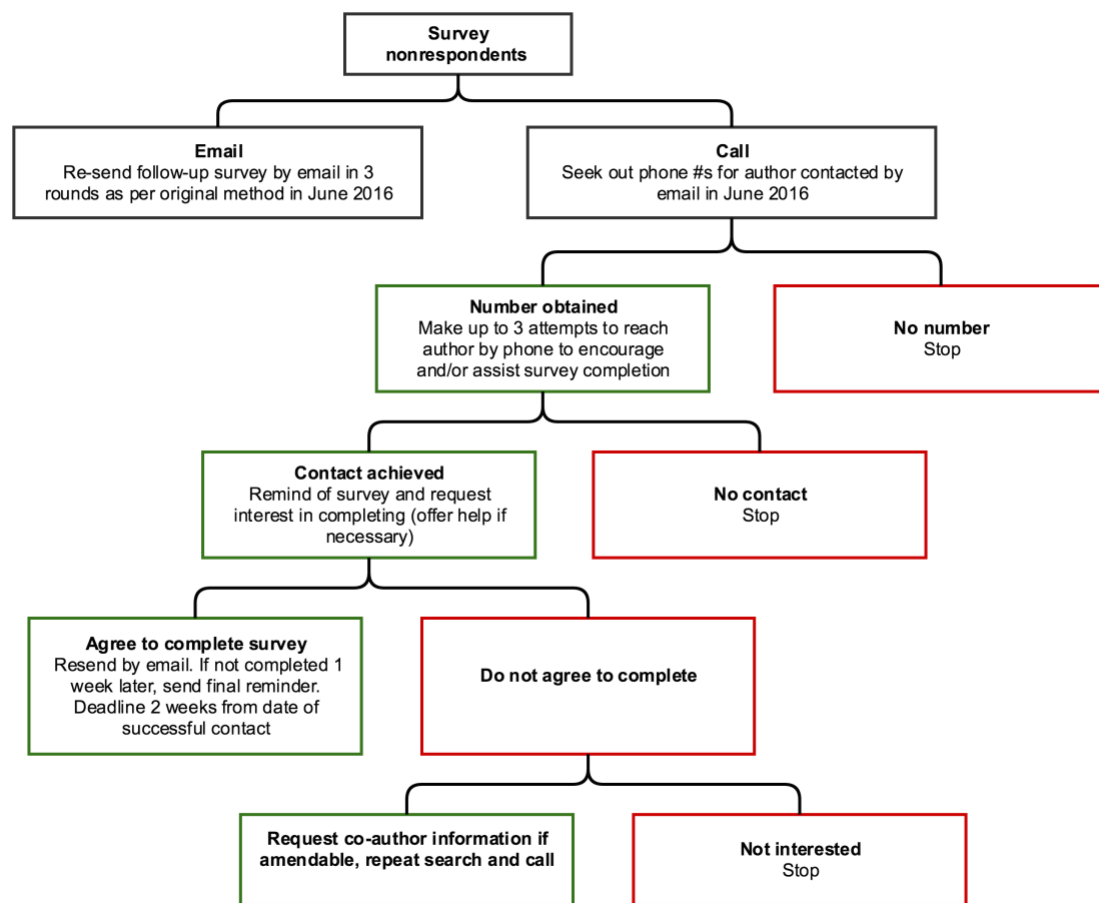


Figure 4 Overview of email and call interventions

3.5.6 Delivery of interventions

One unblinded investigator (KJD) delivered the interventions, managed data, and performed the analysis. Primary study authors were blinded to the intervention, as they were unaware of the trial comparing different methods of author contact. Most author contact was attempted from the Ottawa Hospital Research Institute (OHRI) in Ottawa, Canada; one author was called offsite outside of research hours to accommodate the 13-hour time difference to Tokyo, Japan.

3.5.7 Ethics review

We conceptualized our trial as a *Study Within A Review (SWAR)*, that is, a study undertaken alongside a systematic review to generate evidence on the methods of

future systematic reviews [7]. Waiver of consent is common in these types of studies due to minimal risk on participants and potential for consent to substantially bias findings. We therefore sought and obtained ethics approval from the Ottawa Health Science Network Research Ethics Board to waive consent and perform a debrief with trial participants following trial completion (Protocol ID: 20180429-01H).

3.5.8 Outcomes

The primary outcome of the study was the response proportion, defined as the number of authors who completed the survey over the total number of authors assigned to the intervention. As the primary outcome was binary and clustering of studies for each author was accounted for at the design phase of the trial, an author could only be coded as completing or not completing, regardless of the number of surveys submitted. The secondary outcomes of the study were the time to deliver the interventions and associated costs. We measured time using an online timer and rounded to the nearest minute. If time was measured as less than one minute, we rounded up; this occurred in searching for some author phone numbers that were found immediately or when we telephoned an author and the call rang but was not answered. We calculated the total intervention cost by multiplying observed time to deliver each intervention by the salary range for a research assistant position that may typically contact authors during the review process. The salary range (CAD \$23.14-\$27.26/hour) was based on 2017 values at the OHRI. We calculated the cost per completed survey as the total intervention cost divided by the number of completed surveys in that intervention group. Post hoc, we calculated the time taken by authors to complete the survey by summing the individual authors access times (obtained using the time stamps for accessing the survey pages). Of note, this approach may have overestimated the time authors spent completing the survey because it assumed authors were completing the survey the entire time the survey webpages were open on their devices.

3.5.9 Statistical analysis

We first estimated the effect of telephone and email contact on responses rates in each arm. We then compared the odds of completing the survey in the telephone group using exact logistic regression [98] that accounted for stratified randomization by decade [99,100]. The sample size of this trial was not under our control (because the number of non-responding authors among all authors of studies included in the diabetes QI review was not under our control), nevertheless, *before* undertaking the trial, for the sample size available, we calculated the power to detect a difference in response proportions for different magnitudes of difference using a chi-square test of two proportions at a 0.05 level (Supplemental Table B-4). For example, we calculated that if the email intervention led to a 10% response rate and the telephone intervention led to an additional response rate between 15-20%, our study's power to detect the difference with would range between 75 and 92%.

3.6 Results

3.6.1 Author contact flow diagram

We present a flow diagram summarizing the author contact process in Figure 5³. Author characteristics are reported in Table 2. We assigned 87 authors to the telephone group and 89 to the email group. We report results of the telephone and email groups in Tables 3-4 and Figure 6.

3.6.2 Telephone contact

We delivered the telephone intervention between October 14th, 2016 and April 13th, 2017. We were unable to obtain phone numbers for 11 of the 87 authors randomized to the telephone group. Of these, we were able to find a phone number for a coauthor in nine cases. We stopped attempts to contact two authors of older papers by telephone contact because we could not find a phone number for the randomized author or

³ A more detailed flow diagram is presented in Supplemental Figure B-1 and described in Supplemental Text B-3

another coauthor on the paper. Where possible, the investigator sought the phone number of the authors' office. For many authors however, we could not retrieve this information and we used phone numbers related to the affiliation of the authors (e.g., academic department, clinical unit, company switchboard). The time difference for authors from Eastern Standard Time ranged from -6 hours (Honolulu, Hawaii) to +16 hours (Melbourne, Australia).

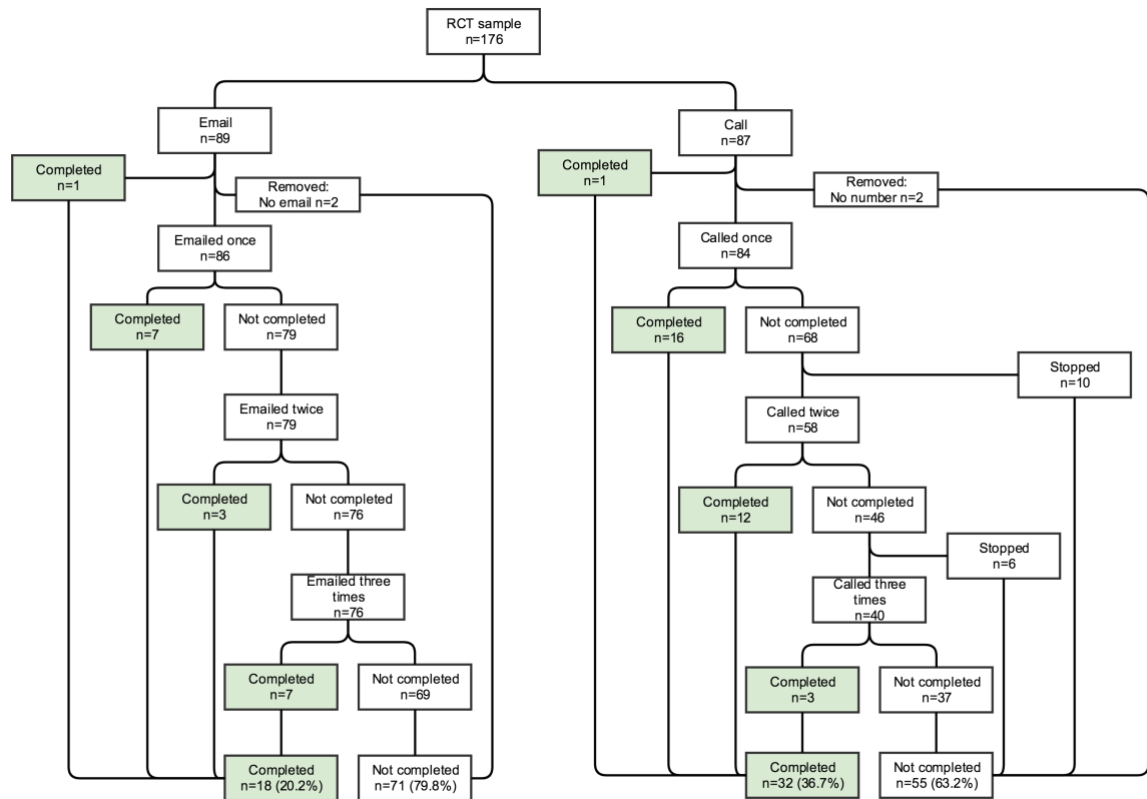


Figure 5 Flow diagram of the randomized controlled trial

Table 2 Author characteristics

	Telephone contact (n=87)	Email contact (n=89)	Total (n=176)
No. authors*			
1980-1989	1	1	2
1990-1999	11	13	24
2000-2009	49	48	97
2010-2014	26	27	53
Duplicate authors			
No. authors with >1 study	8	9	17
No. studies	18	21	39
Region†			
United States	40	39	79
Canada	3	7	10
South America	1	0	1
Europe	28	28	56
Asia	9	9	18
Africa	0	2	2
Australia & Oceania	6	4	10

*Authors were stratified on decade stratum during randomization

† Based on information available in the source publications. If author had multiple publications, information from most recent paper (used for randomization in stratification) was used

Of the 84 authors called once, 16 completed the survey and 10 were excluded from further contact, leaving 58 authors to be called twice. Reasons for stopping further attempts of telephone contact varied and are detailed in Supplemental Text B-3. In general, we stopped contact with authors when contact with phone numbers was unsuccessful and no further contact information for the author or coauthor could be found, or if an author (or their assistant) indicated that they were too busy. In a few cases, author contact was stopped due to investigator error (e.g., contact to one author was stopped because they had completed but not submitted the survey; contact to another author was stopped after a time change error led to calling them in the middle of the night. Protocol violations and implementation challenges are described in Supplemental Text B-4. Of the 58 authors who we called twice, 12 completed the survey and 6 were stopped from further contact, leaving 40 to be called a final and third time. Of the 40 authors called three times, 3 authors completed the survey. Overall, there

were 10 changes in authors (i.e., to coauthors) and 16 changes in phone numbers; contact changes were the result of either additional searches (when calls failed) or information provided by authors (e.g., of a more suitable coauthor to contact). In total, 32 authors in the telephone contact group completed the survey (31 during the intervention and 1 pre-randomization; 32/87 (36.7%) after 182 calls. We observed the telephone intervention took 20 hours (1201 minutes) to deliver (6.60 minutes/call) including the time to search for numbers and email the survey (Table 4). Based on this, the estimated cost to deliver the telephone intervention ranged between \$463-\$546 in total and \$15-\$18 per completed survey. Authors in the telephone contact group took a median of 28 minutes (range 8-901 minutes) to complete the survey⁴.

3.6.3 Email contact

We delivered the email intervention between October 25th and November 10th, 2016. We removed two authors of older papers from email contact because we could not find an email for the randomized author or another coauthor on the paper. We emailed 86 authors once; 81 emails went through, and 5 failed to send (i.e., emails that went through for contacts 1-3, but failed on attempt 4). We found new emails for the same author in all five cases of failed emails and resent the first email to these authors. Of these 86 authors emailed once, 7 completed the survey, leaving 79 authors to be sent a second email. Of the 79 authors sent a second email, 3 authors completed the survey, leaving 76 authors to be sent a final email reminder. Finally, of the 76 authors sent a third email, 7 authors completed the survey. In total, 18 authors in the email contact group completed the survey (17 during the intervention and 1 pre-randomization; 18/89 (20.2%)) after 241 emails. The email intervention took a total of 10 hours (601 minutes) to deliver (2.49 minutes/email). Based on this, the estimated cost to deliver the email intervention ranged between \$232-\$273 (\$13-\$15/completed survey). Authors in the

⁴ The high upper range in observed time to complete surveys in the telephone group is likely due to a loaded survey being left open on a web browser; 901 minutes was an extreme outlier. If removed, the range was 8-389).

email group took a median of 32 minutes (range 14-166 minutes) to complete the survey.

Table 3 Results of contact by telephone vs. email on survey completion

	Telephone contact (n=87)	Email contact (n=89)
No. attempts (telephone or email)		
Attempt 1	84	86
Attempt 2	58	79
Attempt 3	40	76
Total	182	241
No. completed (Response rate)		
Pre trial*	1 (NA)	1 (NA)
Post Attempt 1†	16 (19.0%)	7 (8.1%)
Post Attempt 2†	12 (20.7%)	3 (3.8%)
Post Attempt 3†	3 (7.5%)	7 (9.2%)
Total‡	32 (36.8%)§	18 (20.2%)
No. completed (N strata)		
1980-1989	0 (1)	0 (1)
1990-1999	2 (11)	3 (13)
2000-2009	23 (49)	8 (48)
2010-2014	7 (26)	7 (27)
Total time to deliver intervention (min)		
Attempt 1	511 (6 mins/attempt)	351 (4 mins/attempt)
Attempt 2	452 (8 mins/attempt)	162 (2 mins/attempt)
Attempt 3	238 (6 mins/attempt)	88 (1 min/attempt)
Intervention total	1201 (7 mins/attempt)	601 (2 mins/attempt)
Intervention total/number completed	38	33
Total cost to deliver intervention (\$ lower - \$ upper)		
Attempt 1	\$197-\$232	\$135-\$159
Attempt 2	\$174-\$205	\$62-\$74
Attempt 3	\$92-\$108	\$34-\$40
Intervention total	\$463-\$546	\$232-\$273
Intervention total/completed	\$15-\$17	\$13-\$15
Time to complete survey by authors		
Mean (SD)	87 (165)	44 (38)
Median (Min-max)	29 (8-901)	32 (14-166)

*2 surveys (1 in each arm) were completed prior to delivery of the call and email interventions

†n completed/N attempted contact by telephone or email

‡n completed/N randomized contact by telephone or email

§Adjusted odds ratio in favour of the telephone intervention, adjusted for strata groups: OR 2.26; 95% CI 1.10-4.76)

||Time to deliver intervention attempt includes time to search, call and, email for telephone intervention; and prepare, send and manage flow of emails for email intervention

Table 4 Details on time (minutes) to deliver telephone and email interventions

Telephone contact (n=87)					Email contact (n=89)				
Attempt	Search	Call	Email	Total*	Prepare	Send	Failed	Response flow	Total*
1	450	222	148	820	240	48	63	0	351
2	42	140	57	239	20	125	0	17	162
3	19	90	33	142	0	88	0	0	88
Total	511	452	238	1201	260	261	63	17	601

*Total time to deliver Attempt n=1,2,3 (e.g., for Attempt 1 in Telephone contact, Total time was calculated by summing the time taken to search, contact, and email during Attempt 1).

3.6.4 Comparison of telephone and email contact

Based on observed response proportions of 36.7% for the telephone contact group and 20.2% for the email contact group (difference 16.5%), we calculated an adjusted OR of a 2.26 (95% CI 1.10-4.76), indicating a significant advantage to the telephone group in the odds of completing the survey. Compared to the email contact group however, the telephone contact intervention took more time to deliver (20 hours in total vs. 10 hours) and was more expensive (average total cost \$504 vs. \$252).

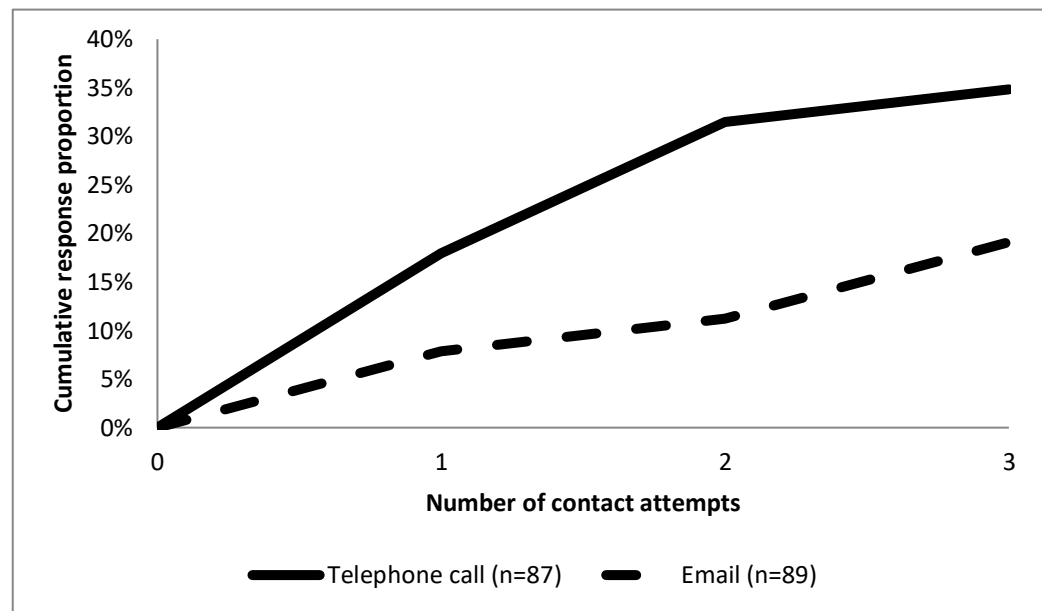


Figure 6 Cumulative response proportion of telephone and email contact

3.7 Discussion

Incomplete reporting of primary study methods and results poses substantial challenges for systematic reviewers. Although reviewers will often attempt to contact authors to obtain additional information, evidence to inform the decision to use of one author contact strategy over another is limited. We sought to compare the effectiveness of two commonly used author contact strategies (telephone vs. email) with respect to author response rates and associated research costs in a parallel group trial. We found that authors randomized to contact by telephone were significantly more likely to respond to our request for additional information compared to those randomized to contact by email, although calling authors required more investigator time and was more expensive. To our knowledge, this is the first study to empirically compare the effectiveness of telephone vs. continued emailing of non-responding authors.

The response rates (20.2% in the email group and 36.7 % in the telephone group; difference 16.5%) in our trial are similar with those in one uncontrolled study [96] and one non-randomized comparative study [91]. For example, one study compared author contact by letter, email, or both, found a 47% response rate among authors contacted by email alone [91]. Unlike our study however, the calculation of response rate excluded over a third of the authors with no contact information (n=95), and thus would likely have been lower if the complete sample of authors were randomized to contact groups and included in calculations as in our study. In another observational study, investigators achieved a 38% response rate by contacting authors using a mixed modes approaches (three contact attempts by email, one attempt by phone, one final attempt by email) [96]. The response rate for the sequential email-phone intervention was consistent with findings from our telephone group and supported evidence from survey literature on the added value of mixed mode contact for survey non-responders compared to continued use of a single mode [101].

Although, contacting authors by telephone increased the response rate compared to email, we found implementation of the telephone intervention (including searching, calling, and emailing) took substantially longer and was more difficult to standardize. First, searching for phone numbers often took a long time and was not always straightforward. While some publications included telephone numbers for corresponding authors, most did not, and had to be sought through web searches that took between 1-21 minutes per article. During these searches, we found it easier to locate phone numbers for authors affiliated with research universities and institutes compared to authors working in industry or clinical settings. Our protocol did not define an upper limit to when searches should be ceased if unsuccessful, and thus it was difficult to know when to stop if no number could be found (or a less ideal number was found). Second, we found that calling authors up to three times in diverse time zones took much longer to deliver and track as compared to sending a group of emails in one sitting. Third, there were unexpected challenges in reaching authors directly by telephone that were not anticipated by our protocol. In many cases, the assistant of an author in the telephone group requested an email of our request be sent to the assistant and/or the author. In other cases, the authors' voicemail requested email contact.

Our study has several limitations. First, we contacted authors in English, which may have posed a barrier for author contact by the investigator and/or survey completion by the author. As all studies included in the review were all published in English, we believe authors had a sufficient working knowledge of English to read the request delivered in the email intervention and complete the survey. Bias due to non-English language may have been introduced in the telephone arm since communicating with non-English receptionists, automated switchboards, or even author voicemails was considerably more challenging. Since actually connecting with authors directly usually led to them completing the survey, this would have led to an underestimation of the response rate for the telephone group. Second, one researcher who was unblinded to the study hypothesis delivered both interventions and collected and analyzed study data. Although

the researcher sought to deliver interventions according to protocol and measure outcomes (particularly time) without bias, two protocol violations occurred in the delivery of the telephone intervention and several measurements of time were missed in both groups. The two protocol violations were the result of the investigator emailing two authors who could not be reached by telephone due to incorrect numbers. In one case, the investigator replied to a pre-RCT email from the author to troubleshoot the survey URL. In the other, the investigator sent a new email to a corrected email address obtained during searches for the authors' telephone number. As both authors went on to complete the survey, these protocol violations may have led to an overestimate of the effect of the telephone intervention. Given the magnitude of the effect in favour of the telephone intervention, we do not expect these protocol violations to have substantially biased or interpretations of the effect size. Errors in the measurement of time primarily occurred during follow-up of contact attempts (e.g., responding to author emails, following up with authors in the telephone group who agreed to complete the survey). Errors were estimated to be equally distributed between groups, and therefore also would not be expected to change interpretations. Third, the email intervention was delivered in a narrow time frame of one month while the telephone intervention was delivered over six months. As the trial was not completely parallel, response proportions may have been differentially influenced by temporal trends. As the email response proportion of the email group in the trial (20.2%) was consistent with the response rate prior to the trial (27.2%), we do not expect that temporal trends markedly biased results.

If author contact is to remain an expected standard of systematic review practice, further work is needed to optimize its efficiency and effectiveness. In addition to evidence informing the method of author contact, further evidence is needed on the added value of information obtained through author contact, such as studies that have evaluated the effect of author contact on completeness of reporting [96,102]. In future work using obtained responses to the survey here, we intend to compare agreement between author and reviewer coding of interventions and explore the extent to which

additional information on effect modifiers can be included in meta-analyses. Further evidence is needed to inform the extent to which review resources should be dedicated to obtaining ideal data through author contact as compared to the conduct and/or optimization of other review processes. For example, methodological texts cite varying reasons for contacting authors to capture, clarify, and confirm various information along the review pipeline. Are systematic reviewers expected to perform all these tasks and if so, how vigorously should they pursue them? We showed a wide range of time may be required to locate and contact some authors; while extra effort undoubtedly increased the time and costs required for author contact, dedicating extra resources to some of the more challenging cases may be needed to avoid bias [38]. While increased automation of systematic review processes [28,103] may free up additional time for reviewers to allot on contacting authors, further work is needed to understand the trade-offs that will be necessary as novel methods are adopted in different aspects of the review process.

3.8 Conclusion

Contacting non-responding authors of included studies for additional information by telephone increases response compared to repeat emailing but is more time consuming and expensive. We anticipate that findings from this randomized trial may help inform methodological and budgeting decisions for future systematic reviews, especially those planning multiple waves of review updating. Future work will explore the added value of the information obtained on review findings.

Acknowledgements

The authors would like to acknowledge members of the scientific team and knowledge users listed in Supplemental Table B-1 for their role in developing the author survey, Mr. Anton Saarimaki (Ottawa Hospital Research Institute) for his help in developing and managing the online survey platform, and Mr. Jordi Pardo Pardo (Ottawa Hospital Research Institute) for his invaluable feedback on early drafts of this manuscript. We would also like to acknowledge Mr. Samir Nasrali, Ms. Pauline Barbeau, Mr. Mostafa

Alabousi who assisted with data abstraction of study characteristics and outcome data for the diabetes quality improvement review, some of which was used for the purpose of this study.

Declaration of interests

None to declare

Author contributions

The project was conceived by KJD and JMG and developed with critical input from NMI, IJD, and DM. KJD performed all data acquisition, analysis, and interpretation with guidance and feedback from IJD, JMG, NMI, and DM. KJD drafted the manuscript and coordinated revisions. All authors critically revised the manuscript and approved the final version to be submitted.

Details of ethical approval

This study was approved the Ottawa Health Science Network Research Ethics Board (Protocol ID: 20180429-01H).

Registration

The trial is registered on the SWAR (Study Within A Review) repository; Registration number SWAR11

(<https://www.qub.ac.uk/sites/TheNorthernIrelandNetworkforTrialsMethodologyResearch/SWATSWARInformation/Repositories/SWARStore/>)

Role of funding sources

This study is supported by a grant from the Canadian Institutes of Health Research (CIHR, FRN-123345) and a research fellowship held by KJD (Frederick Banting and Charles Best Canada Graduate Scholarship; GSD-134936). NMI is funded by a New Investigator Award from the Canadian Institutes of Health Research. JMG holds a Canada Research Chair in

Health Knowledge Transfer and Uptake and is funded by a Foundation Grant from the Canadian Institutes of Health Research. Funders played no role in the design, conduct or reporting of the study.

The following supplemental information for Manuscript 1 can be found in Appendix B:

Supplemental Table B-1	Review team and knowledge users
Supplemental Table B-2	CONSORT 2010 checklist of information to include when reporting a randomised trial
Supplemental Text B-1	Randomization code
Supplemental Text B-2	Additional details on telephone intervention
Supplemental Table B-3	Comparison of telephone contact and email interventions using Behaviour change techniques
Supplemental Table B-4	Power calculations
Supplemental Figure B-1	Detailed author flow diagram
Supplemental Text B-3	Additional details on the delivery of the telephone and email contact interventions
Supplemental Text B-4	Protocol violations or potential errors and implementation challenges

**CHAPTER 4 [Manuscript 2] Descriptive study of the
proportion of studies with unit of analysis errors in a
systematic review of diabetes quality improvement trials**

4.0 [Manuscript 2] Descriptive Study Of The Proportion Of Studies With Unit Of Analysis Errors In A Systematic Review Of Diabetes Quality Improvement Trials

4.1 Preface to Manuscript 2

Study 2 presents a descriptive analysis of the proportion of cluster-randomized trials with unit of analysis errors in the diabetes quality improvement systematic review. In updating the diabetes quality improvement review, there were concerns that unit of analysis errors may not have been adequately addressed within the previous version of the review. I conceived the idea for this study, with input from Dr. Taljaard, Dr. Grimshaw, and Dr. Ivers to address these concerns and get an accurate assessment of the extent of unit of analysis errors in the diabetes quality improvement literature.

I developed the original protocol for the study using a sample of cluster-randomized trials in our review (n=26 cluster randomized trials reporting mean glycemetic control, HbA1c) that I extracted with Dr. Taljaard. Katrina Sullivan, a Research Associate at the Ottawa Hospital Research Institute, was my second reviewer for the data abstraction for the remainder of the outcomes. Dr. Taljaard acted as a third reviewer to resolve conflicts or clarify statistical interpretations. I conducted all analyses, interpreted the results, and drafted the manuscript. Ms. Sullivan, Dr. Taljaard, Dr. Ivers, Dr. Moher, and Dr. Grimshaw contributed to the interpretation of the study results and edited the manuscript. In the acknowledgements section of the manuscript, I acknowledge Mr. Samir Nasralli, Ms. Pauline Barbeau, Mr. Mostafa Alabousi who assisted with data abstraction of study characteristics and outcome data for the diabetes quality improvement review, some of which was used for the purpose of this study.

4.2 Title page

Title: Descriptive study of the proportion of studies with unit of analysis errors in a systematic review of diabetes quality improvement trials

Authors: Danko KJ^{1,2}, Sullivan KJ¹, Taljaard M^{1,2}, Ivers NM^{3,4,5}, Moher D^{1,2,6}, Grimshaw JM^{1,2,6}

Affiliations:

¹Clinical Epidemiology Program, Ottawa Hospital Research Institute, Ottawa, Canada

²School of Epidemiology and Public Health, University of Ottawa, Ottawa, Canada

³Family Practice Health Centre, Women's College Research Institute, Toronto, Canada

⁴Institute for Health Systems Solutions and Virtual Care, Women's College Hospital, Toronto, Canada

⁵Department of Family and Community Medicine, Dalla Lana School of Public Health, Institute of Health Policy Management and Evaluation, University of Toronto, Toronto, Canada

⁶Department of Medicine, University of Ottawa, Ottawa, Ontario, Canada

Corresponding author:

Kristin Julianna Danko

The Ottawa Hospital – General Campus, Box 711

501 Smyth Road

Ottawa, Ontario, Canada

K1H 8L6

Email: kdanko@ohri.ca

4.3 Abstract

Background: Complex healthcare interventions are often evaluated using cluster-randomized trials (CRTs) that randomize groups of individuals to an intervention. Analyses of CRTs that fail to account for within-cluster correlations of individual data have 'unit of analysis' errors characterized by overly precise estimates of intervention effect. Meta-analysis of effect estimates with unit of analysis errors can lead to biased summary estimates and overly precise confidence intervals. The objective of this study was to determine the proportion of studies with unit of analysis errors in a large sample of CRTs (n=73) included in a review of diabetes quality improvement trials.

Methods: For each of the 13 review outcomes, two researchers independently assessed the proportion of studies for which: 1) analysis accounting for clustering was used; 2) appropriate intervention effect estimates and standard errors could be extracted or derived (continuous outcomes only); and 3) an estimate of the ICC was reported. For continuous outcomes, we sought an estimate of the mean difference and its adjusted standard error (or p-value or confidence interval to calculate an adjusted standard error). For dichotomous outcomes, we sought an estimate of the ICC to obtain the adjusted number of events and effective total sample size.

Results: The frequency of studies appropriately accounting for the cluster randomized design varied across outcomes ranging from 1 of 4 studies (25%) reporting the proportion of patients experiencing harms to 21 of 26 studies (81%) reporting mean low-density lipoprotein. Few studies reporting continuous outcomes reported an adjusted estimate of the standard error (range 3-4%); calculation of adjusted standard errors from reported information was feasible for only 48-58% of studies. The prevalence of studies reporting an estimate of the ICC was low across all outcomes (range 0% to 23%).

Conclusions: CRTs pose important methodological challenges for meta-analysis. Reviewers need to identify unit of analysis errors and adjust estimates accordingly. Many CRTs in our study conducted appropriate analysis but did not provide adequate details to extract adjusted data for meta-analysis. Authors of future CRTs should report ICCs for all reported outcomes and, ideally, adjusted standard errors.

Keywords: unit of analysis errors, cluster randomized trials, complex interventions, meta-analysis, intraclass correlation coefficient, intracluster correlation coefficient

4.4 Background

Complex healthcare interventions are often evaluated using cluster-randomized trials (CRTs) that randomize groups (clusters) of individuals to an intervention but assess the impact of the intervention using data from individuals within groups [15,104,105]. CRTs create statistical challenges for trialists as individual responses within a cluster tend to be correlated, thus violating the assumption of independence required for most standard statistical tests to be valid [16]. To obtain valid inferences of effect, the degree of within-cluster correlation must be accounted for during analysis of a CRT [17,18]; failure to do so leads to 'unit of analysis' errors [19] characterized by overly precise estimates of intervention effect [16]. Meta-analysis of effect estimates with unit of analysis errors can lead to biased summary estimates and overly precise confidence intervals [16]. Reviewers must therefore identify and correct unit of analysis errors prior to meta-analysis. The objective of this study was to determine the proportion of studies with unit of analysis errors and availability of data to correct unit of analysis errors in a large systematic review of diabetes quality improvement (QI) trials.

The methods recommended by Cochrane to identify and correct unit of analysis errors in systematic review require a multistep process (Figure 7) in which reviewers must first determine whether analysis was appropriate, and then either extract adjusted estimates directly, or extract unadjusted estimates for external approximate correction [64]. Corrections aim to approximately account for the 'design effect' [72] of the CRT by reducing the sample size of the study to its 'effective sample size' [73] or inflate the observed variance to account for the between-cluster heterogeneity. The design effect is calculated using an estimate of the average cluster size (m) and an estimate of the intraclass (intracluster) correlation coefficient (ICC; typically denoted by the Greek letter ρ) as defined by the following: $(1 + [m - 1] * \rho)$ [16]. Once study-level data are corrected, meta-analysis may proceed as usual.

There are several challenges to operationalizing the assessment, extraction, and correction of CRTs. First, CRTs are complicated designs and different studies may use different statistical methods to adjust for the cluster effect (e.g. multilevel models, generalized estimating equations [GEE], etc.). Reviewers may find it challenging to distinguish appropriate cluster analysis methods from inappropriate methods and as a result fail to correctly identify unit of analysis errors. For this reason, the Cochrane Handbook recommends reviewers seek statistical advice in determining the appropriateness of cluster analysis methods [64].

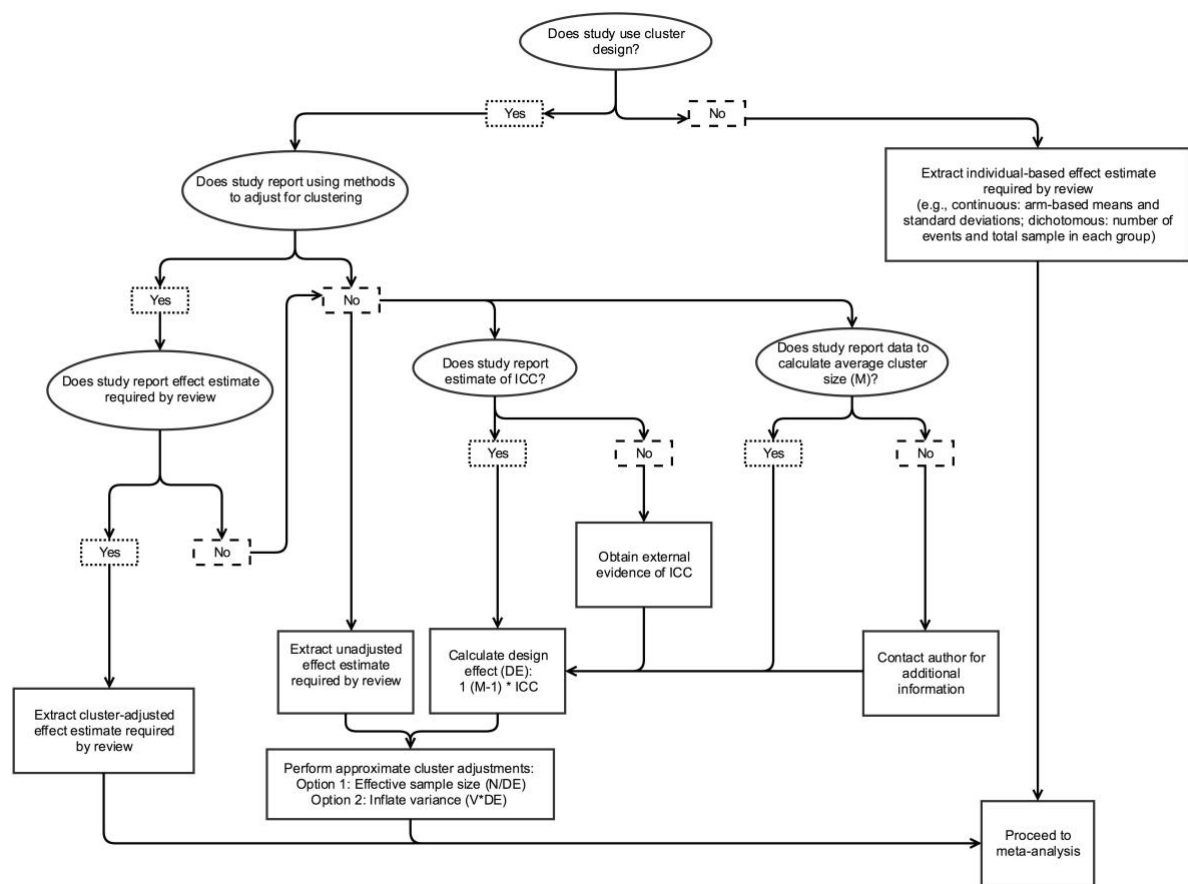


Figure 7 Process for identifying and correcting unit of analysis errors of data to be synthesized in meta-analysis

Second, even if a reviewer can correctly identify appropriate cluster analysis methods, CRTs may fail to report enough information about analysis methods for reviewers to

make this assessment (e.g., vague statement about clustering being taken into account but specific method not described; unclear if adjustment made for all reported outcomes or just primary outcomes). Third, CRTs may conduct appropriate analyses and report methods with sufficient detail to determine they were appropriate, but report adjusted data in such a way that cannot it be readily extracted and used in meta-analytic models. For example, standard methods for synthesizing continuous effect estimates prefer a measure of the study mean and its standard error [74]. Although methods exist to convert reported measures of spread (e.g., 95% confidence interval) to a standard error for standard pairwise individual RCT data, the extent to which these calculations can or should be applied to summary statistics from CRT data is unclear. Forth, many CRTs fail to report an estimate of the ICC, and thus reviewers will need to use an external estimate of the ICC to perform approximate corrections.

Richardson et al. [20] assessed the methodological quality and reporting of 50 Cochrane reviews that included CRTs published between June and November 2013. Only 26% of reviews stated whether included study results adjusted for the cluster design, and where studies did appropriately adjust for clustering, only 68% of reviews correctly extracted adjusted effect estimates. Most concerning, was the fact that in studies with unit of analysis errors, only 38% of reviews excluded unadjusted results from the meta-analysis. Based on the findings of this methodological review, the authors believed that, “review authors do not understand or are not aware of appropriate methods and reporting in reviews including [cluster]-RCTs” and called for a “consolidated and comprehensive set of guidelines to help improve the quality of published reviews including cluster randomized trials” including a possible PRISMA extension for CRTs.

4.4.1 Case example: Diabetes quality improvement review

In 2012, team members JMG, NMI, and DM published a systematic review that assessed the effect of QI interventions on a range of 13 process- and intermediate-patient outcomes [55]; 48 of the 142 (34%) included trials were CRTs. A recent update of the

review searched up to the end of 2014 and included 73 of 272 (27%) CRTs. As unit of analysis errors may not have been adequately addressed within the previous version of this review, we sought to accurately determine the extent of unit of analysis errors in the complete sample of diabetes CRTs. For studies that were analyzed appropriately, we sought to determine the extent to which we could extract adjusted effect estimates and estimates of the ICC.

4.5 Methods

For each of the 13 review outcomes in the sample of 73 CRTs [37,80,106–176], we assessed the proportion of studies for which: 1) appropriate analysis was used (i.e., clustering was adjusted for); 2) adjusted effect sizes could be extracted (or calculated) and 3) an estimate of the ICC was reported. All assessments and extractions were performed by two independent researchers (KJD and KJS) and confirmed through consensus. Where consensus was not achieved, or we experienced challenges in interpreting CRT reports, we sought advice from a statistician with expertise in cluster methods (MT), as recommend in the Cochrane Handbook [64]. Specific methods for the assessment and extraction of CRTs in our review are reported below.

4.5.1 Did the study report the use of appropriate methods to adjust for clustering?

We examined the methods section of CRT reports for evidence of appropriate analysis. A CRT was accepted as having conducted appropriate analysis of the outcome in question if they reported explicitly that clustering was accounted for, for example, if they used generalized mixed effects regression (random effects, multi-level, or hierarchical modeling), Generalized Estimating Equations (GEE), analysis of cluster-level summaries, or test-statistics adjusted for clustering. While some studies applied adjustments to all study outcomes, others adjusted for outcomes selectively (e.g., reported an adjusted standard error for the primary outcome, but an unadjusted standard error for a secondary outcome). We were therefore careful to assess the appropriateness of

analysis for each outcome uniquely (e.g., examining the methods described in analysis section and description of results for the specific outcome in question).

Table 5 Outcomes reported in CRTs included in the diabetes QI systematic review

Outcome	CRTs included in the review		
	No. of CRTs	N population	N clusters
<i>Continuous</i>			
Glycated hemoglobin (mean)	54	86233	3809
Systolic blood pressure (mean)	40	72474	3128
Diastolic blood pressure (mean)	34	50060	2218
Low-density lipoprotein (mean)	26	62346	2269
<i>Dichotomous</i>			
Acetylsalicylic acid (proportion use)	10	16842	1141
Statin (proportion use)	10	23885	1251
Anti-hypertensive (proportion use)	11	26363	1850
Retinopathy (proportion screened)	27	72318	2156
Foot (proportion screened)	24	42973	1757
Renal (proportion screened)	20	45751	1310
Hypertension control (proportion controlled)	20	50007	2010
Smoking (proportion smoking)	14	24851	2159
Harms (proportion to experience harms)	4	2258	549

4.5.2 Did study report effect estimate required by the review?

The review had four continuous and nine dichotomous outcomes (Table 5). If a study was assessed as having conducted appropriate analysis, we sought to extract for continuous outcomes: a mean difference and its adjusted standard error and for dichotomous outcomes: the number of events, sample size per group and an estimate of the ICC. If the CRT reported a confidence interval or p-value from which an adjusted standard error could be derived, we extracted these values and calculated an adjusted standard error for the mean difference according the methods described in Supplemental Text C-1.

4.5.3 Did the study report an estimate of the ICC?

For all outcomes we determined whether an estimate of the ICC was provided.

We calculated the proportion of studies reporting appropriate analysis methods, proportion of adjusted continuous data that could be extracted (or calculated), and proportion of studies reporting an estimate of the ICC to correct unit of analysis errors. Proportions were calculated as a fraction of the total number of studies reporting the outcome.

4.6 Results

4.6.1 Did the study report the use of appropriate methods to adjust for clustering?

Overall, 54 of 73 (75%) studies included in the review used appropriate methods to adjust for clustering for at least one outcome. Across outcomes, an average of 67% of studies adjusted for clustering, ranging from 25% for proportion of patients to experience harms to 81% for mean low-density lipoprotein (LDL) (Figure 8). For the main review outcome, glycemic control (mean HbA1c), 17 of 54 studies (69%) were assessed to have appropriately adjusted for clustering. The most common methods reported to adjust for clustering were mixed models (n=21) and GEE (n=11).

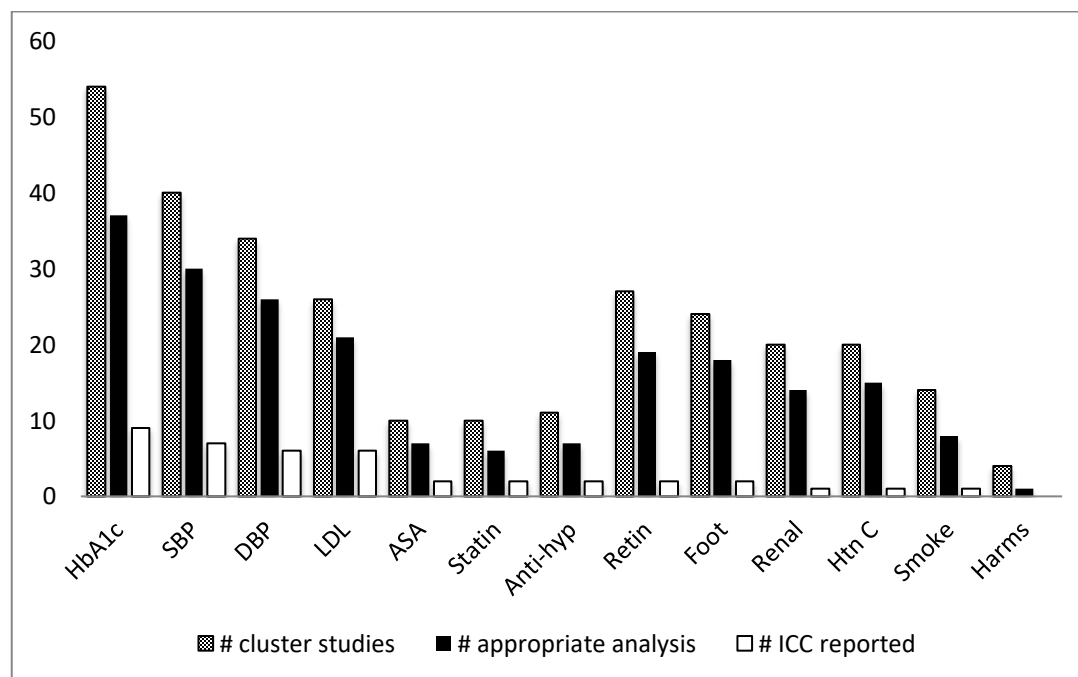


Figure 8 Number of CRTs with appropriate analysis and reporting across outcomes

4.6.2 Did study report effect estimate required by the review?

Although two-thirds of studies were analyzed appropriately, we found few studies reported adjusted standard errors of the mean difference (for continuous outcomes) or estimates of the ICC (for dichotomous outcomes). For example, only 5% of studies that used appropriate methods to analyze HbA1c reported an adjusted standard error that could be directly extracted for meta-analysis; similar results were found for other continuous outcomes, systolic blood pressure (3%), diastolic blood pressure (4%), and LDL (5%) (Table 6). Reporting of ICCs across dichotomous outcomes ranged from 0% (0/4) of studies that reported evidence on harms to 20% (2/10) of studies that reported evidence on proportion of patients to be prescribed statins (Figure 8). A larger portion of appropriately analyzed studies reported confidence intervals and p-values that adjusted for clustering. Using this data, we were able to derive adjusted standard errors for 67-73% of appropriately analyzed studies that reported a continuous outcome (Table 6).

Table 6 Analysis and reporting of continuous outcomes in cluster randomized trials in a diabetes quality improvement review

	HbA1c (n=54)	SBP (n=40)	DBP (n=34)	LDL (n=26)
Reported analysis did not account for the cluster randomized design	17 (31%)*	10 (25%)*	8 (24%)*	5 (19%)*
Reported analysis accounted for the cluster randomized design	37 (69%)*	30 (75%)*	26 (76%)*	21 (81%)*
Appropriate SE for the MD was reported	2 (5%)†	1 (3%)†	1 (4%)†	1 (5%)†
Appropriate SE for the MD was not reported but could be derived from information provided in report	26 (70%)†	20 (67%)	19 (73%)†	15 (71%)†
Appropriate SE for the MD was not reported and could not be derived	9 (24%)†	9 (30%)†	6 (23%)†	5 (24%)†
Total number of trials requiring adjustment to conduct meta-analysis of the MD	26 (48%)*	19 (48%)*	14 (41%)*	10 (38%)*
Due to inappropriate analysis	17 (65%)‡	10 (53%)‡	8 (57%)‡	5 (50%)‡
Due to incomplete reporting	9 (35%)‡	9 (47%)‡	6 (43%)‡	5 (50%)‡

DBP: diastolic blood pressure; HbA1c: glycated hemoglobin; LDL: low-density lipoprotein; MD: mean difference; SBP: systolic blood pressure; SE: standard error

*n/N studies reporting outcome

†n/N studies reporting outcome that were appropriately analyzed

‡n/N studies requiring adjustment to conduct meta-analysis of the MD

Of the remaining studies that conducted appropriate analysis, we could not extract or derive an adjusted standard error (23% of studies reporting diastolic blood pressure to 30% of studies reporting systolic blood pressure). In some cases, adjusted standard errors could not be obtained because the study did not report adjusted p-value or confidence interval; in other cases it was due to missing data (e.g., did not report the number of clusters in groups needed to calculate degrees of freedom), characteristics of the data (e.g., mean difference equal to 0), or unique analyses that could not be used to perform derivations (e.g., 3-way ANOVA). Thus, due to inappropriate analysis or incomplete reporting, we observed 38-48% of all studies reporting a continuous outcome (10/26 studies reporting low density lipoprotein to 26/54 studies reporting HbA1c) needed to be corrected prior to meta-analysis. Complete details on assessed appropriateness of methods, derivation of standard errors, and reporting of ICCs for continuous outcomes are reported in Supplemental Table C-1. Details on the assessed appropriateness of methods and reporting of ICCs for dichotomous outcomes are reported in Supplemental Table C-2.

4.6.3 Did the study report an estimate of the ICC?

The percentage of CRTs that reported estimates of the ICC was low, ranging from 0% of studies reporting evidence on harms to 23% of studies reporting evidence on low-density lipoprotein. The majority of unit of analysis errors identified above needed to be corrected using an external estimate of the ICC.

4.7 Discussion

In this study, we determined the proportion of CRTs included in a sample of diabetes QI trials that appropriately adjusted for the cluster design and reported details of analysis with sufficient detail that their appropriateness could be assessed. In addition, we assessed the proportion of studies for which we could extract (or derive) adjusted standard errors for meta-analysis of continuous outcomes, and the proportion of studies for which we could extract estimates of the ICC for dichotomous outcomes. Finally, we

determined the proportion of ICCs reported across continuous outcomes that could be used to correct remaining unit of analysis errors. We found that while most studies reported using appropriate methods to adjust for clustering, many failed to report the necessary data required to use the methods recommended by the Cochrane Handbook [5]. Thus, after our best efforts to extract CRT evidence corrected for clustering, 38-48% of continuous outcomes and 80-100% dichotomous outcomes needed adjustment before meta-analysis.

Our findings on the frequency of appropriate analysis of our review outcomes (25%-81%;) and reporting of ICCs (0%-23%) are consistent with earlier reviews of published CRTs [68,104,105,177]. For example, in a review of 152 CRTs in primary healthcare published between 1997 and 2000, 59% of studies accounted for clustering in analyses and 7% reported estimates of ICC [105]. In a later review of 300 CRTs published between 2000 and 2008, 70% of studies accounted for clustering in analyses and 16% reported estimates of ICC [68]. More recently, in a review of 29 CRTs evaluating complex interventions in general practice settings published between 1999 and 2014, 93% accounted for clustering in analysis and 14% reported estimates of ICC (although this was improved to 39% after contacting authors) [104]. Interestingly, one review found CRTs were more likely to have reported using appropriate methods to adjust for clustering when a statistician or epidemiologist was listed as a co-author on the report [177].

There are several limitations of our study. First, the assessment of study conduct was based on published CRT reports. It is possible that CRTs appropriately adjusted for clustering but failed to report methods of adjustment with sufficient detail to support reviewers' assessment. As misclassification of a unit of analysis error would lead to the over-adjustment of an effect estimate from a CRT, and thus lead to a more conservative summary effect from the meta-analysis, we were purposively conservative in our assessments of CRT analysis. Second, our study focused on unit of analysis errors only and did not assess (or attempt to correct) other methodological challenges associated

with CRTs [68]. For example, we did not assess the use of restricted randomization procedures for cluster allocation (e.g., matching, stratification) [178] or account for the use of different baseline covariates included in adjusted analysis models (in addition to the cluster effect) [179]. Finally, our calculations to derive adjusted standard errors may have introduced errors into review data due to human error or incorrect assumptions. To limit this, two reviewers performed calculations and we chose conservative assumptions (e.g., t-distribution to account for the smaller sample of clusters).

Our findings have implications for both primary studies and evidence synthesis of CRTs (Box 2). It is clear that a sizeable proportion of CRTs still fail to conduct (or report) appropriate analysis to account for clustering effects. Future initiatives, such as revisions the CONSORT extension for CRTs [180], should continue to educate trialists' on the importance of cluster adjustment for valid inference and reporting methods of adjustment - not only from the perspective of interpretation of the primary CRT, but importantly, from the perspective of supporting efficient and valid evidence synthesis (which in turn may be used to improve the design and conduct of future primary CRTs). In addition, primary CRTs should report an estimate of the ICC for all outcomes analyzed (not just the primary outcome, as recommended in the CONSORT statement) and an estimate of the adjusted standard error (not just the 95% confidence interval, as recommended in the CONSORT statement) [180]. With respect to evidence synthesis of CRTs, further work is needed to produce consolidated guidance on specific methods for handling clustered data in extractions and meta-analysis. Until that time, reviewers should continue to seek substantial statistical support from experts in cluster methods to ensure the adequate assessment of CRT analyses and extraction (or correction) of CRT evidence. Where calculations are performed to convert reported estimates to adjusted standard errors, reviewers should report methods used. Finally, reviewers should extract and report estimates of all reported ICCs used to inform their corrections of unit of analysis errors (or help inform other reviewers corrections).

Box 2 Recommendations for the conduct and reporting of future CRTs and evidence synthesis of CRTS

Recommendations for future CRTs

- Use appropriate method to adjust for clustering for all outcomes
- Report method of analysis used to adjust for clustering for all outcomes with enough detail to support reviewer assessments of whether an appropriate method was used
- Report adjusted standard errors for all outcomes
- Report estimates of the ICC for all outcomes

Recommendations for future systematic review and meta-analysis of CRTs

- Seek ongoing statistical advice to guide how CRTs will be assessed, extracted, and synthesized
- Assess method of analysis for each outcome reported in a CRT
- Derive adjusted estimates of standard error where necessary and report methods used for calculations
- Extract estimates of all ICCs to inform correction of unit of analysis errors

4.8 Conclusion

CRTs pose important methodological challenges for systematic reviews and meta-analyses. Reviewers need to be aware of potential unit of analysis errors and adjust estimates accordingly. While many CRTs trials in the diabetes QI review adjusted for clustering, study reports did not give enough information to extract adjusted estimates (or correct unadjusted estimates) required for meta-analysis. Guidance for conduct and reporting of primary CRT and secondary synthesis of CRTs could be improved to more efficiently support the synthesis process and inform future trials.

Acknowledgements

The authors would like to acknowledge Mr. Samir Nasralli, Ms. Pauline Barbeau, Mr. Mostafa Alabousi who assisted with data abstraction of study characteristics and

outcome data for the diabetes quality improvement review, some of which was used for the purpose of this study.

Conflict of interest

None to declare

The following supplemental information for Manuscript 2 can be found in Appendix C:

Supplemental Text C-1	Procedure used to derive approximate standard errors for continuous outcomes from reported p-values or confidence intervals
Supplemental Table C-1	Study details of included cluster randomized trials in the diabetes quality improvement review (continuous outcomes)
Supplemental Table C-2	Study details of included cluster randomized trials in the diabetes quality improvement review (dichotomous outcomes)

**CHAPTER 5 [Manuscript 3] Development Of An ICC Database
And Imputation Model Addressed Unit Of Analysis Errors In A
Systematic Review Of Diabetes Quality Improvement Trials**

5.0 [Manuscript 3] Development Of An ICC Database And Imputation Model Addressed Unit Of Analysis Errors In A Systematic Review Of Diabetes Quality Improvement Trials

5.1 Preface to Manuscript 3

Study 3 presents a methodological study that investigated the utility of building a database of intracluster correlation coefficients (ICCs) and posterior predictive distribution to impute missing ICCs to correct unit of analysis errors identified in Study 2 (Manuscript 2; Chapter 4). The previous update of the diabetes quality improvement review used a single, somewhat arbitrary, but purposively conservative estimate of the ICC (0.07) to correct unit of analysis errors identified by reviewers. Use of a single ICC estimate however did not make the most of the available ICC evidence or incorporate uncertainty about estimates of the ICC into meta-analysed results. Study 3 therefore sought to collect and synthesize existing evidence on ICCs to improve ICC imputation, and thus, reviews estimates.

I conceived the idea for this study, and developed its protocol, in collaboration with Dr. Dahabreh during my visit to Center for Evidence Synthesis in Health at Brown University June-September 2017. I performed all data collection and extraction. I conducted all analyses, interpreted the results, and drafted the manuscript. Dr. Dahabreh provided statistical advice contributed to the interpretation of results. Dr. Taljaard, Dr. Ivers, and Dr. Grimshaw contributed to the interpretation of the study results and edited the manuscript. In the acknowledgements section of manuscript, I acknowledge Dr. Marion Campbell and Dr. Graeme MacLennan, who provided additional evidence on ICC estimates from the ICC database housed by the Health Services Research Unit at The University of Aberdeen.

5.2 Title page

Title: Development of an ICC database and imputation model addressed unit of analysis errors in a systematic review of diabetes quality improvement trials

Authors: Danko KJ^{1,2}, Taljaard M^{1,2}, Ivers NM^{3,4,5}, Moher D^{1,2,6}, Grimshaw JM^{1,2,6}

Affiliations:

¹Clinical Epidemiology Program, Ottawa Hospital Research Institute, Ottawa, Canada

²School of Epidemiology and Public Health, University of Ottawa, Ottawa, Canada

³Family Practice Health Centre, Women's College Research Institute, Toronto, Canada

⁴Institute for Health Systems Solutions and Virtual Care, Women's College Hospital, Toronto, Canada

⁵Department of Family and Community Medicine, Dalla Lana School of Public Health, Institute of Health Policy Management and Evaluation, University of Toronto, Toronto, Canada

⁶Department of Medicine, University of Ottawa, Ottawa, Canada

Corresponding author:

Kristin Julianna Danko

The Ottawa Hospital – General Campus, Box 711

501 Smyth Road

Ottawa, Ontario, Canada

K1H 8L6

Email: kdanko@ohri.ca

5.3 Abstract

Background: Unit of analysis errors in cluster-randomized trials (i.e., analyses that have not accounted for clustering) must be corrected prior to meta-analysis. Corrections are often made using a single estimate of the intraclass correlation coefficient (ICC) from external literature. We investigated the feasibility of building a database of ICCs and outcome-specific posterior predictive distributions to impute missing ICCs in our review to better account for ICC uncertainty.

Methods: We combined ICC estimates from our review with those obtained from external sources. For outcomes with two or more ICC estimates, we constructed posterior predictive ICC distributions in a Bayesian framework. We compared the impact of correcting unit of analysis errors with a single ICC vs. ICCs drawn from the posterior predictive distribution of one continuous outcome.

Results: Using evidence from 59 ICCs for 12 review outcomes (range 1-10 per outcome) we estimated the posterior predictive ICC distribution for 11 review outcomes. Synthesized results of one continuous outcome were not markedly changed by our approach.

Conclusions: Building a posterior predictive distribution to impute missing ICCs is a feasible approach to correct unit of analysis errors. Further work is needed to establish whether the application of these methods leads to improved review inferences.

Keywords: unit of analysis errors, cluster randomized trials, complex interventions, meta-analysis, intraclass correlation coefficient, intracluster correlation coefficient

5.4 Background

Healthcare interventions are often evaluated using cluster-randomized trials (CRTs) in which groups of individuals are randomized to interventions [15,16,105]. For example, interventions to change healthcare professional behaviour or a healthcare policy will often randomize healthcare providers or healthcare centers, and evaluate the impact of these interventions on the outcomes of patients within clusters [15,76]. As patients within a single cluster tend to be more similar than patients in different clusters, their responses to an intervention are often correlated (i.e., cannot be considered independent observations for the purposes of sample size estimation and analysis). The 'design effect' [72] of cluster trials, defined as function of both the intracluster correlation coefficient (ICC; often defined by the Greek letter ρ) and the average cluster size (m), must be taken into account during both the design and analysis of CRTs [16,17]. Failing to account for the clustering in analysis leads to a 'unit of analysis' error [19], and more specifically, results in overly precise estimates of intervention effect and higher likelihood of spuriously statistical significant results [16]. In some cases, CRTs conduct appropriate analyses but report results in such a way that adjusted effects cannot be abstracted. Meta-analysis that include studies with unit of analysis errors (either due to incorrect analyses or incomplete reporting), can lead to both an overly precise, and biased, pooled effect estimates [64,181]. Unit of analysis errors must therefore be identified and corrected before meta-analysis can proceed [5,71].

Conventional methods for correcting unit of analysis errors in systematic reviews involve first calculating a design effect for each study ($1 + [m - 1] * \rho$) [16] and second, using the design effect to either reduce the sample size to an 'effective sample size' (analyzed $n/\text{design effect}$), or to inflate the variance of an unadjusted effect estimate (variance*design effect). Once corrections have been made, meta-analysis can proceed as usual [64]. As estimates of the ICC are rarely reported by primary studies (particularly those with unit of analysis errors) [18,68], reviewers have commonly used a single estimate of an ICC obtained from external sources (e.g., a similar study, research

databases) [5]. However the use of a single ICC to calculate a design effect does not account for the uncertainty in ICC estimates [182] and can lead to the incorrect estimate of a design effect. If the single ICC is too large then study estimates will be over-adjusted and if it is too small, then study estimates will be under-adjusted. In either case, the ICC is assumed to be known and its uncertainty is not incorporated into the uncertainty of synthesized estimates. To account for the imprecision in the underlying true value of an ICC, Turner et al. have proposed constructing prior distributions of the ICC in a Bayesian framework [79]. Prior distributions of the ICC can be used to account for the uncertainty in the ICC at both design [183] and analysis phases of a CRT [184,185] to better predict sample sizes across a range of ICC values and incorporate uncertainty into model parameters, respectively. In this study, we investigated the feasibility of building a database of ICCs and outcome-specific posterior predictive distributions to impute missing ICCs in our review to better account for ICC uncertainty.

5.4.1 Case example: Diabetes quality improvement review

Diabetes is chronic condition with multiple risk factors and potential complications that warrant attention in clinical practice [49]. Despite a strong evidence base to inform effective diabetes care, evidence indicates the suboptimal performance of a range of diabetes-related clinical behaviors [53,54]. An increasing number of trials have sought to evaluate interventions targeted at the level of the healthcare system and/or healthcare provider to improve quality of diabetes care. In 2012, members of our team (JMG, NMI, DM) undertook a systematic review of 142 trials (published by 2010) that assessed the effect of quality improvement (QI) interventions [55–57] on a range of process (e.g., foot screening) and intermediate patient (e.g., glycemic control) outcomes [55]. Of the 142 included trials, 48 were CRTs with a total sample size 84,865 patients (number of clusters=3,047). The QI interventions were often multicomponent interventions that included one or more of the following 12 QI strategies: audit and feedback, case management, clinician education, clinician reminders, electronic patient registry, facilitated relay, patient education, patient reminders, promotion of self management,

team changes, continuous quality improvement, and financial incentives. To correct unit of analysis errors, the review team used a single ICC (0.07). The choice of the single ICC was somewhat arbitrary and purposively conservative; different ICCs would have led to the calculation of different design effects and resultant effective sample sizes. For example, for the main outcome, mean difference in glycated hemoglobin (HbA1c%), the single ICC reduced the sample size of the 33 CRTs from 34,148 to an effective sample size of 12,052. Had reviewers chosen to use a smaller ICC estimate of 0.027 reported in one of the included CRTs [80], it would have led to a larger effective sample size of 14,786.

Our recent update of the diabetes QI review includes 272 trials with a total sample size of 220,012 patients. Of these trials, 73 are CRTs with an unadjusted sample size of 151,384 patients. We determined the proportion of studies with unit of analysis errors in the updated sample, and proportion of studies that reported estimates of the ICC to correct unit of analysis errors [186]. An important proportion of CRTs failed to account for clustering in their analyses ranging from 19% of CRTs for mean low density lipoprotein (LDL; 5/26 CRTs did not adjust for clustering) to 75% of CRTs for proportion of patients to experience harms (3/4 studies did not adjust for clustering). For the most common outcome HbA1c, 31% of CRTs did not adjust for clustering (17/54 studies). Furthermore, of CRTs that did perform appropriate analyses, most did not report enough information to abstract adjusted data needed for our planned meta-analytic models [22], or an estimate of the ICC to correct unadjusted data. Reporting of the ICC ranged from 0% for harms to 23% for LDL, with an average of 14% of CRTs within outcomes reporting an estimate of the ICC (Table 7). As a result, many of the unit of analysis errors identified in our sample required correction using an external estimate of the ICC.

Table 7 Outcomes assessed in diabetes QI review

Domain	Outcome	Type of outcome	No. CRTs	No. patients	No. ICC's
Glycemic control	Glycated hemoglobin (mean)	Continuous	54	86,233	9
Cardiovascular management	Systolic blood pressure (mean)	Continuous	40	72,474	7
	Diastolic blood pressure (mean)	Continuous	34	50,060	6
	Low-density lipoprotein (mean)	Continuous	26	62,346	6
	Anti-hypertensive (% use)	Dichotomous	11	26,363	2
	Statin (% use)	Dichotomous	10	23,885	2
	Acetylsalicylic acid (% use)	Dichotomous	10	16,842	2
	Hypertension control (% controlled)	Dichotomous	20	50,007	1
Screening	Retinopathy (% screened)	Dichotomous	27	72,318	2
	Foot (% screened)	Dichotomous	24	42,973	2
	Renal (% screened)	Dichotomous	20	45,751	1
Smoking cessation	Smoking (% smoking)	Dichotomous	14	24,851	1
Harms	Harms (% to experience harm)	Dichotomous	4	2,258	0

5.5 Methods

To address unit of analysis errors of CRTs that did not report an internal estimate of the ICC, we investigated the feasibility of building a database of ICCs and outcome-specific posterior predictive distributions to impute missing ICCs in our review. To do so, we adopted the approach proposed by Turner et al. to: 1) define our target population, setting, and outcomes; 2) search and abstract relevant ICCs to build an ICC database; 3) specify a model of ICC estimates; 4) calculate the variance of ICC estimates; and 5) use abstracted ICCs and their calculated variances to construct an outcome-specific posterior distribution of the ICC. To this approach, we added 6) calculate a posterior predictive distribution of the ICC from which random ICCs can be drawn in our meta-analytic models. Finally, 7) we compared the impact of correcting unit of analysis errors with a single ICC vs. ICCs drawn from the posterior predictive distribution of one continuous outcome.

5.5.1 Defining the target population, cluster and outcomes

Based on eligibility criteria of the diabetes QI review, we defined our target setting to be primary care and our target population to be patients with diabetes. Our target

outcomes included any of the 13 outcomes included in the review. As ICCs representing the exact specification of these targets may be rare (as seen in Table 7), we adopted Turner et al.'s approach of capturing ICC estimates for 'similar' outcomes, cluster types, and populations [79]. As such, we also included estimates from non-diabetic primary care patients for outcomes related to cardiovascular care (e.g., mean systolic blood pressure; proportion of patients with hypertension control).

5.5.2 Searching and abstracting relevant ICCs to build an ICC database

To build an ICC database, we supplemented ICCs identified in the diabetes QI review (Table 7) with those obtained from contacting colleagues with expertise in CRT design and analysis (Dr. Marion Campbell, Dr. Monica Taljaard). All ICCs obtained from additional sources were screened for relevance against our defined target criteria. For relevant ICCs, we abstracted data on the: ICC estimate ($\hat{\rho}_i$), its source (author, year of publication, study design), number of clusters (k_i), number of patients (N_i), study country, clinical setting, and patient population. One investigator (KJD) performed all searches and abstractions for supplemental ICCs.

5.5.3 Specifying a model of ICC estimates

For each outcome, we assumed observed estimates of the ICC, $\hat{\rho}_i$, to be realized values of study-specific ICCs, ρ_i , with a shared underlying logit-transformed normal distribution according to the following:

$$\begin{aligned}\text{logit}(\rho_i) &\sim N(\mu_i, \hat{\sigma}_i), i = 1, \dots, n_{\rho_i} \\ \mu_i &\sim N(d, \tau^2).\end{aligned}$$

Under such a model, underlying ICC values (ρ_i) are assumed to be 'exchangeable', and therefore while not identical, similar enough to believe they do not vary systematically [187]. Each underlying μ_i has a variance, $\hat{\sigma}_i$, calculated from empirical data below, and can be conceived as representing an independent draw from the common random

distribution, $N(d, \tau^2)$, and the objective of the model is to estimate the values of d and τ^2 that define this distribution using a Bayesian framework. As underlying ICCs (ρ_i) are often very small, estimates of their realized values ($\hat{\rho}_i$) may occasionally be negative. In practice, negative ICCs are typically censored at 0 [76]. For these reasons, and to avoid computational challenges, ICC estimates equal to or less than 0 were abstracted and reported in the database but not included in the distribution model.

5.5.4 Calculating the variance of ICC estimates

Estimates of the ICC can vary in precision based on the number of clusters and patients within clusters used to estimate them. In practice, the precision of the ICC is rarely reported. We calculated the empirical variance ($\hat{\sigma}^2$) of ICC estimates using Swiger's approach [188] using abstracted data on the number of patients (N_i) and number of clusters (k_i) according to the following formula:

$$\hat{\sigma}_i^2 = \frac{2(N_i - 1)(1 - \hat{\rho}_i)^2(1 + ((N_i/k_i) - 1)\hat{\rho}_i)^2}{(N_i/k_i)^2 (N_i - k_i)(k_i - 1)}$$

5.5.5 Constructing outcome-specific posterior distribution of the ICC

For outcomes with two or more ICC estimates, we fitted the specified model in a Bayesian framework to estimate the posterior distribution of the ICC, $N(d, \tau^2)$, for each review outcome. As true variances are unknown, we assumed calculated variances could be used in their place as is convention [189]. For all models, we assumed vague priors for the mean (d ; $N(0, 10^3)$) and its standard deviation (τ ; $\text{Uniform}[0, 2]$). Models were fitted with Markov Chain Monte Carlo (MCMC) methods using JAGS software [190] called from R. The first 100,000 model iterations were discarded as burn-in and the following 100,000 iterations were used to obtain the posterior distributions of the d and τ parameters of interest. We assessed parameter convergence using the Brooks-Gelman-

Rubin diagnostic [191] and ensured values of <1.001 for all parameters monitored. Parameter estimates are reported as median and 95% credible intervals (CrIs).

5.5.6 Constructing outcome-specific posterior predictive distribution of the ICC

To incorporate uncertainty of ICC posterior distribution into imputations of the ICC, we used posterior distributions to construct posterior predictive distributions from which missing ICCs could be randomly drawn according to the following:

$$\begin{aligned}\text{logit}(d_{\text{pred}}) &\sim N(d, \tau^2) \\ \text{logit}(\hat{\rho}_{i_{\text{missing}}}) &\sim N(d_{\text{pred}}, \tau^2_{\text{pred}})\end{aligned}$$

Model code for the posterior and posterior predictive distributions is available in Supplemental Text D-1.

5.5.7 Comparison of correcting unit of analysis errors using a single ICC vs. ICCs drawn from the posterior predictive distribution

We compared the impact of correcting unit of analysis errors using a single ICC vs. ICCs drawn from the posterior predictive distribution on meta-analyzed effects of one continuous outcome (HbA1c). We used data from 114 trials (241 arms) that reported mean HbA1c at baseline and follow-up representing data from 48,969 patients included in the most recent published update of the review [55]⁵. We assumed the effect of the QI interventions to be the additive effect of QI strategies (i.e., components) and modeled the average change associated with the presence of each QI strategy using a hierarchical multivariate meta-regression model in a Bayesian framework [192]. The only difference between model implementations in our comparison was the approach used to impute missing ICCs. In the single ICC approach, missing ICCs were set to one value, $\hat{\rho}_{i_{\text{missing}}} = 0.07$, which was used to calculate the design effect and inflate variances for each group mean. In the posterior predictive distribution approach, missing ICCs (on the logit scale)

⁵ Review of operational algorithms for data extraction resulting in small changes to the extracted data. Because of these changes, our results do not perfectly match those reported in 2012.

were randomly drawn from the posterior predictive distribution, $\text{logit}(\hat{\rho}_{i_{missing}}) \sim N(d_{pred}, \tau^2_{pred})$, and after back-transformation (performed in the model), were used to calculate design effects and inflate variances the same as single ICC method. Code for both models is available in Supplemental Text D-2.

5.6 Results

5.6.1 Building the ICC database

We identified 42 ICCs from 12 CRTs included in the diabetes systematic review [80,125,131,134,142–144,149,153,155,165,172]. Twenty-six additional sources were obtained from colleagues (or their references) and screened [77,182,193–216]. Based on our target population, setting, and outcome criteria, we abstracted 17 additional ICC estimates from 6 sources (1 survey [n=5] [214], 1 database [n=1] [203], 1 registry [n=1] [210], and 3 CRTs [n=10] [211–213]) to result in a completed database of 59 ICC estimates (Table 8). ICC estimates varied from 0.000 to 0.129 for continuous outcomes and 0.000 to 0.330 for dichotomous outcomes and exhibited wide range of calculated variances (0.018 to 0.401) due to diverse numbers clusters (6 to 121) and patients (231 to 8808) across studies.

Table 8 Database of ICC estimates meeting target criteria obtained from a systematic review of diabetes quality improvement trials and additional sources

Source*	Reported outcome	Review outcome	Patient population	$\hat{\rho}_i$	N_i	k_i	$\hat{\theta}^2_i$
Parker et al. [211]	DBP (mean)	DBP	General	0.129	1175	16 ⁺	0.136
Parker et al. [211]	LDL (mean)	LDL	General	0.006	1175	19 ⁺	0.112
Parker et al. [211]	SBP (mean)	SBP	General	0.048	1175	16 ⁺	0.136
Parker et al. [211]	Smoking status (%)	Smoking	General	0.036	1175	16 ⁺	0.131
Thompson et al. [203]	Smoking status (%)	Smoking	General	0.118	4893	61	0.034
Elley et al. [212]	DBP (mean)	DBP	General [‡]	0.046	878	42	0.051
Elley et al. [212]	SBP (mean)	SBP	General [‡]	0.018	878	42	0.051
Elley et al. [212]	Smoking (%)	Smoking	General [‡]	0.055	878	42	0.051
Fahey & Peters [210]	Htn-c (%)	Htn-c	Hypertension	0.064	882 [§]	18	0.120
Littenberg et al. [214]	DBP (mean)	DBP	Diabetes	0.017	999	69	0.031
Littenberg et al. [214]	Htn-c (%)	Htn-c	Diabetes	0.069	999	69	0.031
Littenberg et al. [214]	LDL (mean)	LDL	Diabetes	0.045	7834	71	0.029
Littenberg et al. [214]	SBP (mean)	SBP	Diabetes	0.042	999	69	0.031
Littenberg et al. [214]	Creatinine on time (%)	Renal	Diabetes	0.288	8808	74	0.028
Feder et al. [213]	Prescribed beta-blocker (%)	Anti-hyp	CHD	0.060	328	52	0.046
Feder et al. [213]	Prescribed cholesterol lowering drugs (%)	Statin	CHD	0.000	328	52	n/a
Feder et al. [213]	Prescribed aspirin (%)	ASA	CHD	0.000	328	52	n/a
Frijling et al. [155]	Foot exam (%)	Foot	Diabetes	0.330	1449	121	0.018
Frijling et al. [155]	Eye exam (%)	Retinopathy	Diabetes	0.200	1449	121	0.018
Kinmonth et al. [149]	HbA1c (mean)	HbA1c	Diabetes	0.047	231	40	0.062
O'Connor et al. [80]	HbA1c (mean)	HbA1c	Diabetes	0.027	754	12	0.185
O'Connor et al. [80]	SBP (mean)	SBP	Diabetes	0.009	754	12	0.185
O'Connor et al. [80]	Annual foot exam (%)	Foot	Diabetes	0.013	754	12	0.185
O'Connor et al. [80]	Annual kidney function test (%)	Renal	Diabetes	0.019	754	12	0.185
O'Connor et al. [80]	Annual dilated examination (%)	Retinopathy	Diabetes	0.088	754	12	0.185
O'Connor et al. [80]	Aspirin use >3 times/week (%)	ASA	Diabetes	0.000	754	12	n/a
O'Connor et al. [80]	LDL (mean)	LDL	Diabetes	0.000	754	12	n/a
Bebb et al. [165]	DBP (mean)	DBP	Diabetes	0.039	1420	42	0.050
Bebb et al. [165]	SBP (mean)	SBP	Diabetes	0.034	1420	42	0.050
Bebb et al. [165]	Prescribed antihypertensives (%)	Anti-hyp	Diabetes	0.050	1420	42	0.050
MacLean et al. [125]	HbA1c (mean)	HbA1c	Diabetes	0.055	7412	64	0.032
Sönnichsen et al. [131]	DBP (mean)	DBP	Diabetes	0.045	1489	6	0.401
Sönnichsen et al. [131]	HbA1c (mean)	HbA1c	Diabetes	0.003	1489	6	0.401
Sönnichsen et al. [131]	LDL (mean)	LDL	Diabetes	0.046	1489	6	0.401

Sönnichsen et al. [131]	SBP (mean)	SBP	Diabetes	0.054	1489	6	0.401
O'Connor et al. [134]	DBP (mean)	DBP	Diabetes	0.0002	731	11	0.203
O'Connor et al. [134]	HbA1c (mean)	HbA1c	Diabetes	0.0002	1092	11	0.202
O'Connor et al. [134]	LDL (mean)	LDL	Diabetes	0.0002	868	11	0.202
O'Connor et al. [134]	SBP (mean)	SBP	Diabetes	0.0002	894	11	0.202
Shah et al. [172]	Blood pressure $\leq 130/80$ (%)	Htn-c	Diabetes	0.089	1592	80	0.027
Shah et al. [172]	Prescription for ACEI/ARB (%)	Anti-hyp	Diabetes	0.060	1592	80	0.027
Shah et al. [172]	Prescription for statin (%)	Statin	Diabetes	0.123	1592	80	0.027
Steyn et al. [142]	HbA1c (mean)	HbA1c	Diabetes	0.100¶	421	18	0.123
Blackberry et al. [143]	DBP (mean)	DBP	Diabetes	0.088	374	59	0.041
Blackberry et al. [143]	HbA1c (mean)	HbA1c	Diabetes	0.098	440	59	0.040
Blackberry et al. [143]	LDL (mean)	LDL	Diabetes	0.058	366	59	0.041
Blackberry et al. [143]	SBP (mean)	SBP	Diabetes	0.096	375	59	0.041
Blackberry et al. [143]	Current smoker (%)	Smoke	Diabetes	0.043	379	59	0.041
Vaccaro et al. [144]	DBP (mean)	DBP	Diabetes	0.123	1292	9	0.252
Vaccaro et al. [144]	HbA1c (mean)	HbA1c	Diabetes	0.066	1292	9	0.252
Vaccaro et al. [144]	LDL (mean)	LDL	Diabetes	0.016	1292	9	0.252
Vaccaro et al. [144]	SBP (mean)	SBP	Diabetes	0.043	1292	9	0.252
Vaccaro et al. [144]	Any one anti-hypertensive treatment (%)	Anti-hyp	Diabetes	0.021	1292	9	0.252
Vaccaro et al. [144]	Antiplatelet treatment (%)	ASA	Diabetes	0.105	1292	9	0.252
Vaccaro et al. [144]	Statins (%)	Statin	Diabetes	0.046	1292	9	0.252
Frei et al. [153]	DBP (mean)	DBP	Diabetes	0.001**	299	30	0.076
Frei et al. [153]	HbA1c (mean)	HbA1c	Diabetes	0.001**	303	30	0.076
Frei et al. [153]	LDL (mean)	LDL	Diabetes	0.040	300	30	0.076
Frei et al. [153]	SBP (mean)	SBP	Diabetes	0.019	300	30	0.076

$\hat{\rho}_i$ =estimate of intracluster correlation coefficient in the ith study; N_i =number of patients in ith study; k_i =number of clusters in ith study; $\hat{\sigma}_i^2$ =calculated variance for the ith study; ACEI: ace inhibitor; ARB: Angiotensin II receptor blocker; Anti-hyp: Anti-hypertensive treatment; CHD: coronary heart disease; DBP: diastolic blood pressure; HbA1c: glycated haemoglobin; Htn-c: hypertension control; LDL: low-density lipoprotein cholesterol; SBP: systolic blood pressure

*Sourced obtained from Dr. Monica Taljaard included Parker et al., Thompson et al., Littenberg et al. Sources obtained from Dr. Marion Campbell (from the Health Services Research Unit database included Fahey & Peters, Feder et al.

†number of clusters calculated from number of patients and average cluster size

‡defined as “less-active 40-79 year olds”

§number of patients calculated from number of clusters and average cluster size

||reported as “small, with values of ICC <0.0002”

¶based on data from previous local study

**reported as <0.001

5.6.2 Constructing posterior and posterior predictive distributions

After removal of four ICC estimates equal to 0, we had data from 55 ICCs to inform the ICC distributions for 11 of the 13 outcomes (range 2 to 10 per outcome). Although we found three estimates of the ICC for the proportion of patients prescribed aspirin, two of the estimates were equal to 0, and thus no distribution could be modeled for this outcome. Additionally, we found no estimates of the ICC for the proportion of patients to experience harm from a QI intervention. Estimated median and 95% CrI's of the posterior distributions ranged from 0.014 (0.004, 0.050) for mean LDL to 0.151 (0.028, 0.468) for proportion of patients to have retinopathy screening (Table 9). The posterior predictive distributions ranged from 0.014 (0.0004, 0.387) for LDL to 0.153 (0.008, 0.749) for retinopathy screening, and thus while the medians of the distributions were generally equivalent, the CrIs of the posterior predictive distribution were larger.

Table 9 Estimated distributions of the ICC and predicted ICC*

Outcome	No. $\hat{p}_i$	Posterior distribution of the ICC		Posterior predictive distribution of the ICC	
		Median (95% CrI)	Mean (SD)	Median (95% CrI)	Mean (SD)
<i>Intermediate outcomes</i>					
SBP	10	0.020 (0.007, 0.056)	0.023 (0.0129)	0.021 (0.0007, 0.386)	0.059 (0.105)
HbA1c	9	0.015 (0.005, 0.049)	0.018 (0.012)	0.015 (0.0004, 0.387)	0.055 (0.106)
DBP	9	0.020 (0.006, 0.063)	0.024 (0.015)	0.020 (0.0005, 0.457)	0.067 (0.121)
LDL	7	0.014 (0.004, 0.050)	0.017 (0.013)	0.014 (0.0004, 0.343)	0.0489 (0.096)
Smoke	4	0.059 (0.022, 0.139)	0.064 (0.031)	0.059 (0.007, 0.344)	0.083 (0.092)
Htn-c	3	0.076 (0.035, 0.149)	0.080 (0.031)	0.076 (0.017, 0.270)	0.089 (0.072)
<i>Process outcomes</i>					
Anti-hyp	4	0.052 (0.024, 0.088)	0.052 (0.017)	0.052(0.011, 0.173)	0.060 (0.052)
Statin	2	0.088 (0.014, 0.330)	0.104 (0.081)	0.089 (0.004, 0.628)	0.133 (0.153)
Retinopathy	2	0.151(0.028, 0.468)	0.167 (0.109)	0.153 (0.008, 0.749)	0.199 (0.180)
Renal	2	0.087 (0.010, 0.460)	0.124 (0.118)	0.087 (0.002, 0.802)	0.176 (0.213)
Foot	2	0.080 (0.008, 0.458)	0.118 (0.118)	0.080 (0.002, 0.809)	0.173 (0.216)

CrI: credible intervals; DBP: diastolic blood pressure; HbA1c: glycated haemoglobin; Htn-c: hypertension control; LDL: low-density lipoprotein cholesterol; SBP: systolic blood pressure; SD: standard deviation

*distributions for proportion of patients to experience harms or be prescribed aspirin not estimated due to insufficient ICC evidence

5.6.3 Comparison of approaches to correct unit of analysis errors

The posterior predictive distribution for HbA1c outcome had a mean of 0.055 and a standard deviation of (0.106) ($\text{logit}(\rho_i) \sim N(-4.16224, 1.8812)$). Results of the

comparison of the two imputation approaches are presented in Table 10. In this example, correction of unit of analysis errors with ICCs obtained from a posterior predictive distribution did not markedly change regression coefficients for the QI strategies. Contrary to our expectations, the precision of the estimates was somewhat narrower when ICCs were imputed using the posterior predictive distribution.

Table 10 Impact of correction of unit of analysis errors with single vs. posterior predictive ICC distribution in the meta-analysis of a continuous outcome (% HbA1c)

	Single ICC ($\rho = 0.07$) Median (95% CrI)	Posterior predictive distribution of the ICC ($\text{logit}(\rho_i) \sim N(-4.16224, 1.8812)$) Median (95% CrI)
Intercept	8.148 (7.968, 8.329)	8.136 (7.958, 8.316)
CM	0.033 (-0.118, 0.173)	0.022 (-0.116, 0.168)
TC	-0.353 (-0.536, -0.175)	-0.352 (-0.530, -0.173)
EPR	-0.139 (-0.380, 0.090)	-0.143 (-0.369, 0.068)
CE	-0.170 (-0.461, 0.104)	-0.181 (-0.441, 0.076)
FR	-0.245 (-0.443, -0.055)	-0.247 (-0.428, -0.063)
PE	-0.098 (-0.282, 0.084)	-0.091 (-0.295, 0.094)
PSM	-0.185 (-0.387, 0.011)	-0.183 (-0.385, 0.015)
PR	-0.009 (-0.236, 0.225)	-0.006 (-0.233, 0.224)
Other	-0.019 (-0.244, 0.178)	0.0218 (-0.212, 0.187)

CE: clinician education; CM: case management; CrI: credible interval; EPR: electronic patient registry; FR: facilitated relay; HbA1c: glycated hemoglobin; PE: patient education; PR: patient reminders; PSM: promotion of self management; TC: Team changes;

5.7 Discussion

CRTs pose substantial challenges for systematic reviews and meta-analyses, particularly when published reports fail to report cluster-adjusted results or an estimate of the ICC to correct unadjusted results. If a study-specific ICC is not reported, reviewers will often use a single estimate of the ICC obtained from external sources to correct unit of analysis errors. Use of a single ICC does not account for the uncertainty of the ICC estimates. We have demonstrated that construction of an ICC database and posterior predictive distribution to impute missing ICCs is feasible and objectively communicates the uncertainty of ICC estimates. Searching for ICC estimates from additional sources led to obtaining 17 additional ICC estimates and the construction of 11 of the 13 outcome distributions (two of which would have otherwise not been feasible since they had only

one ICC estimate in our review). Furthermore, obtaining additional data on the ICCs improved the precision with which the distributions could be estimated. Although we could readily implement the imputation of missing ICCs from the posterior predictive distribution to correct unit of analysis errors in the hierarchical model, synthesized results of one continuous outcome were not markedly changed by our approach.

We expected the use of the posterior predictive distribution would increase uncertainty of synthesized parameters and were surprised to observe the opposite: increased precision of coefficient estimates. In this case, the single estimate of the ICC was more conservative than its distribution, so ICCs drawn from the distribution, while uncertain, were still smaller and led to a smaller inflation of variance than the single ICC. Had the situation been reversed, and the single ICC was too optimistic and the posterior predictive distribution was higher, ICCs drawn from the posterior predictive distribution would have led to a greater inflation of variance than the single ICC. From our perspective, the approach is useful as it: 1) allow reviewers to guard against over/under-adjusting effect estimates that may not reflect the true underlying ICC and 2) incorporates uncertainty of the ICC estimate into the precision of meta-analysis results. The lack of impact on one outcome however should not be used to determine the utility of this approach, and further work, perhaps using simulation studies are needed to establish whether the application of these methods leads to improved review inferences.

There are several limitations to consider. First, supplemental searches for ICC estimates were not systematic. While it is possible that potentially relevant ICCs were missed, we expect this potential bias to be minimal given that we had already conducted a systematic review for studies meeting our ICC criteria. Indeed, five of the six additional sources identified came from non-diabetic general practice populations that we felt were comparable enough to provide evidence on ICCs of our outcomes of interest (e.g., smoking, blood pressure). The other identified source estimated ICCs in a diabetic population using a survey design and therefore would not have been included in our

systematic review [214]. Second, an outcome-specific estimate of the ICC distribution and a posterior predictive distribution to impute missing ICCs can only be constructed if two or more ICC estimates are identified, with increased precision with more ICC estimates. Given sparse ICC evidence, reviewers may need to consider whether it is more appropriate to impute missing ICCs from a highly imprecise posterior predictive distribution, or base corrections on a single ICC from the most 'similar' of the studies identified. Finally, we did not contact authors of included CRTs for information on missing ICCs that may have identified additional ICC estimates. Future updates of the review will consider contacting authors for this information or obtaining patient-level data so we can obtain an estimate of the ICC ourselves.

We anticipate our posterior predictive distribution of diabetes outcomes could benefit both primary studies seeking to estimate sample sizes and future meta-analyses of these trials. Ideally future CRTs of diabetes QI interventions will report ICCs for all reported outcomes and the distributions estimated in this study could be updated and estimated with greater precision.

5.8 Conclusion

Estimating a posterior predictive distribution of ICCs to impute missing ICCs is a feasible approach for incorporating ICC imprecision into the correction unit of analysis errors and ultimately, synthesized estimates.

Acknowledgements

The authors would like to acknowledge Dr. Issa Dahabreh for his work in developing the protocol for this study and statistical advice during analyses.

Conflict of interest

None declared

The following supplemental information for Manuscript 3 can be found in Appendix D:

Supplemental Text D-1	Model for posterior and posterior predictive distribution
Supplemental Text D-2	Code for hierarchical models implemented in comparison ICC from posterior predictive distribution vs. of single ICC

**CHAPTER 6 [Manuscript 4] Evidence Synthesis For Complex
Interventions Using Hierarchical Regression Models**

6.0 [Manuscript 4] Evidence Synthesis For Complex Interventions Using Hierarchical Regression Models

6.1 Preface to manuscript 4

Study 4 presents a methodological study that illustrated the use of hierarchical multivariate meta-regression for quantitative synthesis when estimating the effects of complex interventions and exploring effect heterogeneity. The original idea for this methodological approach was conceived by Dr. Trikalinos and Dr. Dahabreh as part of the update of the diabetes quality improvement review (Ivers et al, Syst Rev 2014;3:88). I developed the detailed protocol for the study in collaboration with Dr. Dahabreh. I led the re-abstraction (and/or confirmation) of data from the original review. Samir Nazarali, a medical student volunteer, was my second reviewer for the data abstraction. I conducted all analyses, interpreted the results, and drafted the manuscript.

I developed the models to flexibly account for variations in study design and their respective missing variances (or estimate of the intraclass correlation coefficient to correct variances for cluster randomized trials). Thus, the model of one continuous outcome (HbA1c) presented in this manuscript uses the posterior predictive distribution estimated in Study 3 (Manuscript 3; Chapter 5).

Throughout, Dr. Dahabreh provided critical feedback on statistical theory and programming. Dr. Trikalinos, Dr. Ivers, Dr. Grimshaw, Dr. Moher, and Dr. Dahabreh, contributed to the interpretation of the study results and edited the manuscript.

6.2 Title page

Title: Evidence synthesis for complex interventions using hierarchical regression models

Authors: Danko KJ^{1,2}, Dahabreh IJ^{3,4}, Trikalinos TA^{3,4}, Ivers NM^{5,6,7}, Moher D^{1,2,8}, Grimshaw JM^{1,2,8}

Affiliations:

¹Clinical Epidemiology Program, Ottawa Hospital Research Institute, Ottawa, Canada

²School of Epidemiology and Public Health, University of Ottawa, Ottawa, Canada

³Center for Evidence Synthesis in Health, Brown University, Providence, United States

⁴Departments of Health Services, Policy & Practice and Epidemiology, Brown University, Providence, United States

⁵Family Practice Health Centre, Women's College Research Institute, Ontario, Canada

⁶Institute for Health Systems Solutions and Virtual Care, Women's College Hospital, Toronto, Canada

⁷Department of Family and Community Medicine, Dalla Lana School of Public Health, Institute of Health Policy Management and Evaluation, University of Toronto, Toronto, Canada

⁸Department of Medicine, University of Ottawa, Ottawa, Canada

Corresponding author:

Kristin Julianna Danko

The Ottawa Hospital – General Campus, Box 711

501 Smyth Road

Ottawa, Ontario, Canada

K1H 8L6

Email: kdanko@ohri.ca

6.3 Abstract

Background: When synthesizing trial evidence about the effects of complex multicomponent interventions, standard meta-analysis methods reduce the data into simple pairwise comparisons (e.g., any combination of components vs. no component) resulting in substantial loss of information and inability to model the association of specific intervention components with the post-treatment outcome mean.

Methods: Our work is motivated by a systematic review of 142 randomized trials of quality improvement interventions involving at least one of 12 component strategies. We specified a series of hierarchical models to assess the association of specific component strategies with the post-treatment outcome mean and compared the results to standard meta-analysis approaches.

Results: Compared to standard methods, the hierarchical modeling approach better reflects the associations of individual components with the post-treatment mean. We show how the models can leverage different assumptions about intervention groupings, possible interactions between components, and effect modification by study-level covariates, to more flexibly model the relationship between components and the post-treatment mean. The modeling strategy also allows us to address particular aspects of the available data such as the large number of studies missing estimates of the sampling variance and the large number of cluster trials with missing estimates of the intraclass correlation coefficient.

Conclusion: Hierarchical modeling is a feasible and potentially more informative approach to synthesizing evidence from clinical trials of complex multi-component interventions. Evidence syntheses using these methods can be used to predict the impact of new complex interventions programs in new settings or support the rational design of future studies.

Keywords: complex interventions, multicomponent interventions, meta-analysis, meta-regression, hierarchical models

6.4 Background

Interventions designed to change healthcare practice and health policy are often ‘complex’, involving multiple components, targeted to various individuals, across dynamic and diverse settings [48]. Although researchers continue to debate the definition of complex intervention [46,47,217], most agree that multiple intervention ‘components’ that may or may not interact with each other, and effect modification by study-specific characteristics (e.g., healthcare setting, patient population) are central features of complex interventions [11].

Decision-makers often want to learn about a range of related complex interventions to understand which components (or combinations of components) work best and under what conditions [86,218]. Systematic reviewers are hard pressed to address such questions about complex interventions with conventional meta-analytic methods that are best applied to synthesis of evidence from a single intervention assessed in a homogenous group of studies. Several authors have proposed the use of hierarchical regression models as an alternative synthesis approach for complex interventions [23,86,88,89].

In this paper, we illustrate the use of hierarchical regression models to synthesize evidence from multiple studies of complex interventions, including estimating the association between individual components and the post-treatment mean and the influence of potential effect modifiers, using data from studies included in a systematic review of diabetes quality improvement interventions. We first introduce the research questions and data from the motivating example (Section 6.5) and discuss limitations of standard methods of meta-analysis for the synthesis of complex interventions (Section 6.6). We then present our approach using hierarchical regression models and discuss issues related to estimation, prediction, and ranking (Section 6.7). We consider extensions to the model that allow for assessment of interactions among intervention components and modification of the components’ effects by study-level covariates and

handling discrete outcomes and missing data, including missing estimates of the intraclass correlation coefficient (ICC) to adjust variance data in cluster RCTs (Section 6.8). We apply the hierarchical model to the motivating example to estimate the effects of intervention components and demonstrate inferential capabilities of ranking and prediction and assessment of interactions and effect modification (Section 6.9). Lastly, we discuss the strengths and limitations of our approach and areas for future research (Section 6.10).

6.5 Motivating example: quality improvement interventions to improve diabetes care

The International Diabetes Federation estimates that 415 million adults were living with diabetes in 2015 and predicts that the prevalence will exceed 640 million by 2040 [219]. People living with diabetes are at increased risk for serious complications (e.g., cardiovascular events, blindness). Despite strong evidence that medical interventions, such as monitoring glycemic control, monitoring micro-vascular complications, and vascular risk factor management improve patient outcomes and reduce costs [50–52], many patients with diabetes do not receive evidence-based care and have sub-optimal control of risk factors [53,54,220]. Quality improvement (QI) interventions seek to address these evidence-to-practice gaps by combining multiple intervention components that target system, provider, or patient factors influencing diabetes care [55].

In 2012, three of us (JMG, NMI, DM) conducted a systematic review of 142 trials to assess the effect of QI interventions across a range of process (e.g., foot screening) and intermediate patient (e.g., glycemic control) outcomes [55]. We coded QI interventions using a taxonomy of 12 components, adapted from Cochrane’s Effective Practice and Organization of Care (EPOC) 2002 taxonomy (Table 11) [55–57]. The majority of trials included in the review evaluated a complex QI intervention with a median of 3 QI components per intervention arm (range 1-8). The review estimated the mean effect of QI intervention a whole (i.e., one or more components versus usual care) on each

observed outcome (e.g., mean difference in HbA1c, a measure of glycemic control) and found QI interventions are effective but effects were highly heterogeneous across studies. For example, based on the random effects meta-analysis of 120 HbA1c studies with a median follow-up of 12 months, QI interventions were associated mean reduction of 0.37% HbA1c (95% CI -0.45 to -0.28; $I^2 = 73\%$). Subgroup analyses assessing the efficacy of QI interventions containing a specific component of interest (e.g., case management) compared to QI interventions not containing that component (e.g., no case management) found improvements associated with most QI components but conflated the effects of co-occurring components. Finally, the meta-analyses did not assess the interactions among intervention components or modification by study-level covariates of the association between components and the outcome. Thus, despite the large number of studies, the review could not explain the observed heterogeneity or provide information about the expected effect of combinations of components were they to be implemented in a new patient population or setting.

Table 11 Taxonomy of Quality Improvement (QI) components

QI component	Definition
Audit and feedback	Summary of clinical performance of health care delivered by an individual clinician or clinic over a specified period, transmitted back to the clinician
Case management	Any system for coordinating diagnosis, treatment, or routine management of patients by a person or multidisciplinary team in collaboration with, or supplementary to, the primary care clinician
Team changes	Changes to the structure or organization of the primary health care team including adding a team member or shared care, use of multidisciplinary teams, or expansion or revision of professional roles
Electronic patient registry	General electronic medical record system or electronic tracking system for patients with condition
Clinician education	Interventions designed to promote increased understanding of principles guiding clinical care or awareness of specific recommendations for a target disorder or population of patients
Clinician reminders	Paper-based or electronic systems intended to prompt a health professional to recall patient-specific information
Facilitated relay of clinical information	Clinical information collected from patients and transmitted to clinicians by means other than the existing medical record
Patient education	Interventions designed to promote greater understanding of a target disorder or to teach specific prevention or treatment strategies, or specific in-person education
Promotion of self management	Provision of equipment or access to resources to promote self-management
Patient	Any effort to remind patients about upcoming appointments or important aspects of

reminders	self care
Continuous quality improvement	Interventions explicitly identified as involving the techniques of continuous QI, total quality management, or plan-do-study-act, or any iterative process for assessing quality problems, developing solutions to those problems, testing their effects, and then reassessing the need for further action
Financial incentives	Interventions with positive or negative financial incentives directed at providers, patients or system-wide changes in reimbursement

6.6 Challenges in meta-analyses for complex interventions

The fundamental challenge to synthesizing evidence from complex multicomponent interventions is sparsity of evidence. Consider the challenge of obtaining sufficient evidence on m intervention components and their combinations alone. Suppose for example, that a decision-maker is interested in adopting a complex intervention comprised of a subset of the available components. Ideally, each unique version of the complex intervention would have been evaluated in one trial, and for the purpose of meta-analysis, in several trials. Better still, each combination would be evaluated in a reasonably diverse range of settings and populations such that a synthesis of their results would reflect the consistency of the effect across different contexts. Under such a scenario, a synthesis of each homogeneously-defined version of a complex intervention (either as separate reviews, or separate analyses within the same review) would be the most appropriate and rigorous meta-analytic approach to inform decision-makers of their relative effects [7]. As the number of components of interest grows, however, the number of possible versions of the intervention grows exponentially: for m components there are 2^m possible complex interventions. In the motivating example of the diabetes QI review we considered 12 QI components, which means there are 4,096 possible QI interventions, most of which will never be evaluated in a trial. Indeed, for the reviews' primary outcome, HbA1c, assessed in 119 trials, the review found evidence for only 83 unique combinations of QI interventions, 66% of which (55/83) were evaluated in only one trial, leaving >4,000 possible complex interventions untested (Figure 9).

To overcome the issue of data sparsity, many reviews (including the diabetes QI review) aggregate versions of a complex intervention in a single meta-analysis and seek to

investigate the effects of intervention components indirectly (e.g., using subgroup analyses). The overall meta-analysis, for example, attempts to answer the question “Do complex interventions as a whole lead to a difference in observed effects?” However, estimates from such an analysis cannot be used by decision-makers to identify combinations of components that are associated with the greatest improvement in the outcome in their target population or setting. Although subgroup analyses and meta-regression are commonly used to explore effect variation due to intervention components, most reviews of complex interventions do not have enough studies to support such analyses beyond one or two components [8,87] and as such may produce misleading results as the presence of co-occurring components would not be controlled for [87].

Another issue in evidence synthesis for complex interventions is that many analysts use statistical approaches that cannot handle multi-arm studies and studies with two or more active-treatment groups, even though trials of complex interventions often compare different (active) versions of complex interventions, sometimes across multiple arms [8,47]. To support the pairwise structure of standard meta-analytic models, many analysts use only two arms from each study (i.e., complex interventions with most versus least number of components), resulting in substantial information loss as well as comparisons that do not have a clear interpretation (e.g., the “most” intense complex intervention in one study can be the least intense intervention in another study).

6.7 Specification of the basic hierarchical regression model, prior specification, and inference

6.7.1 Specification of the basic model

In this section, we propose a meta-analytic modeling approach based on hierarchical multivariable regression that aims to address the important issues of heterogeneity in evidence synthesis of complex interventions by exploring the association of intervention components and study- or arm-level effect modifiers with the post-treatment means [23,89,221]. The approach permits the inclusion of all arms from all studies and allows the association of intervention components with outcomes to vary based on the presence or absence of other components or baseline covariates.

Let Y_{ij} be the post-treatment mean in the j^{th} arm of the i^{th} study, distributed as

$$Y_{ij} \sim N(\mu_{ij}, \theta^2_{ij}), i = 1, \dots, n; j = 1, \dots, n_i; \quad (1)$$

The number of arms in the i^{th} study, n_i , is not restricted, allowing us to model multi-arm trials. Given the large number of components and interactions between them, we will usually have to assume a parsimonious model for μ_{ij} . We begin by considering a linear additive model in terms of the m components:

$$\mu_{ij} = \beta_{0i} + \sum_{k=1}^m \beta_{ki} X_{ijk}, \quad (2)$$

where X_{ijk} denotes the value of the k^{th} component (coded as 0 if absent or 1 if present) in the j^{th} arm of the i^{th} study; henceforth, we refer to the coefficients β_{ki} as “mean differences” because they express the difference in the post-treatment mean when component X_{ijk} is present or absent. The intercept, β_{0i} , represents the post-treatment mean in the absence of intervention [81].

Next, we assume study-specific estimates of component coefficients, $\beta_{ki}, \dots, \beta_{mi}$, to be realizations of underlying random variables from a normal distribution:

$$\beta_{ki} \sim N(\tilde{\beta}_k, \tilde{\tau}_k^2), k = 1, \dots, m, \quad (3)$$

where $\tilde{\beta}_k$ and $\tilde{\tau}_k$ represent the mean and standard deviation of the distribution, respectively, and are to be estimated from the model. Study-intercepts, β_{0i} , are also assumed to be random variables with an underlying normal distribution.

$$\beta_{0i} \sim N(\tilde{\beta}_0, \tilde{\tau}_0^2) \quad (4)$$

We feel the assumption of exchangeable arm-specific parameters is justified in trials of complex interventions where studies commonly evaluate active arms, the distinction between intervention and control arms is ill-defined, and it is common for intervention components to be assessed in all arms. In the case of the diabetes QI review, this corresponds to a single $\tilde{\beta}_0$ coefficient estimating the post-intervention mean in the absence of a QI intervention and 12 $\tilde{\beta}_k$ coefficients estimating the mean change in post-intervention mean associated with each QI strategy. The number of β -coefficients to be estimated in another application of this approach would vary according to expert input and selected framework used to characterize intervention content [21,222,223].

6.7.2 Prior specification

Some features of our approach, such as the incorporation of external information and handling of missing data (see Section 6.8) are most naturally achieved in the Bayesian framework [187,189]. Furthermore, Bayesian hierarchical modeling more fully reflects uncertainty about the model and offers a natural approach for evidence synthesis because the study similarity judgments that systematic reviewers make relate to Bayesian exchangeability assumptions. For these reasons, we opted to use a Bayesian approach to estimate model parameters. To do so, we began by specifying weakly

informative prior distributions for $(\tilde{\beta}_k, \tilde{\tau}_k)$, $k = 0, \dots, m$. Specifically, we used: $\tilde{\beta}_0 \sim N(8, 100)$; $\tilde{\tau}_0 = U(0, 2)$; $\tilde{\beta}_k \sim N(0, 4)$; $\tilde{\tau}_k = U(0, 2)$. When more extensive background knowledge is available about the associations of different components with the post-treatment mean, more informative prior distributions can increase the precision of the results; this may be particularly appealing for the heterogeneity parameters, $\tilde{\tau}_k$.

6.7.3 Inference

We can use the hierarchical regression model described in the previous subsection to: (1) estimate the posterior distribution of the mean difference for each intervention component (and use that posterior distribution for inference); (2) rank the components by the magnitude of the risk differences, for example, to visualize the probability that the mean difference for a component has certain rank (i.e., probability of having greatest mean difference, second greatest, and so on among the components included in the model) [224]; and (3) predict the post-treatment mean in future studies, possibly using combinations of components that have not been previously assessed in trials. Predictive inference can be obtained by examining the posterior predictive distribution for the post-treatment mean for any combination of components:

$$\mu_{new} = \beta_{0_{new}} + \sum_{k=1}^m \beta_{k_{new}} X_{k_{new}} \quad (5)$$

$$\beta_{k_{new}} \sim N(\tilde{\beta}_k, \tilde{\tau}_{\tilde{\beta}_k}^2), k = 0, \dots, m, \quad (6)$$

where $X_{k_{new}}$ denotes the k th component in the new study. The posterior predictive distribution of μ_{new} can be used when designing a new study, because planning decisions can be based on the posterior predictive distribution for a particular study,

$$Y_{new} \sim N(\mu_{new}, \tau_{new}^2), \quad (7)$$

where θ^2_{new} denotes the sampling variance of the new study (which would depend on its sample size).

6.8 Model extensions

In this section we extend the hierarchical regression model to relax the additivity assumptions of Equation (2); handle missing data; and examine discrete outcomes. We use the term *interaction* to describe the differential effect of one intervention component depending on the presence or absence of another intervention component; we use the term *effect modification* to describe the differential effect of one intervention component over levels of a baseline study-covariate (e.g., setting of care, baseline outcome risk, population age).

6.8.1 Interactions among intervention components

We can assess interactions by including product terms among intervention components. For example, to allow for interaction component, r , and all other intervention components, we might specify the following mean model:

$$\mu_{ij} = \beta_{0i} + \sum_{k=1}^m \beta_{ki} X_{ijk} + \sum_{l \neq r} \gamma_{li} X_{ijl} * X_{ijr}. \quad (8)$$

We assume that the interaction coefficients, γ_{li} , are distributed as

$$\gamma_{li} \sim N(\tilde{\gamma}_l, \tilde{\tau}_l^2), l \neq r. \quad (9)$$

Of course, in Bayesian analyses, prior distributions need to be specified for as well. We used the minimally informative distributions $\tilde{\gamma}_l \sim N(0,2)$ and $\tilde{\tau}_l = U(0,2)$.

In most applications, we do not have data to assess all hypothesized interactions or effect modifiers of interest. For example, for m intervention components, there are $m(m - 1)/2$ possible pairwise interactions. In the case of the diabetes QI review, this equates to 66 pairwise interactions formed from 12 QI components. Available studies are unlikely to provide sufficient data to support a fully saturated model (i.e., one that would include all possible interactions) – such a model would have enough coefficients to (non-parametrically) model 4,096 QI combinations, which is equivalent to a network meta-analysis where each possible combination of components is modeled as a separate treatment node. Clearly, most applied meta-analyses of complex interventions cannot identify that many parameters.

6.8.2 Effect modification by study covariates

The model can also be extended to assess effect modification by study-level covariates, such as study setting or other contextual factors, with the addition of appropriate product terms. For example, let Z_i denote an indicator for whether the study was conducted in a for-profit or not-for-profit setting; then, we might use the post-treatment mean model

$$\mu_{ij} = \beta_{0i} + \sum_{k=1}^m \beta_{ki} X_{ijk} + \phi_i Z_i + \sum_{k=1}^m \psi_{ki} Z_i * X_{ijk} \quad (10)$$

where ϕ_i is the “main effect” of the modifier and ψ_{ki} is the “statistical” interaction coefficient for each k th intervention component. We assume the ϕ_i are distributed as

$$\phi_i \sim N(\tilde{\phi}, \tilde{\tau}_{\tilde{\phi}}^2) \quad (11)$$

and the ψ_i are distributed as

(12)

$$\psi_{ki} \sim N(\tilde{\psi}_k, \tau_{\tilde{\psi}_k}^2)$$

Again, in Bayesian analyses, prior distributions need to be specified, and we chose the minimally informative distributions for both $(\tilde{\phi} \sim N(0,4)$ and $\tilde{\tau}_{\tilde{\phi}} = U(0,2)$; $\tilde{\psi}_k \sim N(0,4)$ and $\tau_{\tilde{\psi}_k} = U(0,2)$.

As usual, to obtain reasonable results, meta-analysts will need to make decisions as to what interactions and effect modifiers to pursue, within the limitations of available data [23]. Depending on number of intervention components and effect modifiers of interest, reviewers may consider model reduction strategies such as aggregating intervention components into higher order categories based on substantive knowledge and/or expert opinion or grouping uncommon interventions into a single ‘mixed/other’ category. In a Bayesian framework, reviewers can rely on other meta-analyses of similar interventions or expert opinion [187] to specify prior distributions to improve estimation with sparse data.

6.8.3 Discrete outcomes

The model can be modified to assess the impact of complex interventions on non-linear and non-continuous outcomes (e.g., binary, count) by replacing Equations 1 and 2 with the appropriate link functions and likelihood for the response variable, respectively [4]. For example, for binary response data (e.g., proportion of patients on anti-hypertensive medication), equation (1) is modified such that the count of events in the j th arm of the i th study is a realization of Y_{ij} , a binomial random variable,

$$Y_{ij} \sim \text{Bin}(n_{ij}, p_{ij}), \tag{12}$$

where n_{ij} is the number of individuals and p_{ij} is the probability of event occurrence in the j th arm of the i th study. A common choice for the mean model is a logistic-linear specification,

$$\text{logit}(p_{ij}) = \beta_{0i} + \sum_{k=1}^m \beta_{ki} X_{ijk}. \quad (13)$$

6.8.4 Missing data

Data are often missing in meta-analyses and missingness has to be addressed for valid evidence synthesis [5]. A hierarchical model implemented in a Bayesian framework provides a convenient structure to impute missing data. In our motivating example, the standard error was missing from 68 arms of patient randomized trials and 31 cluster-randomized trials. As many cluster trials did not adequately account for the clustering effect (or report adjusted estimates or an estimate of the ICC to correct unadjusted estimates), we also need to impute an estimate of the ICC for 49 standard errors of the post-intervention mean that required adjustment. Details on this process are provided in Supplemental Text E-1. Briefly, using expert opinion combined with information from the QI intervention review, we imputed missing standard errors from a moderately informative uniform distribution $U(0,2)$. We imputed missing ICCs from a posterior predictive distribution obtained from the synthesis of ICCs in the diabetes QI review and external sources. For example, based on 9 estimates of the ICC for the HbA1c outcome, we obtained a posterior predictive distribution with a logit-transformed mean of -4.16224 and standard deviation of 1.88812.

6.9 Application to the motivating example

To compare our approach to conventional methods of meta-analysis, we implemented a series of hierarchical models to isolate the effects of the QI components and compared them to estimates from standard meta-analysis models. Our applications used data from 114 trials (241 arms) that reported mean HbA1c at baseline and follow-up representing data from 48,969 patients. Studies with missing variances of the post mean for the study arm (or unadjusted variances for cluster-RCTs) were imputed, as demonstrated in Supplemental Text E-1. To facilitate modeling, we (after seeking expert input) opted to

combine QI strategies that were observed infrequently (defined as QI strategies present <10% of arms) in one combined 'Other' category. This category included clinician reminders (9.5%), audit and feedback (8.7%), financial incentives (0.8%), and continuous quality improvement (0.4%). We must also note, that in the process of updating the original systematic review [22], we reviewed all our operational algorithms for data extraction resulting in small changes to the extracted data. Because of these changes, our results do not perfectly match those reported in 2012. The results for the HbA1c outcome reported here use the corrected data from the original set of publications and are intended exclusively for the purpose of illustrating the methods and should not be used for clinical or policy decisions.

The conventional meta-analysis model was fit in R (overall effect and subgroup analyses for each QI strategy) using the meta package [225]; estimates are reported as mean change and 95% confidence intervals (CIs). Hierarchical models were fit using Markov chain Monte Carlo (MCMC) methods with the software JAGS [190] called from R with 100,000 iterations for burn-in and 100,000 iterations to obtain the posterior distribution of parameters of interest. Estimates are reported as median change and 95% credible intervals (CrIs). We used the Brooks-Gelman-Rubin diagnostic to assess parameter convergence [191]. Estimates from hierarchical models are reported as median change and 95% credible intervals.

6.9.1 Comparison of hierarchical meta-regression and pairwise synthesis models

To compare the use of hierarchical models for estimating the specific effect of QI components, we compared parameter estimates from 3 analyses:

- *Analysis 1* used pairwise data (228 arms; 44,375 individuals) and estimated component effects through individual subgroup analyses for each QI component (i.e., the average mean change of studies that included the QI component)

- *Analysis 2* used the same pairwise data as *Analysis 1* but estimated the effect of QI components using the base hierarchical model defined in Equations 1 and 2.
- *Analysis 3* estimated the effect of QI components using the base hierarchical model but included complete data from all available arms (241 arms; 48,969 individuals).

Missing data in *Analysis 1* was handled using conventional methods used in the 2012 review and occurred prior to meta-analysis. Specifically, we used a single value to impute missing standard deviations ([SD]; SD=2.22 representing the 99th percentile of observed data) and a single value to impute missing ICCs (ICC=0.027, obtained from one study in the sample) to correct standard errors from cluster trials that were not appropriately adjusted for the clustering effect [64]. Sensitivity analyses using a less conservative SD (median=1.34) and higher ICC used in the 2012 review (ICC=0.07) did not change the overall mean and precision of the random effects meta-analysis. Missing data in *Analyses 2* and *3* were handled by drawing random values for the standard error and ICC from specified distributions for required observations and was embedded within the model. Details on how trials were uniquely treated by the model, depending on study design and type of missing data, are reported in Supplemental Table E-1. Code for the hierarchical models, that essentially implements the model with different imputation strategies for different subsets of studies depending on their patterns of missingness, appears in Supplemental Text E-2 and Supplemental Text E-3.

We present results from the three analyses in Table 12. Compared to the standard model, QI component effects are much smaller when estimated from the hierarchical meta-regression model (*Analysis 2* vs. *1*). Point estimates of parameters from the hierarchical models were similar (*Analysis 3* vs. *2*), however estimates were slightly more precise in *Analysis 3*, as would be expected from inclusion of additional data. Provided that the hierarchical model is approximately correctly specified, we interpret the reduction in effect sizes when estimated by the hierarchical meta-regression model to be

the better isolation of the expected post-treatment mean reduction associated with unique QI components, controlling for co-occurring components.

Table 12 Summary of results comparing standard and hierarchical regression model

	Analysis 1	Analysis 2*	Analysis 3*
CM	-0.42 (-0.55, -0.29)	0.02 (-0.13, 0.17)	0.02 (-0.12, 0.17)
TC	-0.53 (-0.69, -0.37)	-0.38 (-0.56, -0.20)	-0.35 (-0.53, -0.17)
EPR	-0.37 (-0.53, -0.22)	-0.16 (-0.41, 0.05)	-0.14 (-0.37, 0.07)
CE	-0.23 (-0.37, -0.09)	-0.18 (-0.48, 0.09)	-0.19 (-0.45, 0.06)
FR	-0.40 (-0.54, -0.26)	-0.22 (-0.42, -0.03)	-0.24 (-0.44, -0.06)
PE	-0.44 (-0.56, -0.32)	-0.04 (-0.25, 0.14)	-0.09 (-0.27, 0.09)
PSM	-0.41 (-0.52, -0.30)	-0.20 (-0.41, -0.01)	-0.18 (-0.37, 0.01)
PR	-0.33 (-0.53, -0.14)	0.04 (-0.20, 0.30)	-0.01 (-0.23, 0.21)
Other†	-0.19 (-0.31, -0.06)	-0.00 (-0.24, 0.21)	0.03 (-0.20, 0.18)

Numbers are mean differences (95% CI) for Analysis 1 and mean differences (95% CrI) for Analyses 2 and 3.

CM: case management; CE: clinician education; EPR: electronic patient registry; FR: facilitated relay; TC: team change; PE: patient education; PR: patient reminders; PSM: promotion of self-management

*We used the following prior distributions in the Bayesian analyses: $\tilde{\beta}_0 \sim N(8, 100)$; $\tilde{\tau}_0 = U(0, 2)$;

$\tilde{\beta}_k \sim N(0, 4)$; $\tau_{\tilde{\beta}_k} = U(0, 2)$.

†Other: Combined category for infrequently evaluated component including audit and feedback, clinician reminders, financial incentives and continuous quality improvement

6.9.2 Ranking of intervention components

An example rankogram that can be produced from estimates of a hierarchical meta-regression model is presented in Figure 10. Using the output of Analysis 3 above, the rankogram indicates the probability of each component being the best, second best, and so on among the modeled components with respect to post intervention mean change. Based on this example, the team changes component has a higher probability of ranking as one of the top three QI strategies, while case management appears to rank in the bottom three.

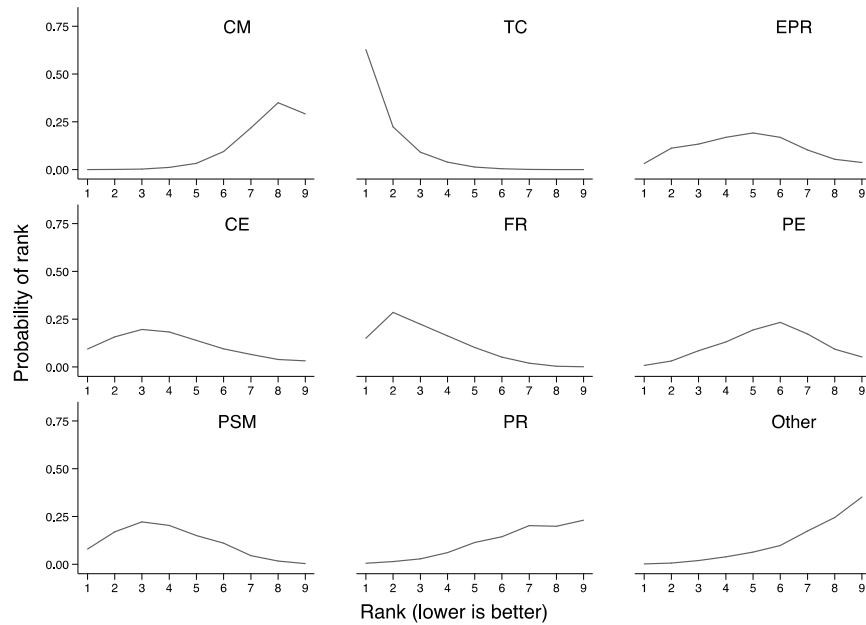


Figure 10 Ranking of QI strategy components

CM: case management; CE: clinician education; EPR: electronic patient registry; FR: facilitated relay; TC: team change; PE: patient education; PR: patient reminders; PSM: promotion of self management

6.9.3 Assessing interactions among components

To demonstrate the assessment of interactions among intervention components, we extended the model in Equations 1 and 2 to Equation 8 to assess pairwise interactions (i.e., non-additive effects) among QI strategies. A series of models were fit that sequentially interacted a single QI strategy, r , with all remaining QI strategies, $l = 1, \dots, q$, corresponding to 8 additional parameters estimated in each of the 9 additional models (Supplemental Text E-4). We present results from the pairwise interaction models in Figure 11. Point estimates of interaction parameters were almost always smaller than the main effects of QI components and had wider uncertainty; thus we did not observe any strong evidence in support of pairwise QI*QI interaction. We therefore opted to maintain our additive assumption of intervention effects in favour of the more parsimonious model of component effects presented in Equation 1.

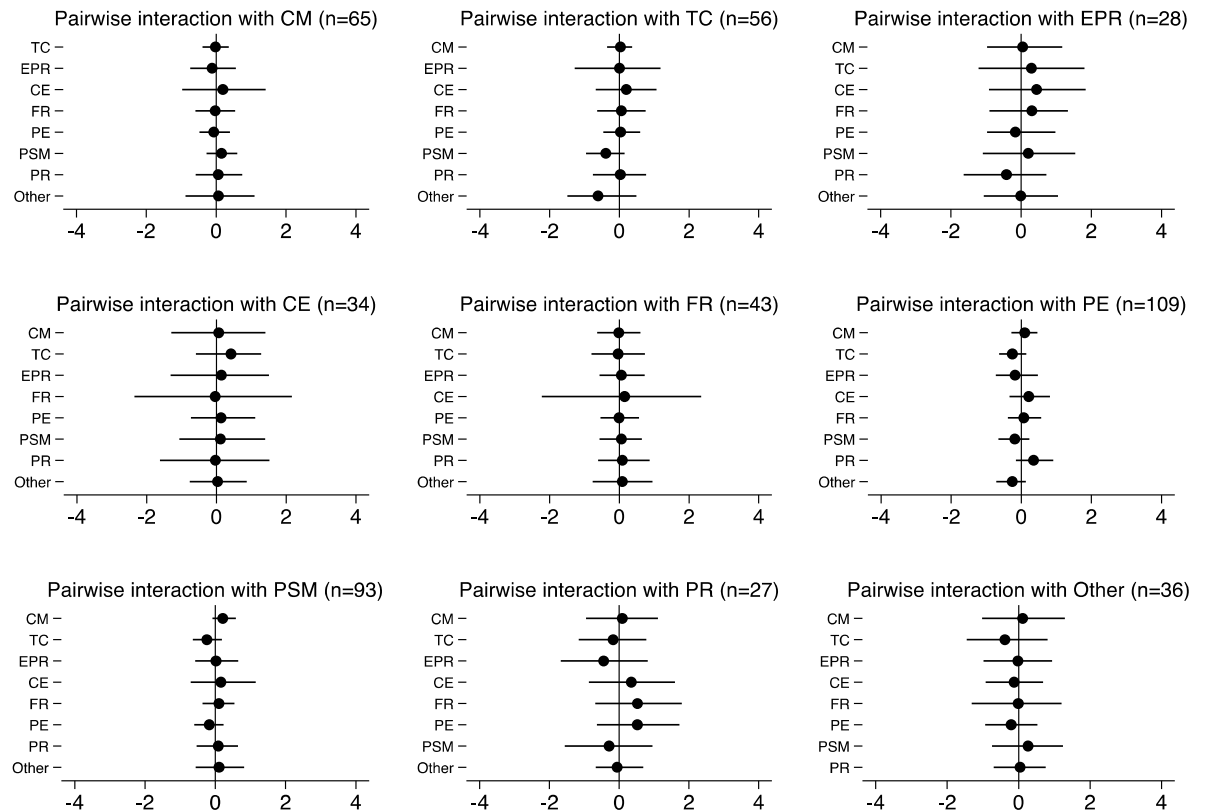


Figure 11 Coefficient estimates of pairwise interactions between QI strategy components*

CM: case management; CE: clinician education; EPR: electronic patient registry; FR: facilitated relay; TC: team change; PE: patient education; PR: patient reminders; PSM: promotion of self-management

*The following priors were used in the Bayesian analyses to estimate parameters in the interaction models:

$\tilde{\beta}_0 \sim N(8, 100)$; $\tilde{\beta}_k \sim N(0, 4)$; $\tilde{\gamma}_l \sim N(0, 2)$; all $\tau = U(0, 2)$. For each model, the reported n indicates the number of observations (arms) for the component of interest (i.e., 65 arms in which case management [CM] was present).

6.9.4 Assessing effect modification by covariates

As the original QI review identified baseline risk as a potential effect modifier of intervention effectiveness, we explored the interaction of baseline HbA1c (coded as a binary and continuous covariate) with each of the QI components as per Equation 10. In the binary model we used glycemic control of 8.0% to delineate between trials with patient populations that were ‘controlled’ and ‘uncontrolled’ at baseline. In the continuous model we calculated a mean centered HbA1c covariate. Both baseline HbA1c

models included an additional 10 interaction parameters (i.e., main effect of baseline HbA1c plus 9 QI * baseline HbA1c interactions; Supplemental Text E-5 and E-6). Although interactions were not significant, baseline risk did appear to modify the effect of some QI strategies (Figure 12) and improve model specification, particularly when treated as a continuous variable (see Section 6.9.6 below). For example, the impact of team changes on post-intervention mean change appears to be slightly greater when delivered in populations with higher baseline risk (Table 13). Compared to the base model however, point estimates for effect modification parameters were all substantially smaller than the main effects and had greater uncertainty. As effect modification with baseline HbA1c did not suggest systematic deviation in effects, we continued to prefer our base model.

Table 13 Effect of QI strategy in arms with controlled vs. uncontrolled HbA1c at baseline*

QI strategy	Uncontrolled† + QI strategy§	Difference in uncontrolled	Controlled‡ + QI strategy¶	Difference in controlled	Difference of difference
CM	8.45 (8.08, 8.83)	-0.18 (-0.48, 0.11)	7.54 (7.26, 7.82)	0.14 (-0.07, 0.35)	0.33 (-0.04, 0.68)
TC	8.35 (7.98, 8.72)	-0.29 (-0.58, 0.01)	7.15 (6.82, 7.47)	-0.25 (-0.54, 0.03)	0.03 (-0.38, 0.44)
EPR	8.63 (7.88, 9.40)	-0.01 (-0.72, 0.73)	7.34 (6.99, 7.72)	-0.06 (-0.37, 0.29)	-0.05 (-0.86, 0.71)
CE	8.53 (7.83, 9.23)	-0.11 (-0.78, 0.57)	7.29 (7.00, 7.60)	-0.11 (-0.39, 0.20)	-0.01 (-0.72, 0.71)
FR	8.42 (8.07, 8.79)	-0.21 (-0.50, 0.10)	7.13 (6.76, 7.51)	-0.27 (-0.60, 0.09)	-0.06 (-0.51, 0.41)
PE	8.55 (8.20, 8.89)	-0.09 (-0.41, 0.23)	7.51 (7.23, 7.79)	0.11 (-0.17, 0.38)	0.20 (-0.25, 0.64)
PSM	8.46 (8.09, 8.82)	-0.17 (-0.50, 0.13)	7.08 (6.73, 7.42)	-0.31 (-0.64, -0.02)	-0.14 (-0.60, 0.30)
PR	8.75 (8.13, 9.40)	0.11 (-0.46, 0.73)	7.17 (6.73, 7.58)	-0.22 (-0.66, 0.13)	-0.34 (-1.10, 0.31)
Other	8.46 (7.75, 9.23)	-0.17 (-0.84, 0.53)	7.43 (7.10, 7.68)	0.03 (-0.25, 0.22)	0.19 (-0.56, 0.89)

CM: case management; CE: clinician education; EPR: electronic patient registry; FR: facilitated relay; TC: team change;

PE: patient education; PR: patient reminders; PSM: promotion of self-management

*The following priors were used in the Bayesian analyses to estimate parameters in the effect modification models:

$\tilde{\beta}_0 \sim N(8, 100)$; $\tilde{\beta}_k \sim N(0, 4)$; $\tilde{\phi} \sim N(0, 4)$; $\tilde{\psi}_k \sim N(0, 4)$; all $\tau = U(0, 2)$.

†post-intervention mean change in patients with uncontrolled HbA1c who did not receive QI strategy, $\beta_0 + \phi$, 8.63 (95% CrI [8.41, 8.86])

‡post-intervention mean change in patients with controlled baseline HbA1c who did not receive QI strategy, β_0 , 7.40 (95% CrI [7.24, 7.57])

§ $\beta_0 + \beta_k + \phi + \psi_k$

|| $\beta_k + \psi_k$

¶ $\beta_0 + \beta_k$

** β_k

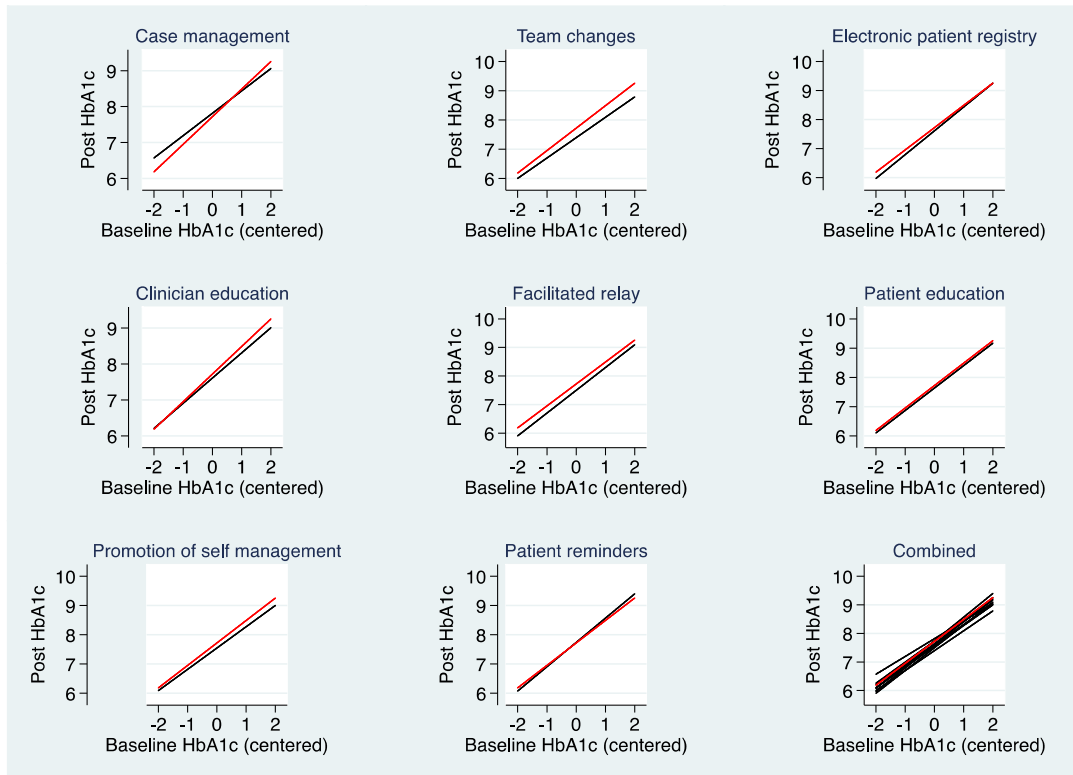


Figure 12. Effect modification by baseline HbA1c

Red line: post-intervention effect in arms that did not receive the QI strategy; black: post-intervention effect in arms that did receive the QI strategy

*The following priors were used in the Bayesian analyses to estimate parameters in the effect modification models: $\tilde{\beta}_0 \sim N(8, 100)$; $\tilde{\beta}_k \sim N(0, 4)$; $\tilde{\phi} \sim N(0, 4)$; $\tilde{\psi}_k \sim N(0, 4)$; all $\tau = U(0, 2)$.

6.9.5 Predicting the effect of a new study

Given that the vast number of combinations of QI components have yet to be realized and that most QI programs involve 3 or more components, one of the most interesting and useful applications of a hierarchical meta-regression approach is being able to predict the mean estimated effect (and uncertainty of that effect) for a specific combination of QI components in a new setting or trial. To demonstrate this application, we present the poster and predictive distribution of a new study ($Y_{ij_{new}}$) of three combinations of QI components that were not evaluated in the diabetes QI review as described in Equations 5-7. For example, team changes, facilitated relay, and electronic patient registry are three components associated with promising differences (relative to the other

components) in post-treatment means (Table 14). However a combination of these components has not been tested in an RCT identified in the literature. We estimated posterior and predictive distributions of the predicted effect of a complex intervention comprised of these components. The estimated posterior distribution and predictive distribution of the post-treatment mean in a new study where no QI component was delivered was 8.14 (95% CrI 7.96, 8.31) and 8.13 (95% CrI 6.46, 9.81), respectively. Compared to no intervention, the complex intervention led to a significant reduction in post-treatment mean, when estimated using either the posterior distribution (-0.78 [95% CrI -1.04, -0.52]; post-treatment mean 7.36 [95% CrI 7.06, 7.66] or the posterior predictive distribution (-0.78 [-1.50, -0.06]; post-treatment mean 7.35 [95% CrI 5.54, 9.16], and thus may represent a promising combination of components to pursue in a future study.

Table 14 Estimated poster and predictive distributions of predicted effects of untested combinations

Untested combination	Median (95% CrI) of the posterior distribution of the post-treatment mean	Difference from baseline	Median (95% CrI) of the posterior predictive distribution of the post-treatment mean	Difference from baseline
CM + EPR + FR	7.79 (7.46, 8.11)	-0.35 (-0.63, -0.07)	7.79 (6.03, 9.55)	-0.34 (-0.93, 0.21)
TC + FR + PSM	7.36 (7.06, 7.66)	-0.78 (-1.04, -0.52)	7.35 (5.54, 9.16)	-0.78 (-1.50, -0.06)
TC + FR + EPR	7.41 (7.07, 7.76)	-0.73 (-1.03, -0.42)	7.41 (5.59, 9.23)	-0.72 (-1.47, -0.01)

CM: Case management; EPR: electronic patient registry; FR: facilitated relay; PSM: promotion of self-management; TC: team changes

6.9.6 Assessment of convergence, model evaluation, robustness

We used the Brooks-Gelman-Rubin diagnostic to assess parameter convergence. In most cases the upper limit of the 95% credibility interval for the Brooks-Gelman-Rubin statistic was <1.1 for all parameters monitored (96.2% of parameters monitored; median 1.016). Gelman-Rubin diagnostic was more stable for parameters estimated in the parsimonious base model (median upper limit 1.008) than those estimated in the interaction and effect modification models (median upper limit 1.017).

We used mixed posterior predictive checks [226] to compare the predicted effects from the models of new studies with the same combination of components observed in our sample, against the observed outcome means. As illustrated in Figure 13, the hierarchical model performs reasonably well in terms of its probability of returning an observed means from parameter estimates. Other standard methods, such as the Deviance Information Criterion (DIC), can also be used to evaluate model performance [226,227]. In general, the assessment approach should be chosen to reflect the goals of the modeling; we think posterior predictive checks are the most relevant approach in our application, because our primary goal was to predict the post-treatment means and mean difference in future implementations of the QI interventions. As Bayesian posterior probabilities can be influenced by specified priors [228] we performed sensitivity tests on our chosen priors and found our findings were robust to these alterations.

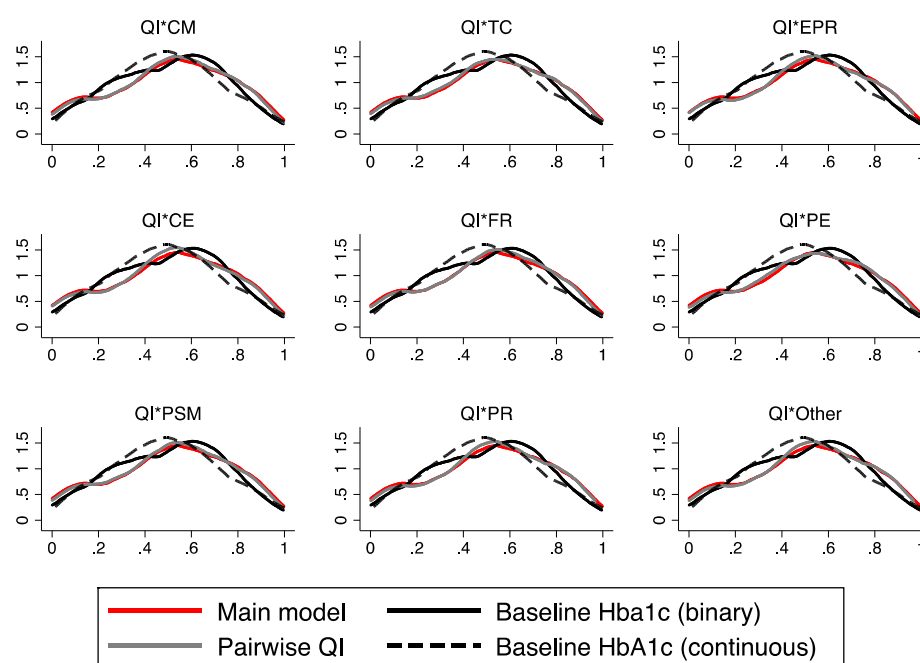


Figure 13 Posterior predictive checks*

*Each panel represents the posterior predictive check of the main model (red), pairwise interaction model (9 interaction models assessing the interaction between one QI strategy and the other 8 QI strategies; grey), effect modification model when baseline Hba1c is treated as binary (black, solid) and effect modification when baseline Hba1c is treated as continuous (black, dashed)

6.10 Discussion

Complex interventions present challenges for evidence synthesis, in particular for estimating the unique effects of components within complex interventions, the prediction of untested combinations of components, and the impact of co-occurring components (i.e., interaction) and population and system factors (i.e., effect modification). Several features of our model offer an improvement over a traditional pairwise approach. By modeling the statistical distribution of data in each study arm our model allows for the inclusion of all data from all studies and the identification of average post-treatment mean associated with each component beyond the typical average effect of heterogeneous interventions. The base model assumes intervention components to be additively and linearly related to the response outcome. This assumption may be unreasonable, and can be relaxed, data permitting, by expanding the model to include appropriate product terms as demonstrated. The model can also be extended to incorporate sophisticated approaches for imputing missing data and can be used to predict the post-treatment means of yet unrealized combinations of components.

A fully saturated comparative network of pairwise differences in our motivating example would require the estimation of 4,096 parameters (i.e., one parameter for each combination of components compared to control), which is not feasible. The hierarchical model in its simplest form allows us to estimate 12 parameters (9 in our application due to the infrequent observation of some strategies) , one for each intervention component of interest. We feel this dramatic reduction of parameters is justified if reviewers (informed by expert opinion) believe the salient aspects of the intervention are captured by the list of components, as they did in this case of the diabetes QI review [22,55] Comparison of estimates between the hierarchical model and standard pairwise analyses suggest our assumptions of the model may be reasonable. For example, while team changes remains the most promising component (based on estimated post-treatment

mean difference and ranking) the post-treatment mean difference of the other components were much more modest, and when additively combined, more consistent with the magnitude of post-treatment observed changes observed in complex intervention QI trials. In addition, the lack of strong evidence in support of pairwise interactions between QI components suggested that the additive assumption of the base model was reasonable. Finally, while baseline risk appeared to qualitatively modify the effect of some QI strategies and suggest slightly improved model fit, there was no strong evidence of systematic deviation in effects due to baseline risk to justify a larger model.

Our primary goal in developing a hierarchical regression model to synthesize evidence from clinical trials of complex interventions is to be able to predict the impact of new complex intervention programs in new settings or support the rational design of future studies. While the model is incomplete and unlikely to be specified perfectly, it may offer more informative estimates to inform decision-makers about QI interventions compared to standard models of synthesis. Our approach builds on earlier works that have sought to use observed effect sizes to construct a predictive function of future effects. Specifically, we see this as an attempt to estimate a 'response surface' as articulated by Rubin [229], and demonstrated in an application to meta-analysis data to estimate predictors of survey response rates by Gelman [23].

The use of such a model is not without its limitations. Foremost, all results are conditional on the model and the structure of the model is ultimately unknown. Substantial heterogeneity of results may remain due to the delivery of interventions in diverse population and health setting contexts, which may also be poorly reported. The addition of further data, whether from new studies or expert opinion, may allow for the better diagnoses of these sources of heterogeneity and refinement of estimated post-treatment means differences. In a way, the approach of the hierarchical model solves the easier problems. The more challenging problems of defining the intervention [59,62,230] and obtaining data from poorly reported publications remain [231,232].

Our approach may be applicable to other healthcare and public health topics of study where complex interventions are more common. Future research should explore comparisons of the hierarchical model with other meta-analytic approaches using datasets beyond diabetes QI RCTs. Simulation studies may be performed to assess bias and performance when the number of available studies is small or the number of components increases. Further model checking could be explored. Importantly though, for any application involving these methods, substantive content expertise is required to identify the appropriate list of components of interest (and their composite combinations), population and setting predictors hypothesized to modify interventions, and outcomes. Input from experts and external evidence is particularly important for model shrinkage (e.g., restricting analyses to only components and combinations of interest, informing priors, etc.). While our models represent our best estimate of effects to date, decision-makers are still faced with the burden of linking synthesized evidence with information on population outcomes, costs, values, and resource limitations unique to their particular setting. Further work in developing decisional models and decision tools to support linking estimates of intervention effects to the wider spectrum of evidence used for decision-making is needed [4,233].

6.11 Conclusion

Hierarchical meta-regression models provide a useful method for reviewers to develop more elaborate models to better learn from existing evidence and explore questions of real interest to decision-makers. Compared to standard meta-analysis models they may allow for the better prediction of the post-treatment mean in future implementations of a complex intervention and support a more rational design of future studies.

Acknowledgements

The authors would like to acknowledge Samir Nazarali, who assisted in data checking and re-extraction of data used in this paper.

Conflict of interest

None declared

The following supplemental information for Manuscript 4 can be found in Appendix E:

Supplemental Text E-1	Imputation of missing variances and estimates of the ICC
Supplemental Table E-1	Analysis bins for hierarchical models
Supplemental Text E-2	Hierarchical model implemented in Analysis 2
Supplemental Text E-3	Hierarchical model implemented in Analysis 3
Supplemental Text E-4	Extension to hierarchical model: Interactions among intervention components
Supplemental Text E-5	Extension to hierarchical model: Effect modification – binary HbA1c
Supplemental Text E-6	Extension to hierarchical model: Effect modification – continuous HbA1c

CHAPTER 7 Discussion

7.0 Discussion

Science is a cumulative process that requires careful consideration of past knowledge to effectively and efficiently guide future applications and discovery [234]. Systematic reviews of the complete set of studies pertaining to a scientific question, and meta-analysis of their observed outcomes, have become standard tools to support robust evidence synthesis of healthcare interventions [5]. While the foundational processes of evidence synthesis are well established (e.g., need for comprehensive searches to identify all relevant studies, need to assess the methodological quality of studies), the specific methods reviewers should adopt to best operationalize review processes are not always certain. Uncertainty of optimal review methods is especially challenging when the evidence to be synthesized comes from evaluations of complex interventions.

In the context of updating a large systematic review of complex diabetes quality improvement (QI) interventions, I identified three methodological challenges relating to evidence *collection*, *abstraction*, and *synthesis* that my research program sought to address. **First**, during data collection, review authors are recommended to attempt to contact study authors to obtain or clarify any information deemed to be missing or unclear, respectively, from study reports. Yet, there is uncertainty over what specific author contact approach should be used to obtain additional information (e.g., contact method, frequency of contact, etc.) and the impact of one approach over another in terms author response and review resources. **Second**, during data abstraction of cluster randomized trials (CRTs), we know reviewers need to determine if clustering was accounted for in the analysis of study effect estimates, and if not, make necessary corrections to unadjusted effect estimates (i.e., estimates with unit of analysis errors) prior to meta-analysis. Yet, identifying and correcting unit of analysis errors is not always straightforward and there is uncertainty on how to best extract adjusted effect estimates from study reports (i.e., when data is not presented in the optimal format needed for meta-analysis) and how to account for uncertainty in the ICC when making approximate corrections of unadjusted data. **Third**, during syntheses, we know that the conventional

meta-analysis models offer reviewers a limited approach to learn about complex interventions since the diversity of complex interventions is collapsed into a single heterogeneous estimate of average effect and methods to explore predictors of heterogeneity (e.g., diverse intervention components) under such a model are limited.

The aim of my doctoral research program was to explore the utility of three methodological approaches to address these challenges and optimize the synthesis of complex interventions using a large systematic review of diabetes QI interventions as a case study. The studies are conceived as applications of ‘meta-research’, an “evolving scientific discipline that aims to evaluate and improve research practices” [235] and specifically included: 1) a randomized controlled trial (RCT) to compare the effects of two author contact strategies on collecting additional data from study authors; 2) a descriptive and methodological study designed to optimize the abstraction and correction of clustered data, respectively; and 3) a methodological study to assess the utility of a hierarchical multivariate regression model to predict complex intervention effects.

7.1 Summary of research findings

The first study investigated the effectiveness of contacting non-responding authors of studies in our systematic review by telephone call as compared to email to request additional information on intervention components and potential effect modifiers. Compared to email, calling authors significantly improved the odds of authors completing our request for additional information (response rate 36.7% vs. 20.2%; difference 16.5%; adjusted odds ratio 2.26 [95% CI 1.10-4.76]), but took more time (20 vs. 10 hours over several months vs. one month in total to deliver the telephone and email interventions, respectively) and was more expensive (approximately \$505 vs. \$253 in total). Additionally, operationalizing the telephone intervention was somewhat more challenging, particularly with respect to contacting non-English authors and reaching authors for whom no direct number could be found. To our knowledge, the author RCT is

the first study to empirically compare the effectiveness of telephone vs. continued emailing of non-responding authors. We see this Study Within A Review (SWAR) [7] as an important contribution to meta-research evidence informing methodological and budgeting considerations of author contact in future systematic reviews.

The second and third studies addressed challenges in abstracting and correcting data from CRTs, first by determining the proportion of studies with unit of analysis errors and availability of data to correct unit of analysis errors, and second, by investigating the utility of building a database of ICCs and posterior predictive distribution to impute missing ICCs to correct unit of analysis errors. As recommended in the Cochrane Handbook, expert advice was sought in both identifying and abstracting data from CRTs, which was performed by two independent researchers. The descriptive analysis presented in Study 2 found that while a high proportion of CRTs in our sample adjusted for clustering effects (67% across outcomes; range 25%-81% within outcomes), few reported enough information to extract standard errors of adjusted effect estimates or an estimate of the ICC to correct unadjusted standard errors or sample sizes. Thus, an important finding of this work is the need to improve the reporting of primary studies to efficiently meet the data needs of evidence synthesis (for example, by requiring the standardized reporting of adjusted effect estimates and an estimate of the ICC for all outcomes). Using evidence on ICCs obtained from Study 2, combined ICCs obtained from additional sources, Study 3 demonstrated the feasibility and utility of building an ICC database and posterior predictive distribution to impute external estimates of the ICC for studies with unit of analysis errors in which the ICC is missing. Although correcting unit of analysis errors using ICCs drawn from the posterior predictive distribution did not markedly change results in our application, our approach offers two methodological advantages: first, it guards against exaggerated adjusting of effect estimates when the use of a single ICC differs substantially from the estimated mean of a posterior predictive distribution, and second, it appropriately accounts for the uncertainty of the ICC estimate in the precision of meta-analysis results.

Finally, the fourth study of this thesis illustrated the use of hierarchical multivariate meta-regression as a quantitative approach to synthesize the effects of complex interventions and explore effect heterogeneity. Using an arm-based analysis of post-intervention means of one continuous outcome, we demonstrate that hierarchical multivariate meta-regression can be used to estimate the average distribution in post-intervention outcome associated with various study covariates (e.g., intervention components, effect modifiers). In other words, the model can be used to estimate a conditional mean function based on the specified covariates hypothesized to drive intervention effects – what Rubin calls a ‘response surface’[229]. To achieve the dramatic reduction in parameters to be modeled, the model leverages strong assumptions (e.g., grouping of variants of components into component categories, ignoring co-interventions not captured by coding taxonomy, assuming additivity and linearity of effects) informed by expert knowledge and data-availability. While the model cannot be used to estimate direct causal effects (as it is unlikely to ever be specified perfectly, or have sufficient data to support a perfect model), it may be used to infer predictions on the average effect one would observe if a combination of specific covariates were implemented in a new setting with a large sample. As such, the hierarchical model may still offer an improvement over conventional meta-analysis models that estimate a distribution of the average effect for all studies, irrespective of components and effect modifiers. In addition to providing decision-makers with a realistic range of outcomes they may expect from a given combination of components, the model may be especially useful in guiding researchers towards the evaluation of promising components with high uncertainty (e.g., clinician education: -0.19 [95% CrI - 0.45, 0.06]) or potentially effective combinations of components that have yet to be tested (e.g., team changes + facilitated relay + promotion of self management: -0.78 ([95% CrI -1.04, -0.52])). While the hierarchical model offers a flexible method to explore variation in effects, the ability to model multiple covariates and estimate precise coefficients of their average effect will be limited by data. Our current base model had

241 data points (i.e., arm means) to model 10 coefficients. Application of this approach may not be feasible for reviews with a small number of included studies and/or a large number of covariates of interest.

A connecting theme across the four studies was that the proposed methods offer improvements over conventional approaches, but take longer and are more complicated. It therefore raises the question of the feasibility and utility of such approaches, particularly if estimates do not change dramatically (as in the ICC imputation study) or even, are substantially reduced (as in the hierarchical model study). However, to focus on just the results of our single application to one outcome in our review would be miss the larger picture of what the methods add – more data (Study 1), more informed imputations and explicit modeling of uncertainty (Study 3), and better modeling of hypothesized predictors of effects (Study 4). Contacting authors by telephone took longer but resulted in a higher retrieval of additional information. Building an ICC database and imputation model was more extensive than imputation using a single conservative ICC, but resulted in more informed imputations and meta-analysis estimates that appropriately incorporated ICC uncertainty into their parameters. Hierarchical multivariate models led to more realistic estimates of the distribution of observed outcomes due to the presence of intervention components, controlling for the presence of other components and study covariates.

A strength of these methods is not that they are ‘the way’ to conduct data collection, abstraction, and synthesis of complex interventions, but more that they offer transparent, reproducible, and useful methods on which to critique and build. Other investigators may for example choose to compare a different approach to contact authors or assume different priors or likelihood to model intervention effects, leading to different response rates and model inferences, respectively. By conducting more Studies Within A Review (such as the author RCT) and conducting more comparisons of ICC imputation and hierarchical model applications, we can evaluate the impact of methods

on obtained data, synthesized estimates, and ultimately decision-making and patient outcomes.

7.2 Limitations

There are several limitations in this thesis to consider. First, the thesis studies were based on a sample of trials included in an existing diabetes QI review. Therefore, any limitation in the diabetes review, such as exclusion of non-English trials or measurement errors (e.g., coding interventions, abstracting outcome data) would have carried over to the thesis studies. If for example, relevant studies were missed (or components coded within studies), it would have led to reduced precision in the estimate of ICC distributions or average effect of covariate(s). To guard against coding errors, I, along with two experienced investigators (Dr. Andrea Tricco and Dr. Noah M. Ivers) rechecked the QI codes for the former review (including studies up to 2010), and together with Dr. Ivers, we independently coded the QI strategies for trials captured in the update (2010-2014). To further improve coding of intervention components, we plan to compare the agreement between our codes and those given by respondents to the author survey, and where disagreement exists, explore survey responses in more depth to understand the reasons for disagreement.

Second, during the conduct of the author RCT, there were several errors made in the delivery of the telephone intervention and measurement of time for both the telephone and email intervention groups. With respect to intervention delivery, there were two protocol violations in which I emailed authors who could not be reached by telephone due to incorrect numbers. In one case, an email was sent as a reply to a pre-RCT email from the author to troubleshoot the survey URL. In the other, an email was sent to a corrected email identified during searches for the authors' phone number before moving on to a co-author. Both email contacts resulted in the authors completing the survey. While the protocol violations may have led to an overestimate of the intervention effect, the main findings of the study would be the same had the authors not complete the two

surveys. Still caution should be taken in interpreting results, and ideally a future Study Within A Review will seek to replicate a comparison of telephone and email modalities, either among non-responding authors as in our RCT, or to all authors. In addition, there were some errors in the measurement of time, particularly for the time taken during follow-up (e.g., responding to author emails, following up with authors in the telephone group who agreed to complete the survey). These errors aside, I was extremely careful in the measurement of time and believe my failure these instances was non-differential but rather was due to performing review tasks 'automatically' before I remembered to turn on the timer. A rough estimate of the time missing from both groups was comparable suggesting that while the total time and cost to deliver the telephone and email interventions was underestimated, our interpretation of the relative difference between interventions (i.e., that telephone calling took more time and was more costly) would be the same. Taken together, these errors may represent challenges of conducting a SWAR [7] where the appropriate conduct of an RCT of review methods (e.g., protocol adherence, measurement without error) may occasionally conflict with the effective and efficient conduct of the review itself (e.g., obtaining author contact regardless of method, efficiently performing review tasks).

Finally, results from Study 3 and 4 are both conditional on their respective models, and the structures of these models are ultimately unknown. Substantial heterogeneity in the posterior predictive distributions of the ICCs (Study 3) and average effects of components (Study 4) may remain due to the delivery of the interventions in diverse populations, settings, and contexts. The addition of further data, whether from new studies or expert opinion, may allow for the better diagnoses of these sources of heterogeneity and refinement of model estimates.

7.3 Implications for future research

Based on the findings of this thesis, there are several implications for future research involving the conduct and synthesis of complex interventions. These include implications

pertaining to: 1) the diabetes QI review, 2) review methodology for complex interventions in general, and 3) planning and reporting trials evaluating complex interventions.

7.3.1 Implications for the diabetes QI review

For the diabetes QI review, we should first continue to consider ways to refine the response surface of the HbA1c outcome and other review outcomes to be estimated. Further to the addition of more data points (e.g., 262 arms from 122 trials for the HbA1c outcome from studies included in the update to 2014), the response surface for HbA1c may be improved through the refinement of the component covariates (e.g., corrected QI coding based on author survey responses, alternative intervention taxonomies [62]), addition of further effect modifiers (e.g., presence of co-occurring cardiovascular disease in study population), or addition of covariates to model study quality (e.g., use of appropriate random allocation method, protocol registration) [189,229], among others. In particular, we have already demonstrated the feasibility of coding QI interventions using the Behaviour Change Technique Taxonomy Version 1 for the data included in the 2012 review [236]. A behaviour change techniques (BCT) is defined as the “irreducible component of an intervention designed to alter or redirect causal processes that regulate behaviour” [222] and thus may better define the active ingredients leading to provider and behaviour change in the diabetes QI interventions. The next step is to compare the modelled HbA1c response when components are specified as BCTs as opposed to QI strategies. In addition, further updates to the review are already planned as part of a Living Systematic Review initiative [237,238] and will be helpful in providing additional data points (beyond a modest ~500 included in the most recent update) to estimate a more informative and precise response surface. Finally, while our final models for the updated review will represent our best estimate of effects to date, decision-makers are still faced with the burden of linking synthesized evidence with information on population outcomes, costs, values, and resource limitations unique to particular healthcare settings. We have therefore planned further work to develop a decisional

model and decision tool to link estimates of intervention effects from the diabetes QI review to the wider spectrum of evidence on benefits, harms, and costs to predict, for different QI interventions, the long term clinical outcomes, quality-adjusted life expectancy, and resource use and costs at the population level [4,233].

7.3.2 General implications for review methodology of complex interventions

With respect to the methods of evidence synthesis for complex interventions in general, further work is needed to replicate and extend the findings from studies in this thesis to future SWARs and synthesis applications. Methods for author contact for example should continue to be evaluated and optimized. While both telephone and email interventions led to survey completion among non-responding authors, the majority of authors randomized to these interventions did not respond (63.2% and 79.8% in telephone and email arms, respectively), suggesting opportunities for improvement in both strategies. In both cases, additional upfront investment in searching for telephone numbers and emails may help limit authors non-response due to suboptimal (indirect) numbers or expired emails. Ideally, both telephone and email could be reported as data elements in trial registration (e.g., ClinicalTrials.gov) and be updated as information evolves, to avoid reviewers spending time on inefficient Internet searches. In addition, detailed methods (and applications of methods) for abstracting and correcting data from CRTs could be better reported in the literature. As with author contact, reviews should explicitly report details of what they did, including how CRT data was assessed and abstracted data was handled (e.g., calculations used to convert adjusted effect estimates to standard errors, approach used to correct unadjusted effect estimates). Ideally, a review of such thoroughly reported applications could be used to provide more explicit guidance and worked examples for reviewers in the future. Finally, for both the ICC imputation and hierarchical model, model codes can be adapted and applied to other reviews of complex interventions to compare their feasibility and utility for datasets beyond diabetes QI trials.

7.3.3 Implications for planning and reporting trials of complex interventions

If the models are to be believed as approximately accurate, then estimates from both hierarchical models and ICC predictive distributions could be used to plan future trials. For example, trialists could use estimates from hierarchical models to build the ‘best’ predicted versions of complex interventions (e.g., combinations of high ranking components) and evaluate these in head-to-head trials. Alternatively, estimates from the proposed hierarchical models could be used to identify components that may have a clinically meaningful effect but are too imprecise to be certain, and thus require additional evidence. Additionally, if future trials are to be conducted using a CRT design, then predicted ICC distributions could be used to estimate an effective sample size, as per the approach advocated by Turner et al. [79].

7.4 Final conclusions

The results from this thesis suggest three methodological approaches (contacting authors by telephone, imputing missing ICCs using a posterior predictive distribution, estimating complex intervention effects using a hierarchical multivariate meta-regression) can be used to optimize the processes of synthesizing complex interventions. Further work is needed to evaluate the impact of additional study-covariates on explaining residual heterogeneity and testing these methods in other reviews complex interventions.

REFERENCE LIST

- [1] Straus SE, Tetroe J, Graham I. Defining knowledge translation. *CMAJ* 2009;181:165–8. doi:10.1503/cmaj.081229.
- [2] Slutsky J, Atkins D, Chang S, Collins Sharp BA. AHRQ Series Paper 1: Comparing medical interventions: AHRQ and the Effective Health-Care Program. *J Clin Epidemiol* 2010;63:481–3. doi:10.1016/j.jclinepi.2008.06.009.
- [3] Oxford Centre for Evidence-based Medicine - Levels of Evidence (March 2009) - CEBM n.d. <https://www.cebm.net/2009/06/oxford-centre-evidence-based-medicine-levels-evidence-march-2009/> (accessed June 5, 2018).
- [4] Welton NJ, Sutton AJ, Cooper NJ, Abrams KR, Ades AE. Evidence synthesis for decision making in healthcare. Wiley; 2012.
- [5] Higgins J, Green S, editors. *Cochrane Handbook for Systematic Reviews of Interventions*. Version 5. The Cochrane Collaboration; 2011.
- [6] Tricco AC, Tetzlaff J, Moher D. The art and science of knowledge synthesis. *J Clin Epidemiol* 2011;64:11–20. doi:10.1016/j.jclinepi.2009.11.007.
- [7] SWAT/SWAR Information | The Northern Ireland Network for Trials Methodology Research | Queen’s University Belfast n.d. <https://www.qub.ac.uk/sites/TheNorthernIrelandNetworkforTrialsMethodologyResearch/SWATSWARInformation/> (accessed May 23, 2018).
- [8] Pigott T, Noyes J, Umscheid CA, Myers E, Morton SC, Fu R, et al. AHRQ series on complex intervention systematic reviews-paper 5: Advanced analytic methods. *J Clin Epidemiol* 2017. doi:10.1016/j.jclinepi.2017.06.015.
- [9] Petticrew M, Rehfuess E, Noyes J, Higgins JPT, Mayhew A, Pantoja T, et al. Synthesizing evidence on complex interventions: how meta-analytical, qualitative, and mixed-method approaches can contribute. *J Clin Epidemiol* 2013;66:1230–43. doi:10.1016/j.jclinepi.2013.06.005.
- [10] Petticrew M, Anderson L, Elder R, Grimshaw J, Hopkins D, Hahn R, et al. Complex interventions and their implications for systematic reviews: a pragmatic approach.

- J Clin Epidemiol 2013;66:1209–14. doi:10.1016/j.jclinepi.2013.06.004.
- [11] Kühne F, Ehmcke R, Härter M, Kriston L. Conceptual decomposition of complex health care interventions for evidence synthesis: a literature review. *J Eval Clin Pract* 2015;21:817–23. doi:10.1111/jep.12384.
 - [12] Pigott T, Shepperd S. Identifying, documenting, and examining heterogeneity in systematic reviews of complex interventions. *J Clin Epidemiol* 2013;66:1244–50. doi:10.1016/j.jclinepi.2013.06.013.
 - [13] Young T, Hopewell S. Methods for obtaining unpublished data. *Cochrane Database Syst Rev* 2011:MR000027. doi:10.1002/14651858.MR000027.pub2.
 - [14] Siebenhofer A, Erckenbrecht S, Pregartner G, Berghold A, Muth C. How often are interventions in cluster-randomised controlled trials of complex interventions in general practices effective and reasons for potential shortcomings? Protocol and results of a feasibility project for a systematic review. *BMJ Open* 2016;6:e009414. doi:10.1136/bmjopen-2015-009414.
 - [15] Eccles M, Grimshaw J, Campbell M, Ramsay C. Research designs for studies evaluating the effectiveness of change and improvement strategies. *Qual Saf Health Care* 2003;12:47–52.
 - [16] Donner A, Klar N. Design and analysis of cluster randomization trials in health research. John Wiley & Sons; 2000.
 - [17] Donner A, Birkett N, Buck C. Randomization by cluster. Sample size requirements and analysis. *Am J Epidemiol* 1981;114:906–14.
 - [18] Klar N, Donner A. Current and future challenges in the design and analysis of cluster randomization trials. *Stat Med* 2001;20:3729–40.
 - [19] Whiting-O’Keefe QE, Henke C, Simborg DW. Choosing the correct unit of analysis in Medical Care experiments. *Med Care* 1984;22:1101–14.
 - [20] Richardson M, Garner P, Donegan S. Cluster randomised trials in Cochrane reviews: Evaluation of methodological and reporting practice. *PLoS One* 2016;11:e0151818. doi:10.1371/journal.pone.0151818.
 - [21] Caldwell DM, Welton NJ. Approaches for synthesising complex mental health

- interventions in meta-analysis. *Evid Based Ment Heal* 2016;19:16–21.
doi:10.1136/eb-2015-102275.
- [22] Ivers N, Tricco AC, Trikalinos TA, Dahabreh IJ, Danko KJ, Moher D, et al. Seeing the forests and the trees--innovative approaches to exploring heterogeneity in systematic reviews of complex interventions to enhance health system decision-making: a protocol. *Syst Rev* 2014;3:88. doi:10.1186/2046-4053-3-88.
 - [23] Gelman A, Stevens M, Chan V. Regression modeling and meta-analysis for decision making: A cost-benefit analysis of incentives in telephone surveys. *J Bus Econ Stat* 2002;21:213–25. doi:10.2307/1392457.
 - [24] Mulrow CD. Rationale for systematic reviews. *BMJ* 1994;309:597–9.
 - [25] Oxman AD, Guyatt GH. The science of reviewing research. *Ann N Y Acad Sci* 1993;703:125-33; discussion 133-4.
 - [26] Lau J, Antman EM, Jimenez-Silva J, Kupelnick B, Mosteller F, Chalmers TC. Cumulative meta-analysis of therapeutic trials for myocardial infarction. *N Engl J Med* 1992;327:248–54. doi:10.1056/NEJM199207233270406.
 - [27] Glass G V. Primary, secondary, and meta-analysis of research. *Educ Res* 1976;5:3. doi:10.2307/1174772.
 - [28] Wallace BC, Dahabreh IJ, Schmid CH, Lau J, Trikalinos TA. Chapter 12. Modernizing evidence synthesis for evidence-based medicine. *Clin Decis Support* 2014:339–62. doi:10.1016/B978-0-12-398476-0.00012-9.
 - [29] Chalmers I, Enkin M, Keirse MJNC. Preparing and updating systematic reviews of randomized controlled trials of health care. *Milbank Q* 1993;71:411. doi:10.2307/3350409.
 - [30] The AGREE Reporting Checklist: a tool to improve reporting of clinical practice guidelines. *BMJ* 2016;354:i4852. doi:10.1136/BMJ.i4852.
 - [31] Clarke M, Hopewell S, Chalmers I. Reports of clinical trials should begin and end with up-to-date systematic reviews of other relevant evidence: A status report. *J R Soc Med* 2007;100:187–90. doi:10.1258/jrsm.100.4.187.
 - [32] Viechtbauer W. Learning from the past: Refining the way we study treatments. *J*

- Clin Epidemiol 2010;63:980–2. doi:10.1016/j.jclinepi.2010.04.004.
- [33] Cooper NJ, Jones DR, Sutton AJ. The use of systematic reviews when designing studies. Clin Trials 2005;2:260–4. doi:10.1191/1740774505cn090oa.
 - [34] Shea BJ, Reeves BC, Wells G, Thuku M, Hamel C, Moran J, et al. AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. BMJ 2017;358:j4008.
 - [35] Moher D, Shamseer L, Clarke M, Ghersi D, Liberati A, Petticrew M, et al. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. Syst Rev 2015;4:1. doi:10.1186/2046-4053-4-1.
 - [36] Centre for Reviews and Dissemination. PROSPERO: International prospective register of systematic reviews n.d. <https://www.crd.york.ac.uk/prospERO/> (accessed June 7, 2018).
 - [37] Bellary S, O’Hare JP, Raymond NT, Gumber A, Mughal S, Szczepura A, et al. Enhanced diabetes care to patients of south Asian ethnic origin (the United Kingdom Asian Diabetes Study): a cluster randomised controlled trial. Lancet 2008;371:1769–76. doi:10.1016/S0140-6736(08)60764-3.
 - [38] Balshem H, Stevens A, Ansari M, Norris S, Kansagara D, Shamliyan T, et al. Finding grey literature evidence and assessing for outcome and analysis reporting biases when comparing medical interventions: AHRQ and the Effective Health Care Program. 2008.
 - [39] Deeks JJ, Higgins JP, editors. Chapter 7: Selecting studies and collecting data. Cochrane Handb. Syst. Rev. Interv. Version 5.1.0 [updated March 2011], The Cochrane Collaboration; 2011.
 - [40] Centre for Reviews and Dissemination. Systematic reviews: CRD’s guidance for undertaking reviews in health care. vol. 10. 2009. doi:10.1016/S1473-3099(10)70065-7.
 - [41] Cooper C, Booth A, Britten N, Garside R. A comparison of results of empirical studies of supplementary search techniques and recommendations in review methodology handbooks: a methodological review. Syst Rev 2017;6:234.

- doi:10.1186/s13643-017-0625-1.
- [42] Mullan RJ, Flynn DN, Carlberg B, Tleyjeh IM, Kamath CC, LaBella ML, et al. Systematic reviewers commonly contact study authors but do so with limited rigor. *J Clin Epidemiol* 2009;62:138–42. doi:10.1016/j.jclinepi.2008.08.002.
 - [43] Foy R, Sales A, Wensing M, Aarons GA, Flottorp S, Kent B, et al. Implementation science: a reappraisal of our journal mission and scope. *Implement Sci* 2015;10:51. doi:10.1186/s13012-015-0240-2.
 - [44] Grimshaw JM, Eccles MP, Lavis JN, Hill SJ, Squires JE. Knowledge translation of research findings. *Implement Sci* 2012;7:50. doi:10.1186/1748-5908-7-50.
 - [45] French SD, Green SE, O'Connor DA, McKenzie JE, Francis JJ, Michie S, et al. Developing theory-informed behaviour change interventions to implement evidence into practice: a systematic approach using the Theoretical Domains Framework. *Implement Sci* 2012;7:38. doi:10.1186/1748-5908-7-38.
 - [46] Guise J-M, Chang C, Butler M, Viswanathan M, Tugwell P. AHRQ series on complex intervention systematic reviews-paper 1: An introduction to a series of articles that provide guidance and tools for reviews of complex interventions. *J Clin Epidemiol* 2017;90:6–10. doi:10.1016/j.jclinepi.2017.06.011.
 - [47] Anderson LM, Petticrew M, Chandler J, Grimshaw J, Tugwell P, O'Neill J, et al. Considering complexity in systematic reviews of interventions: Introducing a series of methodological articles on considering complexity in systematic reviews of interventions. *J Clin Epidemiol* 2013;66:1205–8. doi:10.1016/j.jclinepi.2013.07.005.
 - [48] Craig P, Dieppe P, Macintyre S, Michie S, Nazareth I, Petticrew M. Developing and evaluating complex interventions: the new Medical Research Council guidance. *BMJ* 2008;a1655. doi:10.1136/bmj.a1655.
 - [49] The National Collaborating Centre for Chronic Conditions. Type 2 diabetes 2008.
 - [50] Diabetes Canada. Diabetes Canada | Clinical Practice Guidelines - Full Guidelines 2018. <http://guidelines.diabetes.ca/fullguidelines> (accessed August 22, 2017).
 - [51] American Diabetes Association. American Diabetes Association Standards of

- medical care in diabetes - 2017. *Diabetes Care* 2017;40.
- [52] Diabetes | Guidance and guideline topic | NICE n.d.
 - [53] Centre LH and SCI. HSCIC: National Diabetes Audit 2010–2011 Report 1: Care Processes and Treatment Targets. 2012.
 - [54] Presseau J, Mackintosh J, Hawthorne G, Francis JJ, Johnston M, Grimshaw JM, et al. Cluster randomised controlled trial of a theory-based multiple behaviour change intervention aimed at healthcare professionals to improve their management of type 2 diabetes in primary care. *Implement Sci* 2018;13:65. doi:10.1186/s13012-018-0754-5.
 - [55] Tricco AC, Ivers NM, Grimshaw JM, Moher D, Turner L, Galipeau J, et al. Effectiveness of quality improvement strategies on the management of diabetes: a systematic review and meta-analysis. *Lancet* 2012;379:2252–61. doi:10.1016/S0140-6736(12)60480-2.
 - [56] Shojania KG, Ranji SR, Shaw LK, Charo LN, Lai JC, Ruskakoff RJ, et al. Closing the Quality Gap : A Critical Analysis of Quality Improvement Strategies. Volume 2 - Diabetes Mellitus Care. Agency for Healthcare Research and Quality; 2004.
 - [57] Cochrane Effective Practice and Organization of Care (EPOC). EPOC Taxonomy 2002. <http://epoc.cochrane.org/epoc-taxonomy>.
 - [58] Hoffmann TC, Eructi C, Glasziou PP. Poor description of non-pharmacological interventions: analysis of consecutive sample of randomised trials. *BMJ* 2013;347:f3755. doi:10.1136/bmj.f3755.
 - [59] Michie S, Fixsen D, Grimshaw JM, Eccles MP. Specifying and reporting complex behaviour change interventions: the need for a scientific method. *Implement Sci* 2009;4:40. doi:10.1186/1748-5908-4-40.
 - [60] Riley BL, MacDonald J, Mansi O, Kothari A, Kurtz D, vonTettenborn LI, et al. Is reporting on interventions a weak link in understanding how and why they work? A preliminary exploration using community heart health exemplars. *Implement Sci* 2008;3:27. doi:10.1186/1748-5908-3-27.
 - [61] Colquhoun H, Leeman J, Michie S, Lokker C, Bragge P, Hempel S, et al. Towards a

- common terminology: a simplified framework of interventions to promote and integrate evidence into health practices, systems, and policies. *Implement Sci* 2014;9:51. doi:10.1186/1748-5908-9-51.
- [62] Lokker C, McKibbin KA, Colquhoun H, Hempel S. A scoping review of classification schemes of interventions to promote and integrate evidence into practice in healthcare. *Implement Sci* 2015;10:220. doi:10.1186/s13012-015-0220-6.
 - [63] Walshe K. Pseudoinnovation: the development and spread of healthcare quality improvement methodologies. *Int J Qual Health Care* 2009;21:153–9. doi:10.1093/intqhc/mzp012.
 - [64] Higgins JP, Deeks JJ, Altman DG on behalf of the CSMG. Chapter 16: Special topics in statistics. In: Higgins J, Green S, editors. *Cochrane Handb. Syst. Rev. Interv.* Version 5., The Cochrane Collaboration; 2011.
 - [65] Ivers NM, Tricco AC, Taljaard M, Halperin I, Turner L, Moher D, et al. Quality improvement needed in quality improvement randomised trials: systematic review of interventions to improve care in diabetes. *BMJ Open* 2013;3. doi:10.1136/bmjopen-2013-002727.
 - [66] Fisher RA. *The design of experiments*. Edinburgh: Oliver and Boyd; 1935.
 - [67] Simpson JM, Klar N, Donnor A. Accounting for cluster randomization: a review of primary prevention trials, 1990 through 1993. *Am J Public Health* 1995;85:1378–83.
 - [68] Ivers NM, Taljaard M, Dixon S, Bennett C, McRae A, Taleban J, et al. Impact of CONSORT extension for cluster randomised trials on quality of reporting and study methodology: review of random sample of 300 trials, 2000–8. *BMJ* 2011;343:d5886–d5886. doi:10.1136/bmj.d5886.
 - [69] Eldridge S, Ashby D, Bennett C, Wakelin M, Feder G. Internal and external validity of cluster randomised trials: systematic review of recent trials. *BMJ* 2008;336:876–80. doi:10.1136/bmj.39517.495764.25.
 - [70] Donner A, Brown KS, Brasher P. A methodological review of non-therapeutic intervention trials employing cluster randomization, 1979–1989. *Int J Epidemiol*

1990;19:795–800.

- [71] Donner A, Piaggio G, Villar J. Statistical methods for the meta-analysis of cluster randomization trials. *Stat Methods Med Res* 2001;10:325–38. doi:10.1177/096228020101000502.
- [72] Kish L. Survey sampling. New York: John Wiley; 1965.
- [73] Rao JN, Scott AJ. A simple method for the analysis of clustered binary data. *Biometrics* 1992;48:577–85.
- [74] Borenstein M, Hedges L V., Higgins JPT, Rothstein HR. A basic introduction to fixed-effect and random-effects models for meta-analysis. *Res Synth Methods* 2010;1:97–111. doi:10.1002/jrsm.12.
- [75] Ukoumunne OC, Gulliford MC, Chinn S, Sterne JA, Burney PG. Methods for evaluating area-wide and organisation-based interventions in health and health care: a systematic review. *Health Technol Assess* 1999;3:iii-92.
- [76] Campbell MJ. Cluster randomized trials in general (family) practice research. *Stat Methods Med Res* 2000;9:81–94. doi:10.1177/096228020000900202.
- [77] Campbell M, Grimshaw J, Steen N. Sample size calculations for cluster randomised trials. Changing Professional Practice in Europe Group (EU BIOMED II Concerted Action). *J Health Serv Res Policy* 2000;5:12–6. doi:10.1177/135581960000500105.
- [78] Feng Z, Grizzle JE. Correlated binomial variates: Properties of estimator of intraclass correlation and its effect on sample size calculation. *Stat Med* 1992;11:1607–14. doi:10.1002/sim.4780111208.
- [79] Turner RM, Thompson SG, Spiegelhalter DJ. Prior distributions for the intracluster correlation coefficient, based on multiple previous estimates, and their application in cluster randomized trials. *Clin Trials* 2005;2:108–18. doi:10.1191/1740774505cn072oa.
- [80] O’Connor PJ, Desai J, Solberg LI, Reger LA, Crain AL, Asche SE, et al. Randomized trial of quality improvement intervention to improve diabetes care in primary care settings. *Diabetes Care* 2005;28:1890–7.
- [81] Viechtbauer W. Conducting Meta-Analyses in *R* with the **metafor** Package. *J Stat*

- Softw 2010;36:1–48. doi:10.18637/jss.v036.i03.
- [82] Noyes J, Gough D, Lewin S, Mayhew A, Michie S, Pantoja T, et al. A research and development agenda for systematic reviews that ask complex questions about complex interventions. *J Clin Epidemiol* 2013;66:1262–70. doi:10.1016/j.jclinepi.2013.07.003.
 - [83] Melendez-Torres GJ, Bonell C, Thomas J. Emergent approaches to the meta-analysis of multiple heterogeneous complex interventions. *BMC Med Res Methodol* 2015;15:47. doi:10.1186/s12874-015-0040-z.
 - [84] Tanner-Smith EE, Grant S. Meta-analysis of complex interventions. *Annu Rev Public Heal* 2018;39:16.1-16.17. doi:10.1146/annurev-publhealth.
 - [85] Borenstein M, Higgins JPT. Meta-analysis and subgroups. *Prev Sci* 2013;14:134–43.
 - [86] Welton NJ, Caldwell DM, Adamopoulos E, Vedhara K. Mixed treatment comparison meta-analysis of complex interventions: psychological interventions in coronary heart disease. *Am J Epidemiol* 2009;169:1158–65. doi:10.1093/aje/kwp014.
 - [87] Sun X, Briel M, Walter SD, Guyatt GH. Is a subgroup effect believable? Updating criteria to evaluate the credibility of subgroup analyses. *BMJ* 2010;340:c117–c117. doi:10.1136/bmj.c117.
 - [88] van Houwelingen H, Senn S. Investigating underlying risk as a source of heterogeneity in meta-analysis. *Stat Med* 1999;16:110–5.
 - [89] Senn S. Hans van Houwelingen and the art of summing up. *Biom J* 2010;52:85–94. doi:10.1002/bimj.200900074.
 - [90] Glasziou P, Altman DG, Bossuyt P, Boutron I, Clarke M, Julious S, et al. Reducing waste from incomplete or unusable reports of biomedical research. *Lancet* 2014;383:267–76. doi:10.1016/S0140-6736(13)62228-X.
 - [91] Gibson CA, Bailey BW, Carper MJ, Lecheminant JD, Kirk EP, Huang G, et al. Author contacts for retrieval of data for a meta-analysis on exercise and diet restriction. *Int J Technol Assess Health Care* 2006;22:267–70. doi:10.1017/S0266462306051105.

- [92] Brown T, Hooper L. Effectiveness of brief contact with authors. XI Cochrane Colloq. Evidence, Heal. Care Cult., Barcelona, Spain: n.d.
- [93] Shojania KG, Ranji SR, McDonald KM, Grimshaw JM, Sundaram V, Rushakoff RJ, et al. Effects of quality improvement strategies for type 2 diabetes on glycemic control: a meta-regression analysis. *JAMA* 2006;296:427–40. doi:10.1001/jama.296.4.427.
- [94] SWAR Store | The Northern Ireland Network for Trials Methodology Research | Queen's University Belfast n.d.
<http://www.qub.ac.uk/sites/TheNorthernIrelandNetworkforTrialsMethodologyResearch/SWATSWARInformation/Repositories/SWARStore/> (accessed May 24, 2018).
- [95] Schulz KF, Altman DG, Moher D. CONSORT 2010 Statement: Updated guidelines for reporting parallel group randomised trials. *J Clin Epidemiol* 2010;63:834–40. doi:10.1016/j.jclinepi.2010.02.005.
- [96] Selph SS, Ginsburg AD, Chou R. Impact of contacting study authors to obtain additional data for systematic reviews: diagnostic accuracy studies for hepatic fibrosis. *Syst Rev* 2014;3:107. doi:10.1186/2046-4053-3-107.
- [97] StataCorp. Stata Statistical Software: Release 14 2015.
- [98] Mehta CR, Patel NR. Exact logistic regression: theory and examples. *Stat Med* 1995;14:2143–60.
- [99] Kahan BC, Morris TP. Improper analysis of trials randomised using stratified blocks or minimisation. *Stat Med* 2012;31:328–40. doi:10.1002/sim.4431.
- [100] Kahan BC, Morris TP. Reporting and analysis of trials using stratified randomisation in leading medical journals: review and reanalysis. *BMJ* 2012;345:e5840–e5840. doi:10.1136/bmj.e5840.
- [101] Dillman DA, Phelps G, Tortora R, Swift K, Kohrell J, Berck J, et al. Response rate and measurement differences in mixed-mode surveys using mail, telephone, interactive voice response (IVR) and the Internet. *Soc Sci Res* 2009;38:1–18. doi:10.1016/j.ssresearch.2008.03.007.

- [102] Higgins J, Soornro M, Roberts I, Clarke M. Collecting unpublished data for systematic reviews: a proposal for a randomised trial. 7th Annu. Cochrane Colloq., Rome: n.d.
- [103] Tsafnat G, Glasziou P, Karystianis G, Coiera E. Automated screening of research studies for systematic reviews using study characteristics. *Syst Rev* 2018;7:64. doi:10.1186/s13643-018-0724-7.
- [104] Siebenhofer A, Paulitsch MA, Pregartner G, Berghold A, Jeitler K, Muth C, et al. Cluster-randomized controlled trials evaluating complex interventions in general practices are mostly ineffective: a systematic review. *J Clin Epidemiol* 2018;94:85–96. doi:10.1016/j.jclinepi.2017.10.010.
- [105] Eldridge SM, Ashby D, Feder GS, Rudnicka AR, Ukoumunne OC. Lessons for cluster randomized trials in the twenty-first century: a systematic review of trials in primary care. *Clin Trials J Soc Clin Trials* 2004;1:80–90. doi:10.1191/1740774504cn006rr.
- [106] Cleveringa FGW, Gorter KJ, van den Donk M, Rutten GEHM. Combined task delegation, computerized decision support, and feedback improve cardiovascular risk for type 2 diabetic patients: a cluster randomized trial in primary care. *Diabetes Care* 2008;31:2273–5. doi:10.2337/dc08-0312.
- [107] van Bruggen R, Gorter KJ, Stolk RP, Verhoeven RP, Rutten GEHM. Implementation of locally adapted guidelines on type 2 diabetes. *Fam Pract* 2008;25:430–7. doi:10.1093/fampra/cmn045.
- [108] Grant RW, Wald JS, Schnipper JL, Gandhi TK, Poon EG, Orav EJ, et al. Practice-linked online personal health records for type 2 diabetes mellitus: a randomized controlled trial. *Arch Intern Med* 2008;168:1776–82. doi:10.1001/archinte.168.16.1776.
- [109] Smith SA, Shah ND, Bryant SC, Christianson TJH, Bjornsen SS, Giesler PD, et al. Chronic care model and shared care in diabetes: randomized trial of an electronic decision support system. *Mayo Clin Proc* 2008;83:747–57. doi:10.4065/83.7.747.
- [110] Thomas KG, Thomas MR, Stroebel RJ, McDonald FS, Hanson GJ, Naessens JM, et al.

- Use of a registry-generated audit, feedback, and patient reminder intervention in an internal medicine resident clinic--a randomized trial. *J Gen Intern Med* 2007;22:1740–4. doi:10.1007/s11606-007-0431-x.
- [111] Piatt GA, Orchard TJ, Emerson S, Simmons D, Songer TJ, Brooks MM, et al. Translating the chronic care model into the community: Results from a randomized controlled trial of a multifaceted diabetes care intervention. *Diabetes Care* 2006;29:811–7.
- [112] Phillips LS, Ziemer DC, Doyle JP, Barnes CS, Kolm P, Branch WT, et al. An endocrinologist-supported intervention aimed at providers improves diabetes management in a primary care site: improving primary care of African Americans with diabetes (IPCAAD) 7. *Diabetes Care* 2005;28:2352–60.
- [113] Harris SB, Leiter LA, Webster-Bogaert S, Van DM, O'Neill C. Teleconferenced educational detailing: diabetes education for primary care physicians. *J Contin Educ Health Prof* 2005;25:87–97. doi:10.1002/chp.13.
- [114] Dijkstra RF, Braspenning JCC, Huijsmans Z, Akkermans RP, van Ballegooie E, ten Have P, et al. Introduction of diabetes passports involving both patients and professionals to improve hospital outpatient diabetes care. *Diabetes Res Clin Pract* 2005;68:126–34. doi:10.1016/j.diabres.2004.09.020.
- [115] Glasgow RE, Nutting PA, King DK, Nelson CC, Cutter G, Gaglio B, et al. Randomized effectiveness trial of a computer-assisted intervention to improve diabetes care. *Diabetes Care* 2005;28:33–9.
- [116] Olivarius NF, Beck-Nielsen H, Andreasen AH, Hørder M, Pedersen PA. Randomised controlled trial of structured personal care of type 2 diabetes mellitus. *BMJ* 2001;323:970–5.
- [117] Smith S, Bury G, O'Leary M, Shannon W, Tynan A, Staines A, et al. The North Dublin randomized controlled trial of structured diabetes shared care. *Fam Pract* 2004;21:39–45.
- [118] Krass I, Armour CL, Mitchell B, Brilliant M, Denaar R, Hughes J, et al. The Pharmacy Diabetes Care Program: assessment of a community pharmacy diabetes service

- model in Australia. *Diabet Med* 2007;24:677–83. doi:10.1111/j.1464-5491.2007.02143.x.
- [119] O’Hare JP, Raymond NT, Mughal S, Dodd L, Hanif W, Ahmad Y, et al. Evaluation of delivery of enhanced diabetes care to patients of South Asian ethnicity: the United Kingdom Asian Diabetes Study (UKADS). *Diabet Med* 2004;21:1357–65. doi:10.1111/j.1464-5491.2004.01373.x.
- [120] Eccles MP, Whitty PM, Speed C, Steen IN, Vanoli A, Hawthorne GC, et al. A pragmatic cluster randomised controlled trial of a Diabetes REcall And Management system: the DREAM trial. *Implement Sci* 2007;2:6. doi:10.1186/1748-5908-2-6.
- [121] O’Connor PJ, Sperl-Hillen J, Johnson PE, Rush WA, Crain AL. Customized feedback to patients and providers failed to improve safety or quality of diabetes care: a randomized trial. *Diabetes Care* 2009;32:1158–63. doi:10.2337/dc08-2247.
- [122] Goderis G, Borgermans L, Grol R, Van Den Broeke C, Boland B, Verbeke G, et al. Start improving the quality of care for people with type 2 diabetes through a general practice support program: a cluster randomized trial. *Diabetes Res Clin Pract* 2010;88:56–64. doi:10.1016/j.diabres.2009.12.012.
- [123] Simmons D, Gamble GD, Foote S, Cole DR, Coster G. The New Zealand Diabetes Passport Study: a randomized controlled trial of the impact of a diabetes passport on risk factors for diabetes-related complications. *Diabet Med* 2004;21:214–7.
- [124] Barceló A, Cafiero E, de Boer M, Mesa AE, Lopez MG, Jimenez RA, et al. Using collaborative learning to improve diabetes care and outcomes: the VIDA project. *Prim Care Diabetes* 2010;4:145–53. doi:10.1016/j.pcd.2010.04.005.
- [125] MacLean CD, Gagnon M, Callas P, Littenberg B. The Vermont Diabetes Information System: A Cluster Randomized Trial of a Population Based Decision Support System. *J Gen Intern Med* 2009;24:1303–10. doi:10.1007/s11606-009-1147-x.
- [126] Heisler M, Hofer TP, Schmittdiel JA, Selby J V, Klamerus ML, Bosworth HB, et al. Improving blood pressure control through a clinical pharmacist outreach program in patients with diabetes mellitus in 2 high-performing health systems: the

- adherence and intensification of medications cluster randomized, controlled pragmatic trial. *Circulation* 2012;125:2863–72.
doi:10.1161/CIRCULATIONAHA.111.089169.
- [127] Groeneveld Y, Petri H, Hermans J, Springer M. An assessment of structured care assistance in the management of patients with type 2 diabetes in general practice. *Scand J Prim Health Care* 2001;19:25–30.
 - [128] Quinn CC, Shardell MD, Terrin ML, Barr EA, Ballew SH, Gruber-Baldini AL. Cluster-randomized trial of a mobile phone personalized behavioral intervention for blood glucose control. *Diabetes Care* 2011;34:1934–42. doi:10.2337/dc11-0366.
 - [129] Estrada CA, Safford MM, Salanitro AH, Houston TK, Curry W, Williams JH, et al. A web-based diabetes intervention for physician: a cluster-randomized effectiveness trial. *Int J Qual Heal Care J Int Soc Qual Heal Care* 2011;23:682–9.
doi:10.1093/intqhc/mzr053.
 - [130] Mehuys E, Van Bortel L, De Bolle L, Van Tongelen I, Annemans L, Remon J-P, et al. Effectiveness of a community pharmacist intervention in diabetes care: a randomized controlled trial. *J Clin Pharm Ther* 2011;36:602–13.
doi:10.1111/j.1365-2710.2010.01218.x.
 - [131] Sönnichsen AC, Winkler H, Flamm M, Panisch S, Kowatsch P, Klima G, et al. The effectiveness of the Austrian disease management programme for type 2 diabetes: a cluster-randomised controlled trial. *BMC Fam Pract* 2010;11:86.
doi:10.1186/1471-2296-11-86.
 - [132] Guldborg TL, Vedsted P, Kristensen JK, Lauritzen T. Improved quality of Type 2 diabetes care following electronic feedback of treatment status to general practitioners: a cluster randomized controlled trial. *Diabet Med* 2011;28:325–32.
doi:10.1111/j.1464-5491.2010.03178.x.
 - [133] Polonsky WH, Fisher L, Schikman CH, Hinnen DA, Parkin CG, Jelsovsky Z, et al. Structured self-monitoring of blood glucose significantly reduces A1C levels in poorly controlled, noninsulin-treated type 2 diabetes: results from the Structured Testing Program study. *Diabetes Care* 2011;34:262–7. doi:10.2337/dc10-1732.

- [134] O'Connor PJ, Sperl-Hillen JM, Rush WA, Johnson PE, Amundson GH, Asche SE, et al. Impact of Electronic Health Record Clinical Decision Support on Diabetes Care: A Randomized Trial. *Ann Fam Med* 2011;9:12–21. doi:10.1370/afm.1196.
- [135] Pape GA, Hunt JS, Butler KL, Siemieniczuk J, LeBlanc BH, Gillanders W, et al. Team-based care approach to cholesterol management in diabetes mellitus: two-year cluster randomized controlled trial. *Arch Intern Med* 2011;171:1480–6. doi:10.1001/archinternmed.2011.417.
- [136] Griffin SJ, Borch-Johnsen K, Davies MJ, Khunti K, Rutten GEHM, Sandbaek A, et al. Effect of early intensive multifactorial therapy on 5-year cardiovascular outcomes in individuals with type 2 diabetes detected by screening (ADDITION-Europe): a cluster-randomised trial. *Lancet (London, England)* 2011;378:156–67. doi:10.1016/S0140-6736(11)60698-3.
- [137] Herrin J, Nicewander DA, Hollander PA, Couch CE, Winter FD, Haydar ZR, et al. Effectiveness of diabetes resource nurse case management and physician profiling in a fee-for-service setting: a cluster randomized trial. *Proc (Bayl Univ Med Cent)* 2006;19:95–102.
- [138] Hetlevik I, Holmen J, Krüger O, Kristensen P, Iversen H, Furuseth K. Implementing clinical guidelines in the treatment of diabetes mellitus in general practice. Evaluation of effort, process, and patient outcome related to implementation of a computer-based decision support system. *Int J Technol Assess Health Care* 2000;16:210–27.
- [139] Hermans MP, Elisaf M, Michel G, Muls E, Nobels F, Vandenberghe H, et al. Benchmarking is associated with improved quality of care in type 2 diabetes: the OPTIMISE randomized, controlled trial. *Diabetes Care* 2013;36:3388–95. doi:10.2337/dc12-1853.
- [140] Wolf MS, Seligman H, Davis TC, Fleming DA, Curtis LM, Pandit AU, et al. Clinic-based versus outsourced implementation of a diabetes health literacy intervention. *J Gen Intern Med* 2014;29:59–67. doi:10.1007/s11606-013-2582-2.
- [141] Morgan MAJ, Coates MJ, Dunbar JA, Reddy P, Schlicht K, Fuller J. The TrueBlue

- model of collaborative care using practice nurses as case managers for depression alongside diabetes or heart disease: a randomised trial. *BMJ Open* 2013;3. doi:10.1136/bmjopen-2012-002171.
- [142] Steyn K, Lombard C, Gwebushe N, Fourie JM, Everett-Murphy K, Zwarenstein M, et al. Implementation of national guidelines, incorporated within structured diabetes and hypertension records at primary level care in Cape Town, South Africa: a randomised controlled trial. *Glob Health Action* 2013;6:20796. doi:10.3402/gha.v6i0.20796.
- [143] Blackberry ID, Furler JS, Best JD, Chondros P, Vale M, Walker C, et al. Effectiveness of general practice based, practice nurse led telephone coaching on glycaemic control of type 2 diabetes: the Patient Engagement And Coaching for Health (PEACH) pragmatic cluster randomised controlled trial. *BMJ* 2013;347:f5272–f5272. doi:10.1136/bmj.f5272.
- [144] Vaccaro O, Franzini L, Miccoli R, Cavalot F, Ardigò D, Boemi M, et al. Feasibility and effectiveness in clinical practice of a multifactorial intervention for the reduction of cardiovascular risk in patients with type 2 diabetes: the 2-year interim analysis of the MIND.IT study: a cluster randomized trial. *Diabetes Care* 2013;36:2566–72. doi:10.2337/dc12-1781.
- [145] Gagliardino JJ, Lapertosa S, Pfirter G, Villagra M, Caporale JE, Gonzalez CD, et al. Clinical, metabolic and psychological outcomes and treatment costs of a prospective randomized trial based on different educational strategies to improve diabetes care (PRODIACOR). *Diabet Med* 2013;30:1102–11. doi:10.1111/dme.12230.
- [146] DePue JD, Dunsiger S, Seiden AD, Blume J, Rosen RK, Goldstein MG, et al. Nurse-community health worker team improves diabetes care in American Samoa: results of a randomized controlled trial. *Diabetes Care* 2013;36:1947–53. doi:10.2337/dc12-1969.
- [147] Adachi M, Yamaoka K, Watanabe M, Nishikawa M, Kobayashi I, Hida E, et al. Effects of lifestyle education program for type 2 diabetes patients in clinics: a

- cluster randomized controlled trial. *BMC Public Health* 2013;13:467.
doi:10.1186/1471-2458-13-467.
- [148] Jansink R, Braspenning J, Keizer E, van der Weijden T, Elwyn G, Grol R. No identifiable Hb1Ac or lifestyle change after a comprehensive diabetes programme including motivational interviewing: a cluster randomised trial. *Scand J Prim Health Care* 2013;31:119–27. doi:10.3109/02813432.2013.797178.
- [149] Kinmonth AL, Woodcock A, Griffin S, Spiegel N, Campbell MJ. Randomised controlled trial of patient centred care of diabetes in general practice: impact on current wellbeing and future disease risk. The Diabetes Care From Diagnosis Research Team. *BMJ* 1998;317:1202–8.
- [150] Sieber W, Newsome A, Lillie D. Promoting self-management in diabetes: efficacy of a collaborative care approach. *Fam Syst Health* 2012;30:322–9.
doi:10.1037/a0030714.
- [151] Saenz A, Brito M, Moron I, Torralba A, Garcia-Sanz E, Redondo J. Development and validation of a computer application to aid the physician’s decision-making process at the start of and during treatment with insulin in type 2 diabetes: a randomized and controlled trial. *J Diabetes Sci Technol* 2012;6:581–8.
doi:10.1177/193229681200600313.
- [152] Juul L, Maindal HT, Zoffmann V, Frydenberg M, Sandbaek A. Effectiveness of a training course for general practice nurses in motivation support in type 2 diabetes care: a cluster-randomised trial. *PLoS One* 2014;9:e96683.
doi:10.1371/journal.pone.0096683.
- [153] Frei A, Senn O, Chmiel C, Reissner J, Held U, Rosemann T. Implementation of the chronic care model in small medical practices improves cardiovascular risk but not glycemic control. *Diabetes Care* 2014;37:1039–47. doi:10.2337/dc13-1429.
- [154] Steventon A, Bardsley M, Doll H, Tuckey E, Newman SP. Effect of telehealth on glycaemic control: analysis of patients with type 2 diabetes in the Whole Systems Demonstrator cluster randomised trial. *BMC Health Serv Res* 2014;14:334.
doi:10.1186/1472-6963-14-334.

- [155] Frijling BD, Lobo CM, Hulscher MEJL, Akkermans RP, Braspenning JCC, Prins A, et al. Multifaceted support to improve clinical decision making in diabetes care: a randomized controlled trial in general practice. *Diabet Med* 2002;19:836–42.
- [156] Kiefe CI, Allison JJ, Williams OD, Person SD, Weaver MT, Weissman NW. Improving quality improvement using achievable benchmarks for physician feedback: a randomized controlled trial. *JAMA* 2001;285:2871–9.
- [157] McDermott RA, Schmidt BA, Sinha A, Mills P. Improving diabetes care in the primary healthcare setting: a randomised cluster trial in remote Indigenous communities. *Med J Aust* 2001;174:497–502.
- [158] Ward A, Kamien M, Mansfield F, Fataavitch B. Educational feedback in the management of type 2 diabetes in general practice. *Educ Gen Pr* 1996;7:142–50.
- [159] Persell SD, Denecke-Dattalo TA, Dunham DP, Baker DW. Patient-directed intervention versus clinician reminders alone to improve aspirin use in diabetes: a cluster randomized trial. *Jt Comm J Qual Patient Saf* 2008;34:98–105.
- [160] Kulkarni K, Castle G, Gregory R, Holmes A, Leontos C, Powers M, et al. Nutrition practice guidelines for type 1 diabetes mellitus positively affect dietitian practices and patient outcomes. The Diabetes Care and Education Dietetic Practice Group. *J Am Diet Assoc* 1998;98:62-70; quiz 71-2.
- [161] Dijkstra R, Braspenning J, Grol R. Implementing diabetes passports to focus practice reorganization on improving diabetes care. *Int J Qual Health Care* 2008;20:72–7. doi:10.1093/intqhc/mzm051.
- [162] New JP, Mason JM, Freemantle N, Teasdale S, Wong L, Bruce NJ, et al. Educational outreach in diabetes to encourage practice nurses to use primary care hypertension and hyperlipidaemia guidelines (EDEN): a randomized controlled trial. *Diabet Med* 2004;21:599–603. doi:10.1111/j.1464-5491.2004.01213.x.
- [163] McClellan WM, Millman L, Presley R, Couzins J, Flanders WD. Improved diabetes care by primary care physicians: results of a group-randomized evaluation of the Medicare Health Care Quality Improvement Program (HCQIP). *J Clin Epidemiol* 2003;56:1210–7.

- [164] Ilag LL, Martin CL, Tabaei BP, Isaman DJM, Burke R, Greene DA, et al. Improving diabetes processes of care in managed care. *Diabetes Care* 2003;26:2722–7.
- [165] Bebb C, Kendrick D, Coupland C, Madeley R, Stewart J, Brown K, et al. A cluster randomised controlled trial of the effect of a treatment algorithm for hypertension in patients with type 2 diabetes. *Br J Gen Pract* 2007;57:136–43.
- [166] Perria C, Mandolini D, Guerrera C, Jefferson T, Billi P, Calzini V, et al. Implementing a guideline for the treatment of type 2 diabetics: results of a cluster-randomized controlled trial (C-RCT). *BMC Health Serv Res* 2007;7:79. doi:10.1186/1472-6963-7-79.
- [167] Reiber GE, Au D, McDonell M, Fihn SD. Diabetes quality improvement in Department of Veterans Affairs Ambulatory Care Clinics: a group-randomized clinical trial. *Diabetes Care* 2004;27 Suppl 2:B61-8.
- [168] Kirwin JL, Cunningham RJ, Sequist TD. Pharmacist recommendations to improve the quality of diabetes care: a randomized controlled trial. *J Manag Care Pharm* 2010;16:104–13. doi:10.18553/jmcp.2010.16.2.104.
- [169] Sequist TD, Fitzmaurice GM, Marshall R, Shaykevich S, Marston A, Safran DG, et al. Cultural competency training and performance reports to improve diabetes care for black patients: a cluster randomized, controlled trial. *Ann Intern Med* 2010;152:40–6. doi:10.7326/0003-4819-152-1-201001050-00009.
- [170] Schnipper JL, Linder JA, Palchuk MB, Yu DT, McColgan KE, Volk LA, et al. Effects of documentation-based decision support on chronic disease management. *Am J Manag Care* 2010;16:SP72-81.
- [171] Meigs JB, Cagliero E, Dubey A, Murphy-Sheehy P, Gildesgame C, Chueh H, et al. A controlled trial of web-based diabetes disease management: the MGH diabetes primary care improvement project. *Diabetes Care* 2003;26:750–7.
- [172] Shah BR, Bhattacharyya O, Yu CHY, Mamdani MM, Parsons JA, Straus SE, et al. Effect of an Educational Toolkit on Quality of Care: A Pragmatic Cluster Randomized Trial. *PLoS Med* 2014;11:e1001588. doi:10.1371/journal.pmed.1001588.

- [173] Dickinson WP, Dickinson LM, Nutting PA, Emsermann CB, Tutt B, Crabtree BF, et al. Practice Facilitation to Improve Diabetes Care in Primary Care: A Report From the EPIC Randomized Clinical Trial. *Ann Fam Med* 2014;12:8–16. doi:10.1370/afm.1591.
- [174] Vidal-Pardo JI, Pérez-Castro TR, López-Álvarez XL, Santiago-Pérez MI, García-Soidán FJ, Muñiz J. Effect of an educational intervention in primary care physicians on the compliance of indicators of good clinical practice in the treatment of type 2 diabetes mellitus [OBTEDIGA project]. *Int J Clin Pract* 2013;67:750–8. doi:10.1111/ijcp.12145.
- [175] Wagner EH, Grothaus LC, Sandhu N, Galvin MS, McGregor M, Artz K, et al. Chronic care clinics for diabetes in primary care: a system-wide randomized trial. *Diabetes Care* 2001;24:695–700.
- [176] Peterson KA, Radosevich DM, O'Connor PJ, Nyman JA, Prineas RJ, Smith SA, et al. Improving diabetes care in practice: findings from the TRANSLATE trial. *Diabetes Care* 2008;31:2238–43. doi:10.2337/dc08-2034.
- [177] Diaz-Ordaz K, Froud R, Sheehan B, Eldridge S. A systematic review of cluster randomised trials in residential facilities for older people suggests how to improve quality. *BMC Med Res Methodol* 2013;13:127. doi:10.1186/1471-2288-13-127.
- [178] Ivers NM, Halperin IJ, Barnsley J, Grimshaw JM, Shah BR, Tu K, et al. Allocation techniques for balance at baseline in cluster randomized trials: a methodological review. *Trials* 2012;13:120. doi:10.1186/1745-6215-13-120.
- [179] Wright N, Ivers N, Eldridge S, Taljaard M, Bremner S. A review of the use of covariates in cluster randomized trials uncovers marked discrepancies between guidance and practice. *J Clin Epidemiol* 2015;68:603–9. doi:10.1016/j.jclinepi.2014.12.006.
- [180] Campbell MK, Piaggio G, Elbourne DR, Altman DG. Consort 2010 statement: extension to cluster randomised trials. *BMJ* 2012;345:e5661. doi:10.1136/bmj.e5661.
- [181] Donner A, Klar N. Issues in the meta-analysis of cluster randomized trials. *Stat*

- Med 2002;21:2971–80. doi:10.1002/sim.1301.
- [182] Campbell MK, Fayers PM, Grimshaw JM. Determinants of the intracluster correlation coefficient in cluster randomized trials: the case of implementation research. *Clin Trials* 2005;2:99–107. doi:10.1191/1740774505cn071oa.
 - [183] Turner RM, Toby Prevoost A, Thompson SG. Allowing for imprecision of the intracluster correlation coefficient in the design of cluster randomized trials. *Stat Med* 2004;23:1195–214. doi:10.1002/sim.1721.
 - [184] Spiegelhalter DJ. Bayesian methods for cluster randomized trials with continuous responses. *Stat Med* 2001;20:435–52.
 - [185] Turner RM, Omar RZ, Thompson SG. Bayesian methods of analysis for cluster randomized trials with binary outcome data. *Stat Med* 2001;20:453–72.
 - [186] Danko K, Taljaard M, Sullivan K, Ivers N, Grimshaw J. How pervasive are unit of analysis errors in cluster randomized trials? A review of diabetes quality improvement RCTs. *Glob. Evid. Summit*, Cape Town, South Africa: 2017.
 - [187] Spiegelhalter D, Abrams K, Myles J. Bayesian approaches to clinical trials and health-care evaluation. West Sussex: John Wiley & Sons Ltd; 2004.
 - [188] Swiger LA, Harvey WR, Everson DO, Gregory KE. The variance of intraclass correlation involving groups with one observation. *Biometrics* 1964;20:818. doi:10.2307/2528131.
 - [189] Higgins JPT, Thompson SG, Spiegelhalter DJ. A re-evaluation of random-effects meta-analysis. *J R Stat Soc Ser A Stat Soc* 2009;172:137–59. doi:10.1111/j.1467-985X.2008.00552.x.
 - [190] Plummer M. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. 2003.
 - [191] Brooks SP, Gelman A. General methods for monitoring convergence of iterative simulations. *J Comput Graph Stat* 1998;7:434–55. doi:10.1080/10618600.1998.10474787.
 - [192] Danko K, Dahabreh I, Trikalinos T, Ivers N, Grimshaw J. Making sense of complex interventions: Application of hierarchical multivariate meta- regression to explore

- heterogeneity in diabetes quality improvement interventions International. Conf. Heal. Policy Stat., Charleston, South Carolina: 2018.
- [193] Taljaard M, Donner A, Villar J, Wojdyla D, Velazco A, Bataglia V, et al. Intraclass correlation coefficients from the 2005 WHO Global Survey on Maternal and Perinatal Health: implications for implementation research. *Paediatr Perinat Epidemiol* 2008;22:117–25. doi:10.1111/j.1365-3016.2007.00901.x.
 - [194] Cook JA, Bruckner T, MacLennan GS, Seiler CM. Clustering in surgical trials--database of intraclass correlations. *Trials* 2012;13:2. doi:10.1186/1745-6215-13-2.
 - [195] Pals SL, Beaty BL, Posner SF, Bull SS. Estimates of intraclass correlation for variables related to behavioral HIV/STD prevention in a predominantly African American and Hispanic sample of young women. *Health Educ Behav* 2009;36:182–94. doi:10.1177/1090198108327731.
 - [196] Janjua NZ, Khan MI, Clemens JD. Estimates of intraclass correlation coefficient and design effect for surveys and cluster randomized trials on injection use in Pakistan and developing countries. *Trop Med Int Health* 2006;11:1832–40. doi:10.1111/j.1365-3156.2006.01736.x.
 - [197] Hade EM, Murray DM, Pennell ML, Rhoda D, Paskett ED, Champion VL, et al. Intraclass correlation estimates for cancer screening outcomes: estimates and applications in the design of group-randomized cancer screening studies. *J Natl Cancer Inst Monogr* 2010;2010:97–103. doi:10.1093/jncimonographs/lgq011.
 - [198] Frank B, Fulton RL, Goldie FC, Hacke W, Weimar C, Lees KR. Intraclass correlation coefficients and reliability of randomized multicenter stroke trials within VISTA. *Int J Stroke* 2014;9:602–6. doi:10.1111/ijss.12123.
 - [199] Lajos GJ, Haddad SM, Tedesco RP, Passini Jr R, Dias TZ, Nomura ML, et al. Intraclass correlation coefficients for the Brazilian Multicenter Study on Preterm Birth (EMIP): methodological and practical implications. *BMC Med Res Methodol* 2014;14:54. doi:10.1186/1471-2288-14-54.
 - [200] Pagel C, Prost A, Lewycka S, Das S, Colbourn T, Mahapatra R, et al. Intraclass

- correlation coefficients and coefficients of variation for perinatal outcomes from five cluster-randomised controlled trials in low and middle-income countries: results and methodological implications. *Trials* 2011;12:151. doi:10.1186/1745-6215-12-151.
- [201] Gulliford MC, Adams G, Ukoumunne OC, Latinovic R, Chinn S, Campbell MJ. Intraclass correlation coefficient and outcome prevalence are associated in clustered binary data. *J Clin Epidemiol* 2005;58:246–51. doi:10.1016/j.jclinepi.2004.08.012.
- [202] Adams G, Gulliford MC, Ukoumunne OC, Eldridge S, Chinn S, Campbell MJ. Patterns of intra-cluster correlation from primary care research to inform study design and analysis. *J Clin Epidemiol* 2004;57:785–94. doi:10.1016/j.jclinepi.2003.12.013.
- [203] Thompson DM, Fernald DH, Mold JW. Intraclass correlation coefficients typical of cluster-randomized studies: estimates from the Robert Wood Johnson Prescription for Health projects. *Ann Fam Med* 2012;10:235–40. doi:10.1370/afm.1347.
- [204] Smeeth L, Ng ES-W. Intraclass correlation coefficients for cluster randomized trials in primary care: data from the MRC Trial of the Assessment and Management of Older People in the Community. *Control Clin Trials* 2002;23:409–21.
- [205] Rowe AK, Lama M, Onikpo F, Deming MS. Design effects and intraclass correlation coefficients from a health facility cluster survey in Benin. *Int J Qual Heal Care J Int Soc Qual Heal Care* 2002;14:521–3.
- [206] Knox SA, Chondros P. Observed intra-cluster correlation coefficients in a cluster survey sample of patient encounters in general practice in Australia. *BMC Med Res Methodol* 2004;4:30. doi:10.1186/1471-2288-4-30.
- [207] Sedgwick JE, Pearce AJ, Gulliford MC. Evaluation of equity in diabetes health care in relation to African and Caribbean ethnicity. *Ethn Health* 2003;8:121–33. doi:10.1080/13557850303560.
- [208] Gulliford MC, Mahabir D. Relationship of health-related quality of life to symptom

- severity in diabetes mellitus: a study in Trinidad and Tobago. *J Clin Epidemiol* 1999;52:773–80.
- [209] Amorim LD, Bangdiwala SI, McMurray RG, Creighton D, Harrell J. Intraclass correlations among physiologic measures in children and adolescents. *Nurs Res* 2007;56:355–60. doi:10.1097/01.NNR.0000289497.91918.94.
 - [210] Fahey TP, Peters TJ. What constitutes controlled hypertension? Patient based comparison of hypertension guidelines. *BMJ* 1996;313:93–6.
 - [211] Parker DR, Evangelou E, Eaton CB. Intraclass correlation coefficients for cluster randomized trials in primary care: the cholesterol education and research trial (CEART). *Contemp Clin Trials* 2005;26:260–7. doi:10.1016/j.cct.2005.01.002.
 - [212] Elley CR, Kerse N, Chondros P, Robinson E. Intraclass correlation coefficients from three cluster randomised controlled trials in primary and residential health care. *Aust N Z J Public Health* 2005;29:461–7.
 - [213] Feder G, Griffiths C, Eldridge S, Spence M. Effect of postal prompts to patients and general practitioners on the quality of primary care after a coronary event (POST): randomised controlled trial. *BMJ* 1999;318:1522–6.
 - [214] Littenberg B, MacLean CD. Intra-cluster correlation coefficients in adults with diabetes in primary care practices: the Vermont Diabetes Information System field survey. *BMC Med Res Methodol* 2006;6:20. doi:10.1186/1471-2288-6-20.
 - [215] Health Services Research Unit. Database of intra-correlation coefficients (ICCs). Excel Sheet n.d. <https://www.abdn.ac.uk/hsru/documents/iccs-web.xls> (accessed August 8, 2017).
 - [216] Geyer J, Davis M, Narayan T. Intraclass Correlation Coefficients of Household Economic and Agricultural Outcomes in Mozambique. *Eval Rev* 2016. doi:10.1177/0193841X16659834.
 - [217] Craig P, Petticrew M. Developing and evaluating complex interventions: reflections on the 2008 MRC guidance. *Int J Nurs Stud* 2013;50:585–7. doi:10.1016/j.ijnurstu.2012.09.009.
 - [218] Anderson LM, Oliver SR, Michie S, Rehfuss E, Noyes J, Shemilt I. Investigating

- complexity in systematic reviews of interventions by using a spectrum of methods. *J Clin Epidemiol* 2013;66:1223–9. doi:10.1016/j.jclinepi.2013.06.014.
- [219] International Diabetes Federation. Diabetes Atlas 2015. <http://www.idf.org/diabetesatlas> (accessed November 12, 2017).
- [220] Hawthorne G, Hrisos S, Stamp E, Elovainio M, Francis JJ, Grimshaw JM, et al. Diabetes care provision in UK primary care practices. *PLoS One* 2012;7:e41562. doi:10.1371/journal.pone.0041562.
- [221] Gelman A, Hill A. Data analysis using regression and multilevel/hierarchical models - University of Ottawa. New York: Cambridge University Press; 2009.
- [222] Michie S, Richardson M, Johnston M, Abraham C, Francis J, Hardeman W, et al. The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. *Ann Behav Med* 2013;46:81–95. doi:10.1007/s12160-013-9486-6.
- [223] Powell BJ, Waltz TJ, Chinman MJ, Damschroder LJ, Smith JL, Matthieu MM, et al. A refined compilation of implementation strategies: results from the Expert Recommendations for Implementing Change (ERIC) project. *Implement Sci* 2015;10:209. doi:10.1186/s13012-015-0209-1.
- [224] Cipriani A, Furukawa TA, Salanti G, Geddes JR, Higgins JP, Churchill R, et al. Comparative efficacy and acceptability of 12 new-generation antidepressants: a multiple-treatments meta-analysis. *Lancet* 2009;373:746–58. doi:10.1016/S0140-6736(09)60046-5.
- [225] Schwarzer G. meta: An R package for meta-analysis. *R News* 2007;7:40–5.
- [226] Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A. Bayesian measures of model complexity and fit. *J R Stat Soc Ser B (Statistical Methodol)* 2002;64:583–639. doi:10.1111/1467-9868.00353.
- [227] Marshall EC, Spiegelhalter DJ. Approximate cross-validators predictive checks in disease mapping models. *Stat Med* 2003;22:1649–60. doi:10.1002/sim.1403.
- [228] Sutton AJ, Abrams KR. Bayesian methods in meta-analysis and evidence synthesis.

- Stat Methods Med Res 2001;10:277–303. doi:10.1177/096228020101000404.
- [229] Rubin DB. Meta-Analysis: Literature synthesis or effect-size surface estimation? *J Educ Stat* 1992;17:363–74. doi:10.2307/1165129.
- [230] Bragge P, Grimshaw JM, Lokker C, Colquhoun H, AIMD Writing/Working Group. AIMD - a validated, simplified framework of interventions to promote and integrate evidence into health practices, systems, and policies. *BMC Med Res Methodol* 2017;17:38. doi:10.1186/s12874-017-0314-8.
- [231] Hoffmann TC, Glasziou PP, Boutron I, Milne R, Perera R, Moher D, et al. Better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide. *BMJ* 2014;348:g1687.
- [232] Pinnock H, Epiphaniou E, Sheikh A, Griffiths C, Eldridge S, Craig P, et al. Developing standards for reporting implementation studies of complex interventions (StaRI): a systematic review and e-Delphi. *Implement Sci* 2015;10:42. doi:10.1186/s13012-015-0235-z.
- [233] Hunink M, Weinstein M, Wittenberg E, MF D, Pliskin J, Wong J, et al. Decision making in health and medicine: integrating evidence and values. Cambridge University Press; n.d.
- [234] Kuhn TS. *The Structure of Scientific Revolutions*. 4th editio. London: The University of Chicago Press; 1962.
- [235] Ioannidis JPA, Fanelli D, Dunne DD, Goodman SN. Meta-research: Evaluation and improvement of research methods and practices. *PLOS Biol* 2015;13:e1002264. doi:10.1371/journal.pbio.1002264.
- [236] Pesseau J, Ivers NM, Newham JJ, Knittle K, Danko KJ, Grimshaw JM. Using a behaviour change techniques taxonomy to identify active ingredients within trials of implementation interventions for diabetes care. *Implement Sci* 2015;10:55. doi:10.1186/s13012-015-0248-7.
- [237] Elliott JH, Synnot A, Turner T, Simmonds M, Akl EA, McDonald S, et al. Living systematic review: 1. Introduction—the why, what, when, and how. *J Clin Epidemiol* 2017;91:23–30. doi:10.1016/j.jclinepi.2017.08.010.

- [238] Karunananthan K, Sullivan K, Danko K, Ivers N, Dahabreh I, Sundaramoorthy J, et al. Transitioning to Living Systematic Reviews: Lessons learned from a large scale review on diabetes quality improvement interventions. Glob. Evid. Summit, Cape Town, South Africa: 2017.
- [239] Hozo SP, Djulbegovic B, Hozo I. Estimating the mean and variance from the median, range, and the size of a sample. BMC Med Res Methodol 2005;5:13. doi:10.1186/1471-2288-5-13.
- [240] McKibbin KA, Lokker C, Wilczynski NL, Ciliska D, Dobbins M, Davis DA, et al. A cross-sectional study of the number and frequency of terms used to refer to knowledge translation in a body of health literature in 2006: a Tower of Babel? Implement Sci 2010;5:16. doi:10.1186/1748-5908-5-16.

APPENDIX A: Certificate of ethics approval

Manuscript 1: Contacting non-responding authors of studies included in systematic reviews by telephone increases response rates compared to repeat emailing: results of a randomized controlled trial

- Ottawa Health Science Network Research Ethics Board (Protocol ID: 20180429-01H)

APPENDIX B: Supplemental information for Manuscript 1

Supplemental Table B-1 Review team and knowledge users

Systematic review team

Jeremy M. Grimshaw (Co-PI)	Senior Scientist, Clinical Epidemiology Program, Ottawa Hospital Research Institute. Full Professor, Department of Medicine, University of Ottawa. Director of the Centre for Implementation Research. Ontario, Canada
Noah M. Ivers (Co-PI)	Family Physician and Scientist, Women's College Hospital. Assistant Professor, Department of Family and Community Medicine, University of Toronto. Innovation Fellow, Women's College Health Institute for Health System Solutions and Virtual Care. Ontario, Canada
Issa J. Dahabreh	Assistant Professor of Health Services, Policy & Practice and Epidemiology, Brown University. Rhode Island, United States
Kristin J. Danko	Research Coordinator, Clinical Epidemiology Program, Ottawa Hospital Research Institute. PhD Candidate, School of Epidemiology and Public Health, University of Ottawa. Ontario, Canada
John N. Lavis	Director, McMaster Health Forum. Professor, Department of Health Evidence and Impact. Ontario, Canada
David Moher	Senior Scientist, Ottawa Hospital Research Institute. Associate Professor, School of Epidemiology and Public Health, University of Ottawa. Ontario, Canada
Kaveh G. Shojania	Scientist, Evaluative Clinical Sciences, Veterans & Community Research Program, Sunnybrook Research Institute. Director, Centre for Quality Improvement & Patient Safety, University of Toronto (U of T). Vice-chair, Quality and Innovation, Department of Medicine, U of T. Full professor, Department of Medicine, U of T. Adjunct Faculty, Institute of Health Policy, Management and Evaluation, U of T. Staff Physician, General Internal Medicine, Sunnybrook Health Sciences Centre. Ontario, Canada
Timothy Ramsay	Scientist, Clinical Epidemiology Program, Ottawa Hospital Research Institute. Assistant Professor, Department of Epidemiology and Community Medicine, Faculty of Medicine, University of Ottawa. Ontario, Canada
Sharon E. Straus	Scientist, Keenan Research Centre of the Li Ka Shing Knowledge Institute, St. Michael's Hospital. Director, Knowledge Translation program, Li Ka Shing Knowledge Institute, University of Toronto.

Andrea C. Tricco	Ontario, Canada Scientist, Li Ka Shing Knowledge Institute, St. Michael's Hospital in Toronto. Ontario, Canada
Thomas A. Trikalinos	Associate Professor of Health Services, Policy & Practice; Director of the Center for Evidence Synthesis in Health at Brown University. Rhode Island, United States

Knowledge users

Alun Edwards	Professor at the University of Calgary in the Division of Endocrinology and Metabolism. Senior Medical Director for the Strategic Clinical Network for Diabetes, Obesity and Metabolism with Alberta Health Services. Alberta, Canada
Michael Hillmer	Executive Director at MOHLTC. Assistant Professor in the Institute of Health Policy, Management and Evaluation at the University of Toronto (formerly: Director of Planning, Research and Analysis Branch at Ministry of Health and Long-Term Care Ontario). Ontario, Canada.
Braden Manns	Professor, University of Calgary, Departments of Medicine & Community Health Sciences. President of the Canadian Society of Nephrology. Nephrologist University of Calgary in the Departments of Medicine and Community Health Sciences. Sware Professor in Health Economics. Principal Investigator, Canadian Kidney Disease Knowledge Translation Network. Alberta, Canada
Alison Paprica	Assistant Professor, Institute of Health Policy, Management and Evaluation, University of Toronto (formerly: Director of Planning, Research and Analysis Branch at Ministry of Health and Long-Term Care Ontario). Ontario, Canada
Peter M. Sargious	Associate Professor, Medicine, University of Calgary. Senior Medical Director of the Diabetes, Obesity and Nutrition Strategic Clinical Network, Alberta Health Services. Alberta, Canada
Marcello Tonelli	Professor University of Calgary, Division of Nephrology. Chair Emeritus of the Canadian Task Force on Preventive Health Care and past President of the Canadian Society of Nephrology. Alberta, Canada
Catherine H. Yu	Associate Scientist, Li Ka Shing Knowledge Institute, St. Michael's Hospital. Staff Endocrinologist, St. Michael's Hospital. Chair of the Dissemination and Implementation Committee of Diabetes Canada's Clinical Practice Guidelines. Ontario, Canada

Supplemental Table B-2 CONSORT 2010 checklist of information to include when reporting a randomised trial

Section/Topic	Item No	Checklist item	Reported on page No
Title and abstract			
	1a	Identification as a randomised trial in the title	23
	1b	Structured summary of trial design, methods, results, and conclusions	24
Introduction			
Background and objectives	2a	Scientific background and explanation of rationale	26-27
	2b	Specific objectives or hypotheses	27
Methods			
Trial design	3a	Description of trial design (such as parallel, factorial) including allocation ratio	27-28
	3b	Important changes to methods after trial commencement (such as eligibility criteria), with reasons	28
Participants	4a	Eligibility criteria for participants	28
	4b	Settings and locations where the data were collected	30
Interventions	5	The interventions for each group with sufficient details to allow replication, including how and when they were actually administered	29-30; Supp Text B-2
Outcomes	6a	Completely defined pre-specified primary and secondary outcome measures, including how and when they were assessed	31
	6b	Any changes to trial outcomes after the trial commenced, with reasons	n/a
Sample size	7a	How sample size was determined	28
	7b	When applicable, explanation of any interim analyses and stopping guidelines	n/a
Randomisation: Sequence generation	8a	Method used to generate the random allocation sequence	28; Supp Text B-1
	8b	Type of randomisation; details of any restriction (such as blocking and block size)	28; Supp Text B-1
Allocation concealment mechanism	9	Mechanism used to implement the random allocation sequence (such as sequentially numbered containers), describing any steps taken to conceal the sequence until interventions were assigned	28; Supp Text B-1
Implementation	10	Who generated the random allocation sequence, who enrolled participants, and who assigned participants to interventions	28
Blinding	11a	If done, who was blinded after assignment to interventions (for example, participants, care	30

		providers, those assessing outcomes) and how	
	11b	If relevant, description of the similarity of interventions	Supp Table B-3
Statistical methods	12a	Statistical methods used to compare groups for primary and secondary outcomes	31-32
	12b	Methods for additional analyses, such as subgroup analyses and adjusted analyses	31-32
Results			
Participant flow (a diagram is strongly recommended)	13a	For each group, the numbers of participants who were randomly assigned, received intended treatment, and were analysed for the primary outcome	32-36; Table 3; Fig 5
	13b	For each group, losses and exclusions after randomisation, together with reasons	32-36; Supp Fig B-1; Supp Text B-3
Recruitment	14a	Dates defining the periods of recruitment and follow-up	32; 35
	14b	Why the trial ended or was stopped	n/a
Baseline data	15	A table showing baseline demographic and clinical characteristics for each group	Table 2
Numbers analysed	16	For each group, number of participants (denominator) included in each analysis and whether the analysis was by original assigned groups	Table 3
Outcomes and estimation	17a	For each primary and secondary outcome, results for each group, and the estimated effect size and its precision (such as 95% confidence interval)	Tables 3-4
	17b	For binary outcomes, presentation of both absolute and relative effect sizes is recommended	Table 3
Ancillary analyses	18	Results of any other analyses performed, including subgroup analyses and adjusted analyses, distinguishing pre-specified from exploratory	37
Harms	19	All important harms or unintended effects in each group	34
Discussion			
Limitations	20	Trial limitations, addressing sources of potential bias, imprecision, and, if relevant, multiplicity of analyses	39-40
Generalisability	21	Generalisability (external validity, applicability) of the trial findings	38-41
Interpretation	22	Interpretation consistent with results, balancing benefits and harms, and considering other relevant evidence	38-41
Other information			
Registration	23	Registration number and name of trial registry	31
Protocol	24	Where the full trial protocol can be accessed, if available	31
Funding	25	Sources of funding and other support (such as supply of drugs), role of funders	41-43

Supplemental Text B-1 Randomization code

```
library (blockrand)

## stratified by decades, # in stratum=#in decades, 2 treatments(A=control;B=treatment)
A <- blockrand(n=2, block.sizes = 1:1, id.prefix='A', block.prefix='A',stratum='1980-89')
B <- blockrand(n=23, block.sizes = 1:2, id.prefix='B', block.prefix='B',stratum='1990-99')
C <- blockrand(n=97, id.prefix='C', block.prefix='C',stratum='2000-09')
D <- blockrand(n=53, id.prefix='D', block.prefix='D',stratum='2010-Present')
my.study <- rbind(A,B, C, D)
my.study

sample(c(0,1), 10, replace = TRUE)

A <- runif(2,min=0,max=1)
A
write.xlsx(A, ".xlsx")
B <- runif(23,min=0,max=1)
B
write.xlsx(B, ".xlsx")
C <- runif(97,min=0,max=1)
C
write.xlsx(C, ".xlsx")
D <- runif(53,min=0,max=1)
D
write.xlsx(D, ".xlsx")

library(xlsx)
write.xlsx(A, ".xlsx")
```

Comments on the randomization process:

The generated random numbers were linked to authors within the 4 decade blocks and sorted from smallest to greatest. Authors in the first half of the block were allocated to the email group while authors in the second half of the block were allocated to the call group. As most block sizes were uneven, an extra author was alternatively assigned across blocks (e.g., block B received 12 control vs. 11 intervention authors; block C received 48 control vs. 49 intervention authors).

Supplemental Text B-2 Additional details on telephone intervention

1. Searching for author phone numbers: One researcher searched for a phone number for the most recent contact author. If a phone number was not identified in the publication, the researcher conducted a Google search. If all searches for a number for the author failed, the researcher moved on to co-authors and repeated the search process until a number for a trial author was found. In cases where a number was not found for any trial author, the study was isolated from further contact. With respect to searches, we collected data on the time taken for the search (minutes), the success of the search (found number or not) and the path of the search (researcher notes). If additional searches were required after call Attempts 1 and 2 data, on these searches were also collected.
2. Telephone authors: The researcher made up to three attempts to reach the author, and when contact was achieved, used a semi-structured guide to direct the conversation. This included:
 - Researcher introduction (name, research institute, project)
 - Asked author if they recalled receiving the survey
 - If necessary, reminded them of the survey, its aim, what was involved
 - Asked the author if they would be interested in completing the survey
 - If yes, instructed that we would resend. Confirmed the most appropriate email address to send survey to
 - If not, asked if they would be willing to provide the phone number of a potentially suitable and willing co-investigator
 - If yes, got contact number for co-author to call in next round
 - If no, thanked author for their time and their study

With respect to calling, data was collected on the number of call attempts (n=1,2,3), the time to make calls (minutes), the success of the call attempt (connected with researcher or not), and researcher notes for each call attempt (who was reached, whether they remembered the survey, different email, etc.).

3. Emailing authors: Personalized emails including the survey link and article PDF were sent to authors who agreed to complete the survey. Authors with multiple studies were sent the unique survey links and PDFs for all their respective studies. The email noted the deadline to complete the survey (two weeks from the email sent date), but instructed authors that they would receive a final reminder if they had not had a chance to complete it after one week. With respect to emailing, we collected data on the number of emails sent and investigator time to send emails (minutes).

Supplemental Table B-3 Comparison of telephone contact and email interventions using Behaviour change techniques

While both telephone and email interventions sought to connect with trial authors, the telephone intervention was comparatively more active than the emailing intervention as demonstrated by the wider variety of Behaviour Change Techniques coded below [222]. For example, while both interventions included instruction on how to perform the behaviour, information about social and environmental consequences, material incentive to perform the behaviour, and a credible source, the telephone intervention additionally included goal setting for the behaviour, social support (unspecified), social support (practical), and prompts/cues. Behaviour change techniques were coded by one researcher (KJD) in collaboration with a behaviour change expert, Dr. Andrea Patey.

Behaviour change technique	Telephone contact	Email contact
Instruction on how to perform behaviour Advise or agree on how to perform the behaviour	In the follow-up email to authors who agreed to complete the survey, and 1 st page of survey (Introduction), information was provided on how to complete the survey	In the email letter and 1 st page of survey (Introduction), information was provided on how to complete the survey
Information about social & environmental consequences Provide information (e.g., written, verbal, visual) about social and environmental consequences of performing the behaviour	In the 1 st page of the survey (Introduction), information was provided about knowledge gaps in the field of diabetes quality improvement and how additional information about their trial may advance the collective knowledge about how to improve quality of care for patients with diabetes.	In the 1 st page of the survey (Introduction), information was provided about knowledge gaps in the field of diabetes quality improvement and how additional information about their trial may advance the collective knowledge about how to improve quality of care for patients with diabetes.
Credible source Present verbal or visual communication from a credible source in favour of or against the behaviour.	During the telephone call, the researcher stated that they were calling on behalf of renowned quality improvement/implementation scientist. On the 1 st page of survey (Introduction), the invitation letter was signed by the same scientist, on behalf of the wider research team. Logos of all affiliated institutions were displayed.	The email to authors was sent from the account of renowned quality improvement/implementation scientist. On the 1 st page of survey (Introduction), the invitation letter was signed by the same scientist, on behalf of the wider research team. Logos of all affiliated institutions displayed.
Material incentive (behaviour) Inform that money, vouchers or other valued objects will be delivered if and only if there has been effort and/or progress in performing the behaviour	On the 1 st page of survey (Introduction), authors were informed that, to thank them for their time and contribution to the study, they would be entered in a raffle to win one of five \$100 gift certificates to Amazon or iTunes.	On the 1 st page of survey (Introduction), authors were informed that, to thank them for their time and contribution to the study, they would be entered in a raffle to win one of five \$100 gift certificates to Amazon or iTunes.
Goal setting (behaviour) Set or agree on a goal defined in terms of the behaviour to be achieved	During the call, the researcher asked if the author would be interested in completing the survey, and if they agreed, sent the survey to them via email to complete with a stated	<i>BCT not identified in email contact</i>

	timeline (2 weeks from date of initial contact; if not completed, 1 reminder email was sent)	
<i>Social support (unspecified)</i> Advise on, arrange or provide social support (e.g., from friends, relatives, colleagues, 'buddies' or staff) or non-contingent praise or reward for performance of the behaviour. It includes encouragement and counseling, but only when it is directed at the behaviour	During the call, the researcher provided general support and encouragement to complete the survey and expressed gratitude to authors for their time and contributions	<i>BCT not identified in email contact</i>
<i>Social support (practical)</i> Advise on, arrange, or provide practice help (e.g., from friends, relatives, colleagues, 'buddies' or staff) for performance of the behaviour	During the call, the researcher provided practice support in completing the survey for authors having difficulties with the survey URL platform	<i>BCT not identified in email contact</i>
<i>Prompts/cues</i> Introduce or define environmental or social stimulus with the purpose of prompting or cueing the behaviour. The prompt or cue would normally occur at the time or place of performance	During the call, the researcher prompted the author if they recalled receiving an email request to complete the survey previously	<i>BCT not identified in email contact</i>

Supplemental Table B-4 Power calculations

Expected power to detect a difference in response proportions for different magnitudes of difference (Δ telephone contact) for a fixed sample (n=175) using a chi-square test of two proportions at a 0.05 level*

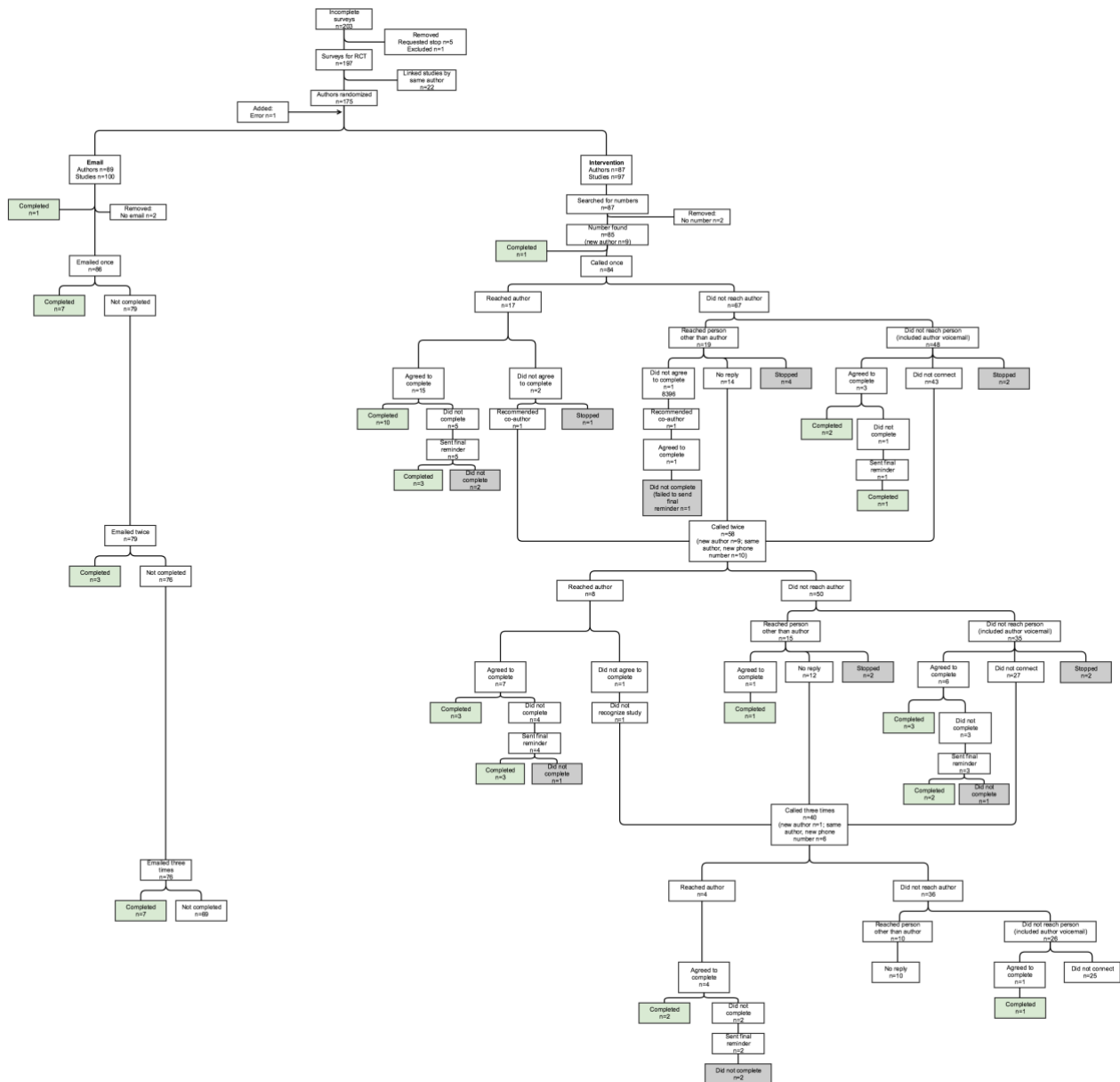
(Δ telephone contact)(n=87)	Email response proportion (n=88)			
	0.0001	0.01	0.1	0.2
0.05	0.5641	0.4374	0.1697	0.1241
0.1	0.8651	0.8014	0.4574	0.3322
0.15	0.9701	0.9515	0.7471	0.6051
0.2	0.9953	0.9916	0.918	0.829
0.3	1.0000	0.9999	0.9974	0.9898

*Calculated in Stata using: power twoproportions p1 p2, n1(n) n2(n)

Supplemental Figure B-1 Detailed author flow diagram

A detailed description of the trial flow is available at the following URL link:

<https://go.gliffy.com/go/share/syvk188u4cw5y6g4nvjs>. The image below represents a printable version of this flow diagram, however the URL was created to allow readers to zoom in on details.



Supplemental Text B-3. Additional details on the delivery of the telephone and email contact interventions

1) Email delivery details

- Post randomization, pre-delivery of email intervention
 - Prior to delivery of the email intervention, we checked all the URL's for control group. During process, we identified 3 studies that had been incorrectly classified prior to randomization (2 as incomplete, 1 as complete)
 - Randomized 1 author for 1 incomplete study (1144) prior to delivery of first round of emails; allocated to control group.
 - Removed 1 author for one study that had been previously completed
 - Of the remaining 88 authors, 2 authors were removed from the email intervention due to no retrievable email
- The email intervention was delivered October 25th to November 10th, 2016. If authors did not reply to the first email request, they received a second and third request one and two weeks later, respectively
- Emailed 86 authors once
 - 81 emails went through
 - 1 email was personalized to author who had emailed previously and completed, but not submitted the survey. Emailed to request permission to submit the completed survey. Author replied to say okay to submit
 - 5 failed to send. New emails were searched for and found for the same author and resent as part of the first round of emails.
 - 1 failed because the inbox was full
 - 2 failed because of institution security
 - 2 failed because email address could not be found
 - 7 authors completed the survey
 - 79 authors did not complete the survey
 - Note on URL link challenges: An error occurred in the mail merge of URL links that was brought to our attention by one of the emailed authors. While the text of the URLs had correctly merged, the embedded (unseen) hyperlink did not. Thus, when authors clicked on the link, it took them to the first survey in the control group. To access their unique survey, authors had to copy/paste the URL into their web browser manually. The problem was missed when the research did URL checks since she also did copy/paste of the URL to avoid the browser pulling up previous surveys. The researcher notified authors of the problem and advised them to copy/paste the URL. The embedded link was removed for Round 2 of emails, and corrected for Round 3.
- Emailed 79 authors twice
 - 79 emails went through
 - 3 authors completed a survey
 - 76 authors did not complete a survey
- Emailed 76 authors three times
 - 76 emails went through
 - 7 authors completed a survey

- 69 authors did not complete a survey

2) Calling delivery details

- Searched for numbers for 87 randomized authors
 - 2 authors removed: could not find number for author or coauthor listed on study
 - 9 authors changed to coauthors: could not find number for author but found number for coauthor listed on study (example reasons: deceased, retired, no information on phone number on research pages)
 - Total of 85 authors had numbers found
 - Contact to 1 author was stopped: survey completed (but not submitted) (10364) (Error: did not follow-up with author to request permission to submit)
- Called 84 authors once
- First calls occurred between November 1st, 2016 and March 28th, 2017
 - 17 authors reached directly
 - 15 authors agreed to complete survey (1 was permission to submit completed survey not submitted)
 - Sent 15 personalized emails to authors in follow-up to call with survey link(s) and article PDF(s)
 - 10 authors completed survey within the week
 - 5 authors did not complete survey within the week
 - Sent 5 reminder emails to authors who did not complete survey
 - 3 authors completed survey
 - 2 authors did not complete the survey
 - 2 authors did not agree to complete
 - 1 author recommended coauthor but did not provide information. Conducted additional search for number of coauthor
 - Contact to 1 author was stopped after a time zone error (12015)
 - 67 authors were not reached directly
 - In 19 cases, connection was made with a non-author
 - 8 non-authors appeared to be an assistant/secretary to the author
 - 4 assistants requested an email of the request be sent to them and/or the author
 - Sent 4 personalized emails to authors and/or their assistants at request
 - 1 author replied and recommended coauthor (PhD student). Coauthor agreed to complete but did not complete after 1 week. Error: Final reminder was not sent (8396)
 - 1 author did not reply. Further contact to author was stopped because assistant said he was busy (5368)
 - 2 authors were called again for the second call (12016 and 10354)
 - 3 assistant took a message (10223, 5254, 12443)

- 1 assistant instructed researcher to call back (11209)
- 11 non-authors were 'Other', typically general numbers for department or clinical offices
 - 5 non-authors were able to give a more direct number to the author or their assistant (5417, 10314, 6575, 12840, 10356)
 - 2 non-authors said they could not connect the call and requested an email of the request be sent
 - Sent 2 personalized emails to authors based on secretary's request
 - Contact to 2 authors was stopped because the hospital said they would not connect (10839, 6534)
 - 1 non-author took a message (2059)
 - 3 non-authors said they did not recognize the name
 - Contact to 1 author was stopped: Could not find new number for author or coauthor (10343)
 - 2 authors changed to coauthors: non-author did not recognize name of author. Searched for and obtained number for coauthor (6459, 5515)
- Of the total 19 cases where connection was made with a non-author:
 - 1 author replied and recommended a co-author (but co-author did not complete)
 - 14 authors did not reply
 - Contact to 4 authors was stopped
- In 48 cases, no connection was made with a person but resulted in the following:
 - 22 author voicemail
 - 9 voicemail other than author/unclear
 - 6 call failed
 - 4 disconnected
 - 3 no answer
 - 2 assistant voicemail
 - 2 automated reception, no connection
- Of the total 48 cases where no connection was made (other than voicemails)
 - 7 emails were sent
 - 3 authors requested emails via their voicemail (701, 1620, 6287). Of note, 2 of 3 provided new email addresses
 - Sent 3 personalized emails to authors
 - 1 author agreed to complete by completing the survey (did not reply) (701)
 - 1 author did not complete survey; stopped (1620)
 - 1 author did not complete survey; called again (6287)

- 2 authors replied to voicemail and agreed to complete survey (10278, 11206)
 - 1 author completed survey within one week (11206)
 - 1 author did not complete survey within one week
 - Sent 1 reminder email to author who did not complete survey (10278)
 - 1 author completed survey (10278)
- 1 assistant requested an email of the request be sent to them
 - Sent 1 personalized email to assistant (10244). Author did not reply and was called again
- Researcher decision: 1 author (1127) could not be reached via phone, as the number appeared incorrect. The author had previously replied to earlier emails to troubleshoot the survey. The researcher replied to the earlier emails to make request, and searched for a new phone number for call 2 [Protocol violation]
- Contact to 1 author was stopped: author could not be reached due to wrong number (1585) but had completed survey previously without submitting. Could not contact coauthor. The author had not replied to earlier email at end of June that asked if they wanted to submit or needed an assistance (did not send email to ask permission to submit).
 - Of 48 did not connect:
 - 3 authors agreed to complete the survey and went on to complete it
 - 43 authors were not connected with/did not reply
 - Contact to 2 authors was stopped

Called 58 authors twice

- Second calls occurred between November 2, 2016 and April 13th, 2017
- Changed to coauthor for 9 studies
- Changed to new phone number for author for 11 studies
 - 8 authors reached directly
 - 7 authors agreed to complete survey
 - Sent 7 personalized emails to authors in follow-up to call with survey link(s) and article PDF(s)
 - 3 authors completed survey within the week
 - 4 authors did not complete survey within the week
 - Sent 4 reminder emails to authors who did not complete survey
 - 3 authors completed the survey
 - 1 author did not complete the survey
 - 1 author did not recognize the study. Conducted additional search for number of coauthor. Called coauthor in Round 3
 - 50 authors were not reached directly
 - In 15 cases, connection was made with a non-author
 - 12 non-authors appeared to be an assistant/secretary to the author

- 1 assistant requested an email of the request be sent to author
 - Sent 1 personalized email to authors at assistants' request. No reply. Stopped (9023)
 - 8 assistants took a message (10354, 12016, 5417, 10223, 12420, 5254, 12297, 2059)
 - 3 assistant instructed researcher to call back (11209, 12443, 553)
- 3 non-authors were 'Other'
 - 1 non-authors were able to give a more direct number (11240)
 - 1 non-author said they did not recognize the name
 - Searched for and found new number for author
 - 1 non-author said it was a wrong number (challenge: language barrier).
 - Contact to 1 author was stopped (9075): could not find new number for author. Possibly retired. (Error: unclear if searched for co-authors)
- Of the total 15 cases where connection was made with a non-author:
 - 1 author completed survey (had received email during Round 1 at request of assistant) (12016)
 - 12 authors did not reply
 - Contact to 2 authors was stopped (9023, 9075)
- In 35 cases, no connection was made with a person but resulted in the following:
 - 14 author voicemail
 - 6 voicemail other than author/unclear
 - 1 call failed
 - 2 disconnected
 - 5 no answer
 - 3 assistant voicemail
 - 2 automated reception, no connection
 - 1 wrong number
 - 1 no details
- Of the total 35 cases where no connection was made (other than voicemails)
 - 8 emails were sent
 - 6 authors replied to voicemail and agreed to complete survey (5573, 8144, 9017, 6366, 1354, 5144)
 - 3 authors completed survey within one week (5573, 6366, 1354)
 - 3 authors did not complete survey within one week (8144, 9017, 5144)
 - Sent 1 reminder email to authors who did not complete survey (8144, 9017, 5144)

- 2 authors completed survey (5144; note 8144 not completed but 6253 by same author was)
- 1 email was re-forwarded to assistant who had requested an email of the request be sent to them in Round 1 (10244) (unclear if Error). Further contact to author was stopped afterwards.
- 1 author requested emails via their voicemail (6287) (same email)
 - Sent 1 personalized email to author
 - Author did not complete survey; stopped
- Of 35 did not connect:
 - 6 authors agreed to complete the survey and 5 went on to complete it
 - 27 authors were not connected with/did not reply
 - Contact to 2 authors was stopped

Called 40 authors three times

- Third calls occurred between November 2, 2016 and April 13, 2017
- Changed to coauthor for 1 study
- Changed to new phone number for author for 6 studies
 - 4 authors reached directly
 - 4 authors agreed to complete survey
 - Sent 4 personalized emails to authors in follow-up to call with survey link(s) and article PDF(s)
 - 2 authors completed survey within the week
 - 2 authors did not complete survey within the week
 - Sent 2 reminder emails to authors who did not complete survey
 - 2 author did not complete a survey
 - 1 author noted upon reviewing the study she realized she was not involved in it
 - 36 authors were not reached directly
 - In 10 cases, connection was made with a non-author
 - 8 non-authors appeared to be an assistant/secretary to the author
 - 3 assistants requested an email of the request be sent to author and/or assistant
 - Sent 1 personalized email to author at assistants' request. No reply. (5417)
 - Sent 1 personalized email to assistant. No reply (10156)
 - Sent 1 personalized email to author and assistant at assistants' request. No reply. (10223)
 - 4 assistants took a message (5943, 12297, 11209, 5254)
 - 1 assistant said author was not available (11292)
 - 2 non-authors were 'Other'

- 1 non-author said author could not be reached at that number. Emailed research coordinator to request new number for author. She replied that she had been in touch with author and author would contact us if interested (11240) (Error: Searching for number after 3rd call)
 - 1 non-author said they could not find authors' name (12420)
- Of the total 10 cases where connection was made with a non-author:
 - 10 authors did not reply
- In 26 cases, no connection was made with a person but resulted in the following:
 - 7 author voicemail
 - 7 voicemail other than author/unclear
 - 1 call failed
 - 2 disconnected
 - 4 no answer
 - 3 automated reception, no connection
 - 1 wrong number
 - 1 message, unclear (not in English)
- Of the total 26 cases where no connection was made (other than voicemails)
 - 1 email was sent
 - Personalized email sent to 1 author (2275), who based on searching for number to call, knew previous email was incorrect. Sent email to corrected email for author before moving on to coauthor. [Protocol violation]
 - 1 author completed the survey within one week
- Of 26 did not connect:
 - 1 author agreed to complete the survey and went on to complete it
 - 25 authors were not connected with/did not reply

Supplemental Text B-4. Protocol violations or potential errors and implementation challenges

Did not do, but should have

- Contact to one author was stopped after calling wrong number and searches did not provide a new number. Unclear if searched for co-authors (9075).
- Did not send final email reminder for 1 author (8396)

Did do, but should not have

- Sent email to two authors who could not be reached by phone due to incorrect numbers. In the first case, an email was sent in reply to a previous email from the author troubleshooting the survey URL (1127). In the second, email was sent to a corrected email identified during searches for the authors phone number before moving on to a coauthor (2275). Both contacts resulted in completion of the survey.

Unclear/no specified in protocol

- 11 emails were sent to assistants and/or authors at the request of an assistant.
 - During Round 1: 7 emails sent round 1 at request of assistant or secretary (8396, 5368, 12016, 10354, 10839, 6534, 10224)
 - During Round 2: 1 email to authors at assistants' request. (9023), 1 email was re-forwarded to assistant who had requested an email of the request be sent to them in Round 1
 - During Round 3, 3 assistants requested an email of the request be sent to author and/or assistant 5417, 10156, 10223
- Voicemail: Did not always leave a voicemail message.
- Contact to one author was stopped because survey was completed but not submitted (10364). Did not follow-up with author to confirm permission to submit. Accordingly, survey marked as incomplete.
- Contact to one author was stopped because survey was completed by not submitted and they could not be reached due to wrong number (1585). Did not think it was appropriate to contact coauthor, but did not email to ask permission to submit. The author had not replied to earlier email at end of June that asked if they wanted to submit or needed assistance.

Implementation challenges

- During the email intervention there were challenges with the survey URL resulting from a merge error. The technical issue was addressed and authors were notified of the problem.
- During the call intervention, one international author was called erroneously in the middle of the night due to a combination of unforeseen events (number obtained from university webpage was for a) a personal number and b) a time zone 4 hours behind the city in which the university was located).

APPENDIX C: Supplemental information for Manuscript 2

Supplemental Text C-1 Procedure used to derive approximate standard errors for continuous outcomes from reported p-values or confidence intervals

For all calculations, we assumed a t-distribution was more appropriate than a normal distribution due to the small number of clusters

Calculating a standard error from a p-value

- Extract intervention effect estimate (i.e., mean difference)
- Extract reported p-value
- Determine corresponding critical value from t distribution using p-value and degrees of freedom
- Degrees of freedom are given by $(N_1 + N_2 - 2)$ and N_1 and N_2 are the numbers of clusters in the experimental and control arms, respectively
- Critical value can be calculated by entering $=\text{tinv}(\text{p-value}, \text{degrees of freedom})$ in a cell in Microsoft Excel
- Calculate adjusted standard error according to the following:
 - o Intervention effect estimate (e.g., mean difference) / critical value
- Approach is difficult when levels of significance are reported vaguely (e.g., $p < 0.05$ or $p = \text{NS}$) instead of exact p-values
 - o When p-values were reported 'less than', we adopted a conservative approach and used the cut off number (e.g., if $p < 0.05$, we took $p = 0.05$).
 - o Where p-values were not provided or merely reported as "not significant", we did not calculate an adjusted standard error

Example 1. Calculating a standard error from a p-value in the diabetes review – Kinmonth et al. [149]

- o Outcome: mean difference glycated hemoglobin (HbA1c)
- o Report p-value comparing arms at follow-up, adjusted for clustering
- o Mean difference
 - = Intervention mean-Control mean
 - = $7.07 - 7.17$
 - = 0.1
- o Degrees of freedom
 - = # clusters-2
 - = $40 - 2$
 - = 38
- o p-value = 0.31
- o Critical value from t-distribution associated with p-value and degrees of freedom
 - = $\text{tinv}(0.31, 38)$
 - = 1.029
- o Adjusted standard error

= mean difference/critical value
= $0.1/1.029 = 0.097$ (calculated in excel without rounding)
= 0.01

- NB: that in addition to clustering, this estimate also adjusted for: district general hospital, practice list size, and organization of diabetes care

2. Calculating t-value from a confidence interval

- Obtain reported 95% confidence interval
- Typically, the standard error can be calculated according to the following:
 - $SE = (\text{upper limit} - \text{lower limit}) / 3.92$ (for 95% confidence intervals)
- However, in small samples such as a cluster study, confidence intervals should have been calculated using the t-distribution. This means that 3.92 must be replaced with the critical value obtained from the t-distribution the same as above
 - Degrees of freedom: N_1 and $N_2 - 2$

Example 1. Calculating a standard error from a confidence interval in the diabetes review – Smith et al [109]

- Outcome: mean difference systolic blood pressure (SBP)
- Report 95% confidence intervals comparing arms at follow-up, adjusted for clustering
 - $CI \text{ upper limit} - CI \text{ lower limit} / (t \text{ critical})$
 - $(4.52) - (-0.70) / (t\text{-critical})$
 - Critical value from t-distribution associated with $p=0.05$ and degrees of freedom $= 94 - 2 \rightarrow 1.986$
 - $SE = 2.6283$ (calculated in excel without rounding)

Supplemental Table C-1 Study details of included cluster randomized trials in the diabetes quality improvement review (continuous outcomes)

Author (Year)	No. clusters	No. patients	Cluster assessment	HbA1c (n=54)	SBP (n=40)	DBP (n=34)	LDL (n=26)
Griffin et al. (2011) [136]	343	3057	Appropriate analysis Adjusted SE of MD reported Adjusted SE of MD calculated ICC reported	Y (robust variance) N Y (95% CI) N	Y (robust variance) N Y (95% CI) N	Y (robust variance) N Y (95% CI) N	Y (robust variance) N Y (95% CI) N
Vaccaro et al. (2013) [144]	9	1461	Appropriate analysis Adjusted SE of MD reported Adjusted SE of MD calculated ICC reported	Y (mixed model) Y n/a Y (0.0661)	Y (mixed model) Y n/a Y (0.043)	Y (mixed model) Y n/a Y (0.1233)	Y (mixed model) Y n/a Y (0.0158)
Smith et al. (2008) [109]	94	639	Appropriate analysis Adjusted SE of MD reported Adjusted SE of MD calculated ICC reported	Y (GEE model) N N (log scale, post difference=0) N	Y (GEE model) N Y (95% CI) N	Y (GEE model) N Y (95% CI) N	Y (GEE model) N N (log scale) N
Goderis et al. (2010) [122]	90	2495	Appropriate analysis Adjusted SE of MD reported Adjusted SE of MD calculated ICC reported	Y (GEE model) N N (p value, difference in change=0) N	Y (GEE model) N Y (p value, difference in change) N	Y (GEE model) N N (p value, difference in change=0) N	Y (GEE model) N Y (p value, difference in change) N
Pape et al. (2011) [135]	9	6229	Appropriate analysis Adjusted SE of MD reported Adjusted SE of MD calculated ICC reported	Y (vague details) N Y (p value, post difference) N	Y (vague details) N Y (p value, post difference) N	Y (vague details) N N (p value, post difference=0) N	Y (vague details) N Y (p value, post difference) N
Herrin et al. (2006) [137]	22	2155	Appropriate analysis Adjusted SE of MD reported Adjusted SE of MD calculated ICC reported	N n/a n/a N	N n/a n/a N	N n/a n/a N	N n/a n/a N
Jansink et al. (2013) [148]	58	940	Appropriate analysis Adjusted SE of MD reported Adjusted SE of MD calculated ICC reported	Y (multilevel) N Y (95% CI) N	Y (multilevel) N Y (95% CI) N	Y (multilevel) N Y (95% CI) N	Y (multilevel) N Y (95% CI) N
Meigs et al. (2003) [171]	66	598	Appropriate analysis Adjusted SE of MD reported Adjusted SE of MD calculated ICC reported	Y (test statistic adjusted) N Y (p value, difference in change) N	Y (test statistic adjusted) N Y (p value, difference in change) N	Y (test statistic adjusted) N Y (p value, difference in change) N	Y (test statistic adjusted) N Y (p value, difference in change) N
Sönnichsen et al. (2010) [131]	6	1494	Appropriate analysis Adjusted SE of MD reported Adjusted SE of MD calculated ICC reported	Y (mixed model) N Y (95% CI) Y (-0.002)	Y (mixed model) N Y (95% CI) Y (0.002)	Y (mixed model) N Y (95% CI) Y (0.009)	Y (mixed model) N Y (95% CI) Y (0.007)
MacLean et al. (2009) [125]	64	7412	Appropriate analysis Adjusted SE of MD reported Adjusted SE of MD calculated ICC reported	Y (mixed model) N Y (95% CI) Y (0.055)	Y (mixed model) N Y (95% CI) N	Y (mixed model) N Y (95% CI) N	Y (mixed model) N Y (95% CI) N
Estrada et al. (2011) [129]	205	1182	Appropriate analysis Adjusted SE of MD reported Adjusted SE of MD calculated	N n/a n/a	N n/a n/a	N n/a n/a	N n/a n/a

			ICC reported	N	N	N	N
Cleveringa et al. (2008) [106]	55	3391	Appropriate analysis	Y (vague details)	Y (vague details)	Y (vague details)	Y (vague details)
			Adjusted SE of MD reported	N	N	N	N
			Adjusted SE of MD calculated	Y (95% CI)	Y (95% CI)	Y (95% CI)	Y (95% CI)
			ICC reported	N	N	N	N
Blackberry et al. (2013) [143]	59	473	Appropriate analysis	Y (GEE)	Y (GEE)	Y (GEE)	Y (GEE)
			Adjusted SE of MD reported	N	N	N	N
			Adjusted SE of MD calculated	Y (95% CI)	Y (95% CI)	Y (95% CI)	Y (95% CI)
			ICC reported	Y (0.098)	Y (0.096)	Y (0.088)	Y (0.058)
Quinn et al. (2011) [128]	26	213	Appropriate analysis	Y (mixed model)	Y (mixed model)	Y (mixed model)	Y (mixed model)
			Adjusted SE of MD reported	N	N	N	N
			Adjusted SE of MD calculated	Y (p value, difference in change)	N (difference in change; no p-value)	N (difference in change; no p-value)	N (difference in change; no p-value)
			ICC reported	N	N	N	N
Thomas et al. (2007) [110]	78	483	Appropriate analysis	Y (GEE)	Y (GEE)	Y (GEE)	Y (GEE)
			Adjusted SE of MD reported	N	N	N	N
			Adjusted SE of MD calculated	Y (p value, difference in change)	Y (p value, post difference)	Y (p value, post difference)	Y (p value, difference in change)
			ICC reported	N	N	N	N
O'Connor et al. (2011) [134]	11	2556	Appropriate analysis	Y (mixed model)	Y (mixed model)	Y (mixed model)	Y (mixed model)
			Adjusted SE of MD reported	N	N	N	N
			Adjusted SE of MD calculated	Y (p value, difference in change)	Y (p value, difference in change)	Y (p value, difference in change)	Y (p value, difference in change)
			ICC reported	Y (≤ 0.0002)	Y (≤ 0.0002)	Y (≤ 0.0002)	Y (≤ 0.0002)
Adachi et al. (2013) [147]	20	193	Appropriate analysis	Y (mixed model)	Y (mixed model)	Y (mixed model)	Y (mixed model)
			Adjusted SE of MD reported	N	N	N	N
			Adjusted SE of MD calculated	Y (95% CI)	Y (95% CI)	Y (95% CI)	Y (95% CI)
			ICC reported	N	N	N	N
Wolf et al. (2013) [140]	6	486	Appropriate analysis	Y (vague details)	Y (vague details)	Y (vague details)	Y (vague details)
			Adjusted SE of MD reported	N	N	N	N
			Adjusted SE of MD calculated	Y (95% CI)	Y (95% CI)	Y (95% CI)	Y (95% CI)
			ICC reported	N	N	N	N
Frei et al. (2014) [153]	30	326	Appropriate analysis	Y (multilevel)	Y (multilevel)	Y (multilevel)	Y (multilevel)
			Adjusted SE of MD reported	N	N	N	N
			Adjusted SE of MD calculated	Y (95% CI)	Y (95% CI)	Y (95% CI)	Y (95% CI)
			ICC reported	Y (< 0.001)	Y (0.019)	Y (< 0.001)	Y (0.04)
Smith et al. (2004) [117]	30	183	Appropriate analysis	N	N	N	N
			Adjusted SE of MD reported	n/a	n/a	n/a	n/a
			Adjusted SE of MD calculated	n/a	n/a	n/a	n/a
			ICC reported	N	N	N	N
Eccles et al. (2007) [120]	58	3608	Appropriate analysis	Y (GEE)	Y (GEE)	Y (GEE)	Y (GEE)
			Adjusted SE of MD reported	N	N	N	N
			Adjusted SE of MD calculated	Y (95% CI)	Y (95% CI)	Y (95% CI)	Y (95% CI)
			ICC reported	N	N	N	N
Olivarius et al. (2001) [116]	474	1316	Appropriate analysis	Y (GEE)	Y (GEE)	Y (GEE)	Y (GEE)
			Adjusted SE of MD reported	N	N	N	N
			Adjusted SE of MD calculated	N (post difference; p value on log scale)	N (post difference; p value on log scale)	N (post difference; p value on log scale)	N (post difference; p value on log scale)
			ICC reported	N	N	N	N
Dijkstra et al.	9	1350	Appropriate analysis	Y (mixed model)	Y (mixed model)	Y (mixed model)	Y (mixed model)

(2005) [114]			Adjusted SE of MD reported	N	N	N	
			Adjusted SE of MD calculated	Y (p value, difference in change)	N (difference of change; no p-value)	Y (p value, difference in change)	
			ICC reported	N	N	N	
Barceló et al. (2010) [124]	10	307	Appropriate analysis	N	N	N	
			Adjusted SE of MD reported	n/a	n/a	n/a	
			Adjusted SE of MD calculated	n/a	n/a	n/a	
			ICC reported	N	N	N	
van Bruggen et al. (2008) [107]	30	1640	Appropriate analysis	Y (GEE)	Y (GEE)	Y (GEE)	
			Adjusted SE of MD reported	N	N	N	
			Adjusted SE of MD calculated	Y (p value, post difference)	Y (p value, post difference)	Y (p value, post difference)	
			ICC reported	N	N	N	
Kinmonth et al. (1998) [149]	41	360	Appropriate analysis	Y (unclear)	Y (unclear)	Y (unclear)	
			Adjusted SE of MD reported	N	N	N	
			Adjusted SE of MD calculated	Y (p value, post difference)	Y (p value, post difference)	Y (p value, post difference)	
			ICC reported	Y (0.047)	N	N	
Hetlevik et al. (2000) [138]	29	1034	Appropriate analysis	N	N	N	
			Adjusted SE of MD reported	n/a	n/a	n/a	
			Adjusted SE of MD calculated	n/a	n/a	n/a	
			ICC reported	N	N	N	
Simmons et al. (2004) [123]	135	398	Appropriate analysis	Y (unclear)	Y (unclear)	Y (unclear)	
			Adjusted SE of MD reported	N	N	N	
			Adjusted SE of MD calculated	N (not enough information; geometric means)	N (not enough information; geometric means)	N (not enough information; geometric means)	
			ICC reported	N	N	N	
Groeneveld et al. (2001) [127]	15	246	Appropriate analysis	N	N	N	
			Adjusted SE of MD reported	n/a	n/a	n/a	
			Adjusted SE of MD calculated	n/a	n/a	n/a	
			ICC reported	N	N	N	
O'Hare et al. (2004) [119]	6	361	Appropriate analysis	N	N	N	
			Adjusted SE of MD reported	n/a	n/a	n/a	
			Adjusted SE of MD calculated	n/a	n/a	n/a	
			ICC reported	N	N	N	
Piatt et al. (2006) [111]	11	119	Appropriate analysis	Y (mixed model)	Y (mixed model)	Y (mixed model)	
			Adjusted SE of MD reported	N	N	N	
			Adjusted SE of MD calculated	N (difference in change; p value but from ANOVA 3 arm comparison)	N (difference in change; p value but from ANOVA 3 arm comparison)	N (difference in change; p value but from ANOVA 3 arm comparison)	
			ICC reported	N	N	N	
Krass et al. (2007) [118]	56	335	Appropriate analysis	N	N	N	
			Adjusted SE of MD reported	n/a	n/a	n/a	
			Adjusted SE of MD calculated	n/a	n/a	n/a	
			ICC reported	N	N	N	
Hermans et al. (2013) [139]	477	4027	Appropriate analysis	N	N	N	N
			Adjusted SE of MD reported	n/a	n/a	n/a	n/a
			Adjusted SE of MD calculated	n/a	n/a	n/a	n/a
			ICC reported	N	N	N	N
O'Connor et al. (2005) [80]	12	754	Appropriate analysis	Y (mixed model)	Y (mixed model)	Y (mixed model)	
			Adjusted SE of MD reported	N	N	N	
			Adjusted SE of MD calculated	Y (p value, difference in	N (p value, difference in	Y (p value, difference in	

			ICC reported	change) Y (0.027)	change=0) Y (0.009)	change) Y (0.000)
Peterson et al. (2008) [176]	24	8405	Appropriate analysis Adjusted SE of MD reported Adjusted SE of MD calculated ICC reported	Y (mixed model) N N (difference in change; no p value) N	Y (mixed model) N N (difference in change; no p value) N	Y (mixed model) N N (difference in change; no p value) N
Phillips et al. (2005) [112]	345	4138	Appropriate analysis Adjusted SE of MD reported Adjusted SE of MD calculated ICC reported	Y (mixed model) N N (p value, difference in change; no cluster numbers per 4 arms) N	Y (mixed model) N N (p value, difference in change; no cluster numbers per 4 arms) N	Y (mixed model) N N (p value, difference in change; no cluster numbers per 4 arms) N
Heisler et al. (2012) [126]	16	4622	Appropriate analysis Adjusted SE of MD reported Adjusted SE of MD calculated ICC reported	N n/a n/a N	Y (hierarchical model) N N (difference of change, CI; no p value) N	N n/a n/a N
Gagliardino et al. (2013) [145]	36	468	Appropriate analysis Adjusted SE of MD reported Adjusted SE of MD calculated ICC reported	N n/a n/a N	N n/a n/a N	
O'Connor et al. (2009) [121]	123	3703	Appropriate analysis Adjusted SE of MD reported Adjusted SE of MD calculated ICC reported	Y (mixed model) N N (analyzed as factorial study) N		Y (mixed model) N N (analyzed as factorial study) N
Sieber et al. (2012) [150]	21	914	Appropriate analysis Adjusted SE of MD reported Adjusted SE of MD calculated ICC reported	N n/a n/a N		N n/a n/a N
Guldborg et al. (2011) [132]	86	2716	Appropriate analysis Adjusted SE of MD reported Adjusted SE of MD calculated ICC reported	Y (robust variance estimation) N Y (95% CI) N		
Wagner et al. (2001) [175]	35	707	Appropriate analysis Adjusted SE of MD reported Adjusted SE of MD calculated ICC reported	Y (mixed model) N N (p value; post difference=0) N		
Glasgow et al. (2005) [115]	52	886	Appropriate analysis Adjusted SE of MD reported Adjusted SE of MD calculated ICC reported	Y (random effect model) N Y (p value, post difference) N		
Harris et al. (2005) [113]	90	660	Appropriate analysis Adjusted SE of MD reported Adjusted SE of MD calculated ICC reported	Y (mixed model) N Y (p value, post difference) N		
Steyn et al. (2013) [142]	18	456	Appropriate analysis Adjusted SE of MD reported Adjusted SE of MD calculated ICC reported	Y (robust variance estimation) N Y (95% CI) Y (0.1)		

Mehuys et al. (2011) [130]	66	288	Appropriate analysis Adjusted SE of MD reported Adjusted SE of MD calculated ICC reported	N n/a n/a N		
Polonsky et al. (2011) [133]	34	483	Appropriate analysis Adjusted SE of MD reported Adjusted SE of MD calculated ICC reported	Y (mixed model) N Y (p value, difference in change) N		
Kulkarni et al. (1998) [160]	19	54	Appropriate analysis Adjusted SE of MD reported Adjusted SE of MD calculated ICC reported	N n/a n/a N		
Grant et al. (2008) [108]	11	244	Appropriate analysis Adjusted SE of MD reported Adjusted SE of MD calculated ICC reported	N n/a n/a N		
Saenz et al. (2012) [151]	14	697	Appropriate analysis Adjusted SE of MD reported Adjusted SE of MD calculated ICC reported	N n/a n/a N		
DePue et al. (2013) [146]	12	268	Appropriate analysis Adjusted SE of MD reported Adjusted SE of MD calculated ICC reported	Y (mixed model) Y n/a N		
Morgan et al. (2013) [141]	11	156	Appropriate analysis Adjusted SE of MD reported Adjusted SE of MD calculated ICC reported	N n/a n/a N		
Juul et al. (2014) [152]	40	4034	Appropriate analysis Adjusted SE of MD reported Adjusted SE of MD calculated ICC reported	Y (mixed model) N Y (95% CI) N		
Steventon et al. (2014) [154]	112	513	Appropriate analysis Adjusted SE of MD reported Adjusted SE of MD calculated ICC reported	Y (mixed model) N Y (95% CI) N		
Bellary et al. (2008) [37]	21	1486	Appropriate analysis Adjusted SE of MD reported Adjusted SE of MD calculated ICC reported	Y (hierarchical mixed model) N Y (95% CI) N	Y (hierarchical mixed model) N Y (95% CI) N	
Bebb et al. (2007) [165]	42	1534	Appropriate analysis Adjusted SE of MD reported Adjusted SE of MD calculated ICC reported	Y (random effect model) N Y (95% CI) Y (0.034)	Y (random effect model) N Y (95% CI) Y (0.039)	

Supplemental Table C-2 Study details of included cluster randomized trials in the diabetes quality improvement review (dichotomous outcomes)

Author (Year)	No. clusters	No. patients	Cluster assessment	ASA (n=10)	Statin (n=10)	Anti-hyp (n=11)	Retin (n=27)	Foot (n=24)	Renal (n=20)	Htn-c (n=20)	Smoke (n=14)	Harms (n=4)
Hermans et al. (2013) [139]	477	4027	Appropriate analysis ICC reported	N N	N N	N N	Y (MLMM) N	Y (MLMM) N		Y (MLMM) N	Y (MLMM) N	
Griffin et al. (2011) [136]	343	3057	Appropriate analysis ICC reported	N N	N N	N N				N N	N N	
Vaccaro et al. (2013) [144]	9	1461	Appropriate analysis ICC reported	Y (MM) Y (0.1052)	Y (MM) Y (0.0455)	Y (MM) N				Y (MM) N		
Smith et al. (2004) [117]	30	183	Appropriate analysis ICC reported	N N	N N		N N	N N	N N			
Smith et al. (2008) [109]	94	639	Appropriate analysis ICC reported	Y (GEE) N	Y (GEE) N					Y (GEE) N		
Goderis et al. (2010) [122]	90	2495	Appropriate analysis ICC reported	Y (GEE) N	Y (GEE) N						Y (GEE) N	
Eccles et al. (2007) [120]	58	3608	Appropriate analysis ICC reported	Y (GEE) N		Y (GEE) N	Y (GEE) N	Y (GEE) N	Y (GEE) N		Y (GEE) N	
Ilag et al. (2003) [164]	9	284	Appropriate analysis ICC reported	Y (MM) N			Y (MM) N	Y (MM) N	Y (MM) N	Y (MM) N	Y (MM) N	
O'Connor et al. (2005) [80]	12	754	Appropriate analysis ICC reported	Y (MM) Y (0.000)			Y (MM) Y (0.088)	Y (MM) Y (0.013)	Y (MM) Y (0.019)			
Persell et al. (2008) [159]	19	334	Appropriate analysis ICC reported	Y (GEE) N								
Guldborg et al. (2011) [132]	86	2716	Appropriate analysis ICC reported		Y (robust) N	Y (robust) N	Y (robust) N					
Shah et al. (2014) [172]	80	1592	Appropriate analysis ICC reported		Y (GEE) Y (0.123)	Y (GEE) Y (0.06)				Y (GEE) Y (0.089)		
Bellary et al. (2008) [37]	21	1486	Appropriate analysis ICC reported		N N	N N						
Pape et al. (2011) [135]	9	6229	Appropriate analysis ICC reported		Y (vague) N					Y (vague) N		
McDermott et al. (2001) [157]	21	555	Appropriate analysis ICC reported			N N	N N	N N	N N	N N		
Schnipper et al. (2010) [170]	239	5011	Appropriate analysis ICC reported			Y (GEE) N	Y (GEE) N	Y (GEE) N	Y (GEE) N			
Olivarius et al. (2001) [116]	474	1316	Appropriate analysis ICC reported			Y (MM) N				Y (MM) N	Y (MM) N	Y (MM) N
Bebb et al. (2007) [165]	42	1534	Appropriate analysis ICC reported			Y (RE) Y (0.05)						
Herrin et al. (2006) [137]	22	2155	Appropriate analysis ICC reported				N N	N N	N N	N N		
Peterson et al. (2008) [176]	24	8405	Appropriate analysis ICC reported				Y (MM) N	Y (MM) N	Y (MM) N	Y (MM) N		
Dijkstra et al. (2008) [161]	40	2059	Appropriate analysis ICC reported				Y (MMM) N	Y (MMM) N	Y (MMM) N	Y (MMM) N		
Vidal-Pardo et al. (2013) [174]	108	2938	Appropriate analysis ICC reported				Y (unclear) N	Y (unclear) N	Y (unclear) N		N N	
Ward et al. (1996)	139	386	Appropriate analysis				N	N	N			

[158]			ICC reported	N	N	N		
Wagner et al. (2001) [175]	35	707	Appropriate analysis	Y (MM)	Y (MM)	Y (MM)		
			ICC reported	N	N	N		
Dijkstra et al. (2005) [114]	9	1350	Appropriate analysis	Y (MM)	Y (MM)	Y (MM)		
			ICC reported	N	N	N		
Glasgow et al. (2005) [115]	52	886	Appropriate analysis	N	N	N		
			ICC reported	N	N	N		
Harris et al. (2005) [113]	90	660	Appropriate analysis	Y (MM)	Y (MM)	Y (MM)		
			ICC reported	N	N	N		
Jansink et al. (2013)[148]	58	940	Appropriate analysis	Y (ML)	Y (ML)	Y (ML)		
			ICC reported	N	N	N		
Dickinson et al. (2014) [173]	40	822	Appropriate analysis	Y (MM)	Y (MM)	Y (MM)		
			ICC reported	N	N	N		
Meigs et al. (2003) [171]	66	598	Appropriate analysis	Y (GEE)	Y (GEE)		Y (GEE)	
			ICC reported	N	N		N	
Barceló et al. (2010) [124]	10	307	Appropriate analysis	Y (ML)	Y (ML)		Y (ML)	
			ICC reported	N	N		N	
Frijling et al. (2002) [155]	124	1410	Appropriate analysis	Y (ML)	Y (ML)			
			ICC reported	Y (0.2)	Y (0.33)			
Sönnichsen et al. (2010) [131]	6	1494	Appropriate analysis	N	N			
			ICC reported	N	N			
Steyn et al. (2013) [142]	18	456	Appropriate analysis	Y (robust)	Y (robust)			
			ICC reported	N	N			
Kirwin et al. (2010) [168]	8	346	Appropriate analysis	Y (GEE)		Y (GEE)	Y (GEE)	
			ICC reported	N		N	N	
Perria et al. (2007) [166]	252	6290	Appropriate analysis	N		N		
			ICC reported	N		N		
McClellan et al. (2003) [163]	123	22971	Appropriate analysis	N				
			ICC reported	N				
Kiefe et al. (2001) [156]	70	2978	Appropriate analysis		Y (unclear)			
			ICC reported		N			
MacLean et al. (2009) [125]	64	7412	Appropriate analysis			Y (MM)		
			ICC reported			N		
Estrada et al. (2011) [129]	205	1182	Appropriate analysis				Y (MM)	N
			ICC reported				N	N
New et al. (2004) [165]	44	5005	Appropriate analysis				N	
			ICC reported				N	
Reiber et al. (2004) [170]	14	1593	Appropriate analysis				N	
			ICC reported				N	
van Bruggen et al. (2008) [107]	30	1640	Appropriate analysis				Y (GEE)	
			ICC reported				N	
Sequist et al. (2010) [169]	31	7557	Appropriate analysis				Y (GEE)	
			ICC reported				N	
Kinmonth et al. (1998) [149]	41	360	Appropriate analysis					Y (unclear)
			ICC reported					N
Hetlevik et al. (2000) [138]	29	1034	Appropriate analysis					N
			ICC reported					N
Simmons et al. (2004) [123]	135	398	Appropriate analysis					N
			ICC reported					N
Cleveringa et al.	55	3391	Appropriate analysis					Y (vague)

(2008) [114]			ICC reported	N	
Mehuys et al.	66	288	Appropriate analysis	N	
(2011) [130]			ICC reported	N	
Blackberry et al.	59	473	Appropriate analysis	Y (GEE)	
(2013) [143]			ICC reported	Y (0.043)	
Groeneveld et al.	15	246	Appropriate analysis		N
(2001) [127]			ICC reported		N
Quinn et al. (2011)	26	213	Appropriate analysis		N
[128]			ICC reported		N
Polonsky et al.	34	483	Appropriate analysis		N
(2011) [133]			ICC reported		N

ASA: Acetylsalicylic acid; Anti-hyp: Anti-hypertensive; GEE: Generalized estimating equation; Htn-c: Hypertension control; MM: mixed model; MMM: multilevel mixed model; ML: multilevel model; RE: random effects; Retin: Retinopathy;

APPENDIX D: Supplemental information for Manuscript 3

Supplemental Text D-1 Model for posterior and posterior predictive distribution

```
model {  
  for(i in 17:25) {  
    P[i] <- (1/V[i])  
    rho[i] ~ dnorm(mu[i], P[i])  
    mu[i] ~ dnorm(d, prec)  
  }  
  
  d ~ dnorm (0, 0.001)  
  tau ~ dunif (0,2)  
  prec <- (1/tau) * (1/tau)  
  
  d_posterior <- exp(d) / ( 1 + exp(d))  
  logit_rho_pred ~ dnorm(d, prec)  
  rho_pred <- exp(logit_rho_pred) / ( 1 + exp(logit_rho_pred))  
  
}
```

Note: The model loop above (17:25) was for evidence in the ICC database pertaining to outcome HbA1c. The model was run for each outcome using outcome-specific ICCs.

Supplemental Text D-2 Code for hierarchical models implemented in comparison ICC from posterior predictive distribution vs. of single ICC

1) Hierarchical model with imputation of ICC from posterior predictive distribution

```

model {
  for(i in 1:59) {
    for(j in 1:n_arms[i]) {
      y[i,j] ~ dnorm(mu[i,j] , prec.resp[i,j])
      mu[i,j] = beta[i,1] + beta[i,2] * CM[i,j]
                                     + beta[i,3] * TC[i,j]
                                     + beta[i,4] * EPR[i,j]
                                     + beta[i,5] * CE[i,j]
                                     + beta[i,6] * FR[i,j]
                                     + beta[i,7] * PE[i,j]
                                     + beta[i,8] * PSM[i,j]
                                     + beta[i,9] * PR[i,j]
                                     + beta[i,10] * Other[i,j]

      prec.resp[i,j] <- (1/se[i,j])*(1/se[i,j])
    }
    for(m in 1:10) {
      beta[i,m] ~ dnorm(mu_bar[m], prec[m])
    }
  }

  for(i in 60:68) {
    for(j in 1:n_arms[i]) {
      y[i,j] ~ dnorm(mu[i,j] , prec.resp[i,j])
      mu[i,j] = beta[i,1] + beta[i,2] * CM[i,j]
                                     + beta[i,3] * TC[i,j]
                                     + beta[i,4] * EPR[i,j]
                                     + beta[i,5] * CE[i,j]
                                     + beta[i,6] * FR[i,j]
                                     + beta[i,7] * PE[i,j]
                                     + beta[i,8] * PSM[i,j]
                                     + beta[i,9] * PR[i,j]
                                     + beta[i,10] * Other[i,j]

      prec.resp[i,j] <- (1/(se[i,j]*sqrt(corr[i,j])))*(1/(se[i,j]*sqrt(corr[i,j])))
      corr[i,j] <- 1 + (avg_cluster_size[i,j]-1) * ICC[i,j]
      ICC[i,j] <- exp(logit_ICC[i,j]) / ( 1 + exp(logit_ICC[i,j]))
      logit_ICC[i,j] ~ dnorm(-4.16224, (1/(1.8812*1.8812)))
    }
    for(m in 1:10) {
      beta[i,m] ~ dnorm(mu_bar[m], prec[m])
    }
  }

  for(i in 69:100) {
    for(j in 1:n_arms[i]) {
      y[i,j] ~ dnorm(mu[i,j] , prec.resp[i,j])
      mu[i,j] = beta[i,1] + beta[i,2] * CM[i,j]
                                     + beta[i,3] * TC[i,j]
                                     + beta[i,4] * EPR[i,j]
                                     + beta[i,5] * CE[i,j]
                                     + beta[i,6] * FR[i,j]
                                     + beta[i,7] * PE[i,j]
                                     + beta[i,8] * PSM[i,j]
                                     + beta[i,9] * PR[i,j]

```

```

+ beta[i,10] * Other[i,j]

prec.resp[i,j] <- (1/se[i,j])*(1/se[i,j])
se[i,j] ~ dunif(0,2)
}
for(m in 1:10) {
beta[i,m] ~ dnorm(mu_bar[m], prec[m])
}
}

for(i in 101:101) {
for(j in 1:n_arms[i]) {
y[i,j] ~ dnorm(mu[i,j], prec.resp[i,j])
mu[i,j] = beta[i,1] + beta[i,2] * CM[i,j]
+ beta[i,3] * TC[i,j]
+ beta[i,4] * EPR[i,j]
+ beta[i,5] * CE[i,j]
+ beta[i,6] * FR[i,j]
+ beta[i,7] * PE[i,j]
+ beta[i,8] * PSM[i,j]
+ beta[i,9] * PR[i,j]
+ beta[i,10] * Other[i,j]

se[i,j] ~ dunif(0, 2)
prec.resp[i,j] <- (1/(se[i,j]*sqrt(corr[i,j])))*(1/(se[i,j]*sqrt(corr[i,j])))
corr[i,j] <- 1 + (avg_cluster_size[i,j]-1) * ICC[i,j]
}
for(m in 1:10) {
beta[i,m] ~ dnorm(mu_bar[m], prec[m])
}
}

for(i in 102:114) {
for(j in 1:n_arms[i]) {
y[i,j] ~ dnorm(mu[i,j], prec.resp[i,j])
mu[i,j] = beta[i,1] + beta[i,2] * CM[i,j]
+ beta[i,3] * TC[i,j]
+ beta[i,4] * EPR[i,j]
+ beta[i,5] * CE[i,j]
+ beta[i,6] * FR[i,j]
+ beta[i,7] * PE[i,j]
+ beta[i,8] * PSM[i,j]
+ beta[i,9] * PR[i,j]
+ beta[i,10] * Other[i,j]

se[i,j] ~ dunif(0,2)
prec.resp[i,j] <- (1/(se[i,j]*sqrt(corr[i,j])))*(1/(se[i,j]*sqrt(corr[i,j])))
corr[i,j] <- 1 + (avg_cluster_size[i,j]-1) * ICC[i,j]
ICC[i,j] <- exp(logit_ICC[i,j]) / ( 1 + exp(logit_ICC[i,j]))
logit_ICC[i,j] ~ dnorm(-4.16224, (1/(1.8812*1.8812)))
}
for(m in 1:10) {
beta[i,m] ~ dnorm(mu_bar[m], prec[m])
}
}

mu_bar[1] ~ dnorm(8,0.01)
for(m in 2:10) { mu_bar[m] ~ dnorm(0,0.25) }
for(m in 1:10) {
prec[m] = (1/tau[m]) * (1/tau[m])
}
tau[1] ~ dunif(0,2)

```

```

    for(m in 2:10) { tau[m] ~ dunif(0,2) }

    rk <- rank(mu_bar[2:10])

}

```

1) Hierarchical model with single ICC imputation (ICC=0.07)

```

model {
  for(i in 1:59) {
    for(j in 1:n_arms[i]) {
      y[i,j] ~ dnorm(mu[i,j] , prec.resp[i,j])
      mu[i,j] = beta[i,1] + beta[i,2] * CM[i,j]
                                     + beta[i,3] * TC[i,j]
                                     + beta[i,4] * EPR[i,j]
                                     + beta[i,5] * CE[i,j]
                                     + beta[i,6] * FR[i,j]
                                     + beta[i,7] * PE[i,j]
                                     + beta[i,8] * PSM[i,j]
                                     + beta[i,9] * PR[i,j]
                                     + beta[i,10] * Other[i,j]

      prec.resp[i,j] <- (1/se[i,j])*(1/se[i,j])
    }
    for(m in 1:10) {
      beta[i,m] ~ dnorm(mu_bar[m], prec[m])
    }
  }

  for(i in 60:68) {
    for(j in 1:n_arms[i]) {
      y[i,j] ~ dnorm(mu[i,j] , prec.resp[i,j])
      mu[i,j] = beta[i,1] + beta[i,2] * CM[i,j]
                                     + beta[i,3] * TC[i,j]
                                     + beta[i,4] * EPR[i,j]
                                     + beta[i,5] * CE[i,j]
                                     + beta[i,6] * FR[i,j]
                                     + beta[i,7] * PE[i,j]
                                     + beta[i,8] * PSM[i,j]
                                     + beta[i,9] * PR[i,j]
                                     + beta[i,10] * Other[i,j]

      prec.resp[i,j] <- (1/(se[i,j]*corr[i,j]))*(1/(se[i,j]*corr[i,j]))
      corr[i,j] <- 1 + (avg_cluster_size[i,j]-1) * 0.07
    }
    for(m in 1:10) {
      beta[i,m] ~ dnorm(mu_bar[m], prec[m])
    }
  }

  for(i in 69:100) {
    for(j in 1:n_arms[i]) {
      y[i,j] ~ dnorm(mu[i,j] , prec.resp[i,j])
      mu[i,j] = beta[i,1] + beta[i,2] * CM[i,j]
                                     + beta[i,3] * TC[i,j]
                                     + beta[i,4] * EPR[i,j]
                                     + beta[i,5] * CE[i,j]
                                     + beta[i,6] * FR[i,j]

```

```

+ beta[i,7] * PE[i,j]
+ beta[i,8] * PSM[i,j]
+ beta[i,9] * PR[i,j]
+ beta[i,10] * Other[i,j]

prec.resp[i,j] <- (1/se[i,j])*(1/se[i,j])
se[i,j] ~ dunif(0,2)
}
for(m in 1:10) {
beta[i,m] ~ dnorm(mu_bar[m], prec[m])
}
}

for(i in 101:101) {
for(j in 1:n_arms[i]) {
y[i,j] ~ dnorm(mu[i,j] , prec.resp[i,j])
mu[i,j] = beta[i,1] + beta[i,2] * CM[i,j]

+ beta[i,3] * TC[i,j]
+ beta[i,4] * EPR[i,j]
+ beta[i,5] * CE[i,j]
+ beta[i,6] * FR[i,j]
+ beta[i,7] * PE[i,j]
+ beta[i,8] * PSM[i,j]
+ beta[i,9] * PR[i,j]
+ beta[i,10] * Other[i,j]

se[i,j] ~ dunif(0, 2)
prec.resp[i,j] <- (1/(se[i,j]*corr[i,j]))*(1/(se[i,j]*corr[i,j]))
corr[i,j] <- 1 + (avg_cluster_size[i,j]-1) * 0.07
}
for(m in 1:10) {
beta[i,m] ~ dnorm(mu_bar[m], prec[m])
}
}

for(i in 102:114) {
for(j in 1:n_arms[i]) {
y[i,j] ~ dnorm(mu[i,j] , prec.resp[i,j])
mu[i,j] = beta[i,1] + beta[i,2] * CM[i,j]

+ beta[i,3] * TC[i,j]
+ beta[i,4] * EPR[i,j]
+ beta[i,5] * CE[i,j]
+ beta[i,6] * FR[i,j]
+ beta[i,7] * PE[i,j]
+ beta[i,8] * PSM[i,j]
+ beta[i,9] * PR[i,j]
+ beta[i,10] * Other[i,j]

se[i,j] ~ dunif(0,2)
prec.resp[i,j] <- (1/(se[i,j]*corr[i,j]))*(1/(se[i,j]*corr[i,j]))
corr[i,j] <- 1 + (avg_cluster_size[i,j]-1) * 0.07
}
for(m in 1:10) {
beta[i,m] ~ dnorm(mu_bar[m], prec[m])
}
}

mu_bar[1] ~ dnorm(8,0.01)
for(m in 2:10) { mu_bar[m] ~ dnorm(0,0.25) }
for(m in 1:10) {
prec[m] = (1/tau[m]) * (1/tau[m])
}

```

```
tau[1] ~ dunif(0,2)
for(m in 2:10) { tau[m] ~ dunif(0,2) }

rk <- rank(mu_bar[2:10])
```

```
for(m in 1:10) {
  mu_bar_new[m] ~ dnorm(mu_bar[m] , prec[m])
}

}
```

APPENDIX E: Supplemental information for Manuscript 4

Supplemental Text E-1 Imputation of missing variances and estimates of the ICC

Missing estimates of variance

- In extraction, we sought the standard deviation or standard error of post-intervention group mean
- If unavailable, we abstracted other measures of variance and calculated the standard error according to established methods [5,239]
- For meta-analyses, we converted abstracted standard deviations to sample standard errors
- After performing the above, there was missing data for 68 observations (standard errors of group means) from patient randomized trials and missing data from 31 observations from cluster randomized trials (99 of 241 observations; 41%)
- The median and range for observed standard errors was 0.18 (0.08-0.51). We opted to impute missing standard errors from a uniform distribution with parameters 0,2.

Missing estimates of ICC

- There were 23 cluster RCTs that assessed HbA1c in the example dataset
- Of these, no study reported adjusted standard errors of the post group means of HbA1c
- Two studies (included one excluded from the example because it did not report baseline HbA1c) reported estimates of the intraclass correlation coefficient (ICC). We used these ICC to calculate a correction factor to adjust reported standard errors for those two studies
- The remainder of standard errors from cluster RCTs (corresponding to 49 observations) required adjustment from an external ICC estimate
- We opted to impute missing ICCs from a predictive distribution obtained from the synthesis of ICCs in the diabetes QI review and external sources. Based on 9 estimates of the ICC for the HbA1c outcome, we estimated a posterior predictive distribution with a logit-transformed mean of -4.16224 and standard deviation of 1.8812.

Supplemental Table E-1 Analysis bins for hierarchical models

The models presented in Supplemental Text E-2 –E-6 applied the same linear model to different analysis bins. In parsing data this way, we were able to account for nuanced design and reporting challenges, including missing variances and ICCs. The following tables describe the content of the analysis bins, the number of studies and number of arms, and how data within were handled uniquely.

Analysis bins for Analysis 2 (pairwise data to match standard meta-analytic models)

Analysis bin	Data characteristics	# arms (n=228)*	# studies (n=114)	#bin
1	Patient RCT, variance reported Run model	118	59	1-59
2	Cluster RCT, variance reported, cluster adjusted Run model	0	0	
3	Cluster RCT, variance reported, not cluster adjusted, ICC reported Adjust with ICC, then run model	0	0	--
4	Cluster RCT, variance reported, not cluster adjusted, no ICC reported Impute ICC, adjust with ICC, then run model	18	9	60-68
5	Patient RCT, no variance reported Impute SE, then run model	64	32	69-100
6	Cluster RCT, no variance reported, not cluster adjusted, ICC reported Impute SE, adjust with ICC, then run model	2	1	101
7	Cluster RCT, no variance reported, not cluster adjusted, no ICC reported Impute SE, impute ICC, adjust with ICC, then run model	26	13	102-114

*13 arms from 8 multiarm studies were dropped

Analysis 3 (complete data from all arms)†

Analysis bin	Data characteristics	# arms (n=241)	# studies (n=114)	#bin
1	Patient RCT, variance reported Run model	122	59	1-59
2	Cluster RCT, variance reported, cluster adjusted Run model	0	0	
3	Cluster RCT, variance reported, not cluster adjusted, ICC reported Adjust with ICC, then run model	0	0	--
4	Cluster RCT, variance reported, not cluster adjusted, no ICC reported Impute ICC, adjust with ICC, then run model	20	9	60-68
5	Patient RCT, no variance reported Impute SE, then run model	68	32	69-100
6	Cluster RCT, no variance reported, not cluster	2	1	101-101

	adjusted, ICC reported Impute SE, adjust with ICC, then run model			
7	Cluster RCT, no variance reported, not cluster adjusted, no ICC reported Impute SE, impute ICC, adjust with ICC, then run model	29	13	102-114

†Data from Analysis 3 was used for all subsequent hierarchical models

Supplemental Text E-2 Hierarchical model implemented in Analysis 2

```

model {
  for(i in 1:59) {
    for(j in 1:n_arms[i]) {
      y[i,j] ~ dnorm(mu[i,j] , prec.resp[i,j])
      mu[i,j] = beta[i,1] + beta[i,2] * CM[i,j]
                                     + beta[i,3] * TC[i,j]
                                     + beta[i,4] * EPR[i,j]
                                     + beta[i,5] * CE[i,j]
                                     + beta[i,6] * FR[i,j]
                                     + beta[i,7] * PE[i,j]
                                     + beta[i,8] * PSM[i,j]
                                     + beta[i,9] * PR[i,j]
                                     + beta[i,10] * Other[i,j]

      prec.resp[i,j] <- (1/se[i,j])*(1/se[i,j])
    }
    for(m in 1:10) {
      beta[i,m] ~ dnorm(mu_bar[m], prec[m])
    }
  }

  for(i in 60:68) {
    for(j in 1:n_arms[i]) {
      y[i,j] ~ dnorm(mu[i,j] , prec.resp[i,j])
      mu[i,j] = beta[i,1] + beta[i,2] * CM[i,j]
                                     + beta[i,3] * TC[i,j]
                                     + beta[i,4] * EPR[i,j]
                                     + beta[i,5] * CE[i,j]
                                     + beta[i,6] * FR[i,j]
                                     + beta[i,7] * PE[i,j]
                                     + beta[i,8] * PSM[i,j]
                                     + beta[i,9] * PR[i,j]
                                     + beta[i,10] * Other[i,j]

      prec.resp[i,j] <- (1/(se[i,j]*sqrt(corr[i,j])))*(1/(se[i,j]*sqrt(corr[i,j])))
      corr[i,j] <- 1 + (avg_cluster_size[i,j]-1) * ICC[i,j]
      ICC[i,j] <- exp(logit_ICC[i,j]) / ( 1 + exp(logit_ICC[i,j]))
      logit_ICC[i,j] ~ dnorm(-4.16224, (1/(1.8812*1.8812)))
    }
    for(m in 1:10) {
      beta[i,m] ~ dnorm(mu_bar[m], prec[m])
    }
  }

  for(i in 69:100) {
    for(j in 1:n_arms[i]) {
      y[i,j] ~ dnorm(mu[i,j] , prec.resp[i,j])
      mu[i,j] = beta[i,1] + beta[i,2] * CM[i,j]
                                     + beta[i,3] * TC[i,j]

```

```

+ beta[i,4] * EPR[i,j]
+ beta[i,5] * CE[i,j]
+ beta[i,6] * FR[i,j]
+ beta[i,7] * PE[i,j]
+ beta[i,8] * PSM[i,j]
+ beta[i,9] * PR[i,j]
+ beta[i,10] * Other[i,j]

prec.resp[i,j] <- (1/se[i,j])*(1/se[i,j])
se[i,j] ~ dunif(0,2)
}
for(m in 1:10) {
beta[i,m] ~ dnorm(mu_bar[m], prec[m])
}
}

for(i in 101:101) {
  for(j in 1:n_arms[i]) {
    y[i,j] ~ dnorm(mu[i,j] , prec.resp[i,j])
    mu[i,j] = beta[i,1] + beta[i,2] * CM[i,j]
+ beta[i,3] * TC[i,j]
+ beta[i,4] * EPR[i,j]
+ beta[i,5] * CE[i,j]
+ beta[i,6] * FR[i,j]
+ beta[i,7] * PE[i,j]
+ beta[i,8] * PSM[i,j]
+ beta[i,9] * PR[i,j]
+ beta[i,10] * Other[i,j]

    se[i,j] ~ dunif(0, 2)
    prec.resp[i,j] <- (1/(se[i,j]*sqrt(corr[i,j])))*(1/(se[i,j]*sqrt(corr[i,j])))
    corr[i,j] <- 1 + (avg_cluster_size[i,j]-1) * ICC[i,j]
  }
  for(m in 1:10) {
    beta[i,m] ~ dnorm(mu_bar[m], prec[m])
  }
}

for(i in 102:114) {
  for(j in 1:n_arms[i]) {
    y[i,j] ~ dnorm(mu[i,j] , prec.resp[i,j])
    mu[i,j] = beta[i,1] + beta[i,2] * CM[i,j]
+ beta[i,3] * TC[i,j]
+ beta[i,4] * EPR[i,j]
+ beta[i,5] * CE[i,j]
+ beta[i,6] * FR[i,j]
+ beta[i,7] * PE[i,j]
+ beta[i,8] * PSM[i,j]
+ beta[i,9] * PR[i,j]
+ beta[i,10] * Other[i,j]

    se[i,j] ~ dunif(0,2)

```

```

        prec.resp[i,j] <- (1/(se[i,j]*sqrt(corr[i,j])))*(1/(se[i,j]*sqrt(corr[i,j])))
        corr[i,j] <- 1 + (avg_cluster_size[i,j]-1) * ICC[i,j]
        ICC[i,j] <- exp(logit_ICC[i,j]) / ( 1 + exp(logit_ICC[i,j]))
        logit_ICC[i,j] ~ dnorm(-4.16224, (1/(1.8812*1.8812)))
    }
    for(m in 1:10) {
        beta[i,m] ~ dnorm(mu_bar[m], prec[m])
    }
}

mu_bar[1] ~ dnorm(8,0.01)
for(m in 2:10) {      mu_bar[m] ~ dnorm(0,0.25) }
for(m in 1:10) {
    prec[m] = (1/tau[m]) * (1/tau[m])
}
tau[1] ~ dunif(0,2)
for(m in 2:10) {      tau[m] ~ dunif(0,2) }

}

```

Supplemental Text E-3 Hierarchical model implemented in Analysis 3 (base model)

```

model {
  for(i in 1:59) {
    for(j in 1:n_arms[i]) {
      y[i,j] ~ dnorm(mu[i,j] , prec.resp[i,j])
      mu[i,j] = beta[i,1] + beta[i,2] * CM[i,j]
                                     + beta[i,3] * TC[i,j]
                                     + beta[i,4] * EPR[i,j]
                                     + beta[i,5] * CE[i,j]
                                     + beta[i,6] * FR[i,j]
                                     + beta[i,7] * PE[i,j]
                                     + beta[i,8] * PSM[i,j]
                                     + beta[i,9] * PR[i,j]
                                     + beta[i,10] * Other[i,j]

      prec.resp[i,j] <- (1/se[i,j])*(1/se[i,j])
    }
    for(m in 1:10) {
      beta[i,m] ~ dnorm(mu_bar[m], prec[m])
    }
  }

  for(i in 60:68) {
    for(j in 1:n_arms[i]) {
      y[i,j] ~ dnorm(mu[i,j] , prec.resp[i,j])
      mu[i,j] = beta[i,1] + beta[i,2] * CM[i,j]
                                     + beta[i,3] * TC[i,j]
                                     + beta[i,4] * EPR[i,j]
                                     + beta[i,5] * CE[i,j]
                                     + beta[i,6] * FR[i,j]
                                     + beta[i,7] * PE[i,j]
                                     + beta[i,8] * PSM[i,j]
                                     + beta[i,9] * PR[i,j]
                                     + beta[i,10] * Other[i,j]

      prec.resp[i,j] <- (1/(se[i,j]*sqrt(corr[i,j])))*(1/(se[i,j]*sqrt(corr[i,j])))
      corr[i,j] <- 1 + (avg_cluster_size[i,j]-1) * ICC[i,j]
      ICC[i,j] <- exp(logit_ICC[i,j]) / ( 1 + exp(logit_ICC[i,j]))
      logit_ICC[i,j] ~ dnorm(-4.16224, (1/(1.8812*1.8812)))
    }
    for(m in 1:10) {
      beta[i,m] ~ dnorm(mu_bar[m], prec[m])
    }
  }

  for(i in 69:100) {
    for(j in 1:n_arms[i]) {
      y[i,j] ~ dnorm(mu[i,j] , prec.resp[i,j])
      mu[i,j] = beta[i,1] + beta[i,2] * CM[i,j]
                                     + beta[i,3] * TC[i,j]

```

```

+ beta[i,4] * EPR[i,j]
+ beta[i,5] * CE[i,j]
+ beta[i,6] * FR[i,j]
+ beta[i,7] * PE[i,j]
+ beta[i,8] * PSM[i,j]
+ beta[i,9] * PR[i,j]
+ beta[i,10] * Other[i,j]

prec.resp[i,j] <- (1/se[i,j])*(1/se[i,j])
se[i,j] ~ dunif(0,2)
}
for(m in 1:10) {
beta[i,m] ~ dnorm(mu_bar[m], prec[m])
}
}

for(i in 101:101) {
  for(j in 1:n_arms[i]) {
    y[i,j] ~ dnorm(mu[i,j] , prec.resp[i,j])
    mu[i,j] = beta[i,1] + beta[i,2] * CM[i,j]
+ beta[i,3] * TC[i,j]
+ beta[i,4] * EPR[i,j]
+ beta[i,5] * CE[i,j]
+ beta[i,6] * FR[i,j]
+ beta[i,7] * PE[i,j]
+ beta[i,8] * PSM[i,j]
+ beta[i,9] * PR[i,j]
+ beta[i,10] * Other[i,j]

    se[i,j] ~ dunif(0, 2)
    prec.resp[i,j] <- (1/(se[i,j]*sqrt(corr[i,j])))*(1/(se[i,j]*sqrt(corr[i,j])))
    corr[i,j] <- 1 + (avg_cluster_size[i,j]-1) * ICC[i,j]
  }
  for(m in 1:10) {
    beta[i,m] ~ dnorm(mu_bar[m], prec[m])
  }
}

for(i in 102:114) {
  for(j in 1:n_arms[i]) {
    y[i,j] ~ dnorm(mu[i,j] , prec.resp[i,j])
    mu[i,j] = beta[i,1] + beta[i,2] * CM[i,j]
+ beta[i,3] * TC[i,j]
+ beta[i,4] * EPR[i,j]
+ beta[i,5] * CE[i,j]
+ beta[i,6] * FR[i,j]
+ beta[i,7] * PE[i,j]
+ beta[i,8] * PSM[i,j]
+ beta[i,9] * PR[i,j]
+ beta[i,10] * Other[i,j]

    se[i,j] ~ dunif(0,2)

```

```

        prec.resp[i,j] <- (1/(se[i,j]*sqrt(corr[i,j])))*(1/(se[i,j]*sqrt(corr[i,j])))
        corr[i,j] <- 1 + (avg_cluster_size[i,j]-1) * ICC[i,j]
        ICC[i,j] <- exp(logit_ICC[i,j]) / ( 1 + exp(logit_ICC[i,j]))
        logit_ICC[i,j] ~ dnorm(-4.16224, (1/(1.8812*1.8812)))
    }
    for(m in 1:10) {
        beta[i,m] ~ dnorm(mu_bar[m], prec[m])
    }
}

mu_bar[1] ~ dnorm(8,0.01)
for(m in 2:10) {      mu_bar[m] ~ dnorm(0,0.25) }
for(m in 1:10) {
    prec[m] = (1/tau[m]) * (1/tau[m])
}
tau[1] ~ dunif(0,2)
for(m in 2:10) {      tau[m] ~ dunif(0,2) }

rk <-rank(mu_bar[2:10])

for(m in 1:10) {
    mu_bar_new[m] ~ dnorm(mu_bar[m] , prec[m])
}

#predictive distribution
for(i in 1:114) {
    for(j in 1:n_arms[i]) {
        y_new [i,j] ~ dnorm(mu_new[i,j] , prec.resp[i,j])
        mu_new[i,j] <- mu_bar_new[1] + mu_bar_new[2] * CM[i,j]
                                + mu_bar_new[3] * TC[i,j]
                                + mu_bar_new[4] * EPR[i,j]
                                + mu_bar_new[5] * CE[i,j]
                                + mu_bar_new[6] * FR[i,j]
                                + mu_bar_new[7] * PE[i,j]
                                + mu_bar_new[8] * PSM[i,j]
                                + mu_bar_new[9] * PR[i,j]
                                + mu_bar_new[10] * Other[i,j]
    }
}

#probability model predictions exceed observed value
for(i in 1:114) {
    for(j in 1:n_arms[i]){
        p.crossval[i,j]<-step(y_new [i,j]-y[i,j])
    }
}

```

```

#predicting new combinations
#posterior distributions
#COMBO 1: CM + EPR + FR
xmu_posterior_combo_1 <-mu_bar[1] + mu_bar[2] + mu_bar[4] + mu_bar[6]
xdif_posterior_combo_1 <-mu_bar[2] + mu_bar[4] + mu_bar[6]
xbase_posterior_combo_1 <-mu_bar[1]

#COMBO 2: TC + FR + PSM
xmu_posterior_combo_2 <-mu_bar[1] + mu_bar[3] + mu_bar[6] + mu_bar[8]
xdif_posterior_combo_2 <-mu_bar[3] + mu_bar[6] + mu_bar[8]
xbase_posterior_combo_2 <-mu_bar[1]

#COMBO 3: TC + FR + EPR
xmu_posterior_combo_3 <-mu_bar[1] + mu_bar[3] + mu_bar[6] + mu_bar[4]
xdif_posterior_combo_3 <-mu_bar[3] + mu_bar[6] + mu_bar[4]
xbase_posterior_combo_3 <-mu_bar[1]

#posterior predictive distributions
#COMBO 1: CM + EPR + FR
xmu_posterior_predictive_combo_1 <-mu_bar_new[1] + mu_bar_new[2] + mu_bar_new[4] +
mu_bar_new[6]
xdif_posterior_predictive_combo_1 <-mu_bar_new[2] + mu_bar_new[4] + mu_bar_new[6]
xbase_posterior_predictive_combo_1 <-mu_bar_new[1]

#COMBO 2: TC + FR + PSM
xmu_posterior_predictive_combo_2 <-mu_bar_new[1] + mu_bar_new[3] +mu_bar_new[6] +
mu_bar_new[8]
xdif_posterior_predictive_combo_2 <-mu_bar_new[3] + mu_bar_new[6] + mu_bar_new[8]
xbase_posterior_predictive_combo_2 <-mu_bar_new[1]

#COMBO 3: TC + FR + EPR
xmu_posterior_predictive_combo_3 <-mu_bar_new[1] + mu_bar_new[3] + mu_bar_new[6] +
mu_bar_new[4]
xdif_posterior_predictive_combo_3 <-mu_bar_new[3] + mu_bar_new[6] + mu_bar_new[4]
xbase_posterior_predictive_combo_3 <-mu_bar_new[1]

}

```

Supplemental Text E-4. Extension to hierarchical model: Interactions among intervention components

NB: We ran 9 versions of this model representing pairwise interactions between a QI strategy and each of the other QI strategies. Presented below is the pairwise interaction model of case management (CM) with all other QI strategies.

```

model {
  for(i in 1:59) {
    for(j in 1:n_arms[i]) {
      y[i,j] ~ dnorm(mu[i,j] , prec.resp[i,j])
      mu[i,j] = beta[i,1] + beta[i,2] * CM[i,j]
    }
  }
}

```

```

+ beta[i,3] * TC[i,j]
+ beta[i,4] * EPR[i,j]
+ beta[i,5] * CE[i,j]
+ beta[i,6] * FR[i,j]
+ beta[i,7] * PE[i,j]
+ beta[i,8] * PSM[i,j]
+ beta[i,9] * PR[i,j]
+ beta[i,10] * Other[i,j]
+ beta[i,11] * TC[i,j] * CM[i,j]
+ beta[i,12] * EPR[i,j] * CM[i,j]
+ beta[i,13] * CE[i,j] * CM[i,j]
+ beta[i,14] * FR[i,j] * CM[i,j]
+ beta[i,15] * PE[i,j] * CM[i,j]
+ beta[i,16] * PSM[i,j] * CM[i,j]
+ beta[i,17] * PR[i,j] * CM[i,j]
+ beta[i,18] * Other[i,j] * CM[i,j]

    prec.resp[i,j] <- (1/se[i,j])*(1/se[i,j])
  }
  for(m in 1:18) {
    beta[i,m] ~ dnorm(mu_bar[m], prec[m])
  }
}

for(i in 60:68) {
  for(j in 1:n_arms[i]) {
    y[i,j] ~ dnorm(mu[i,j] , prec.resp[i,j])
    mu[i,j] = beta[i,1] + beta[i,2] * CM[i,j]
+ beta[i,3] * TC[i,j]
+ beta[i,4] * EPR[i,j]
+ beta[i,5] * CE[i,j]
+ beta[i,6] * FR[i,j]
+ beta[i,7] * PE[i,j]
+ beta[i,8] * PSM[i,j]
+ beta[i,9] * PR[i,j]
+ beta[i,10] * Other[i,j]
+ beta[i,11] * TC[i,j] * CM[i,j]
+ beta[i,12] * EPR[i,j] * CM[i,j]
+ beta[i,13] * CE[i,j] * CM[i,j]
+ beta[i,14] * FR[i,j] * CM[i,j]
+ beta[i,15] * PE[i,j] * CM[i,j]
+ beta[i,16] * PSM[i,j] * CM[i,j]
+ beta[i,17] * PR[i,j] * CM[i,j]
+ beta[i,18] * Other[i,j] * CM[i,j]

    prec.resp[i,j] <- (1/(se[i,j]*sqrt(corr[i,j])))*(1/(se[i,j]*sqrt(corr[i,j])))
    corr[i,j] <- 1 + (avg_cluster_size[i,j]-1) * ICC[i,j]
    ICC[i,j] <- exp(logit_ICC[i,j]) / ( 1 + exp(logit_ICC[i,j]))
    logit_ICC[i,j] ~ dnorm(-4.16224, (1/(1.8812*1.8812)))
  }
  for(m in 1:18) {

```

```

    beta[i,m] ~ dnorm(mu_bar[m], prec[m])
  }
}

for(i in 69:100) {
  for(j in 1:n_arms[i]) {
    y[i,j] ~ dnorm(mu[i,j] , prec.resp[i,j])
    mu[i,j] = beta[i,1] + beta[i,2] * CM[i,j]
              + beta[i,3] * TC[i,j]
              + beta[i,4] * EPR[i,j]
              + beta[i,5] * CE[i,j]
              + beta[i,6] * FR[i,j]
              + beta[i,7] * PE[i,j]
              + beta[i,8] * PSM[i,j]
              + beta[i,9] * PR[i,j]
              + beta[i,10] * Other[i,j]
              + beta[i,11] * TC[i,j] * CM[i,j]
              + beta[i,12] * EPR[i,j] * CM[i,j]
              + beta[i,13] * CE[i,j] * CM[i,j]
              + beta[i,14] * FR[i,j] * CM[i,j]
              + beta[i,15] * PE[i,j] * CM[i,j]
              + beta[i,16] * PSM[i,j] * CM[i,j]
              + beta[i,17] * PR[i,j] * CM[i,j]
              + beta[i,18] * Other[i,j] * CM[i,j]

    prec.resp[i,j] <- (1/se[i,j])*(1/se[i,j])
    se[i,j] ~ dunif(0,2)
  }
  for(m in 1:18) {
    beta[i,m] ~ dnorm(mu_bar[m], prec[m])
  }
}

for(i in 101:101) {
  for(j in 1:n_arms[i]) {
    y[i,j] ~ dnorm(mu[i,j] , prec.resp[i,j])
    mu[i,j] = beta[i,1] + beta[i,2] * CM[i,j]
              + beta[i,3] * TC[i,j]
              + beta[i,4] * EPR[i,j]
              + beta[i,5] * CE[i,j]
              + beta[i,6] * FR[i,j]
              + beta[i,7] * PE[i,j]
              + beta[i,8] * PSM[i,j]
              + beta[i,9] * PR[i,j]
              + beta[i,10] * Other[i,j]
              + beta[i,11] * TC[i,j] * CM[i,j]
              + beta[i,12] * EPR[i,j] * CM[i,j]
              + beta[i,13] * CE[i,j] * CM[i,j]
              + beta[i,14] * FR[i,j] * CM[i,j]
              + beta[i,15] * PE[i,j] * CM[i,j]

```

```

+ beta[i,16] * PSM[i,j] * CM[i,j]
+ beta[i,17] * PR[i,j] * CM[i,j]
+ beta[i,18] * Other[i,j] * CM[i,j]

se[i,j] ~ dunif(0, 2)
prec.resp[i,j] <- (1/(se[i,j]*sqrt(corr[i,j])))*(1/(se[i,j]*sqrt(corr[i,j]))))
corr[i,j] <- 1 + (avg_cluster_size[i,j]-1) * ICC[i,j]
}
for(m in 1:18) {
beta[i,m] ~ dnorm(mu_bar[m], prec[m])
}
}

for(i in 102:114) {
  for(j in 1:n_arms[i]) {
    y[i,j] ~ dnorm(mu[i,j] , prec.resp[i,j])
    mu[i,j] = beta[i,1] + beta[i,2] * CM[i,j]
+ beta[i,3] * TC[i,j]
+ beta[i,4] * EPR[i,j]
+ beta[i,5] * CE[i,j]
+ beta[i,6] * FR[i,j]
+ beta[i,7] * PE[i,j]
+ beta[i,8] * PSM[i,j]
+ beta[i,9] * PR[i,j]
+ beta[i,10] * Other[i,j]
+ beta[i,11] * TC[i,j] * CM[i,j]
+ beta[i,12] * EPR[i,j] * CM[i,j]
+ beta[i,13] * CE[i,j] * CM[i,j]
+ beta[i,14] * FR[i,j] * CM[i,j]
+ beta[i,15] * PE[i,j] * CM[i,j]
+ beta[i,16] * PSM[i,j] * CM[i,j]
+ beta[i,17] * PR[i,j] * CM[i,j]
+ beta[i,18] * Other[i,j] * CM[i,j]

    se[i,j] ~ dunif(0,2)
    prec.resp[i,j] <- (1/(se[i,j]*sqrt(corr[i,j])))*(1/(se[i,j]*sqrt(corr[i,j]))))
    corr[i,j] <- 1 + (avg_cluster_size[i,j]-1) * ICC[i,j]
    ICC[i,j] <- exp(logit_ICC[i,j]) / ( 1 + exp(logit_ICC[i,j]))
    logit_ICC[i,j] ~ dnorm(-4.16224, (1/(1.8812*1.8812)))
  }
  for(m in 1:18) {
    beta[i,m] ~ dnorm(mu_bar[m], prec[m])
  }
}

mu_bar[1] ~ dnorm(8,0.01)
for(m in 2:10) { mu_bar[m] ~ dnorm(0,0.25) }
for(m in 11:18) { mu_bar[m] ~ dnorm(0,0.50) }
for(m in 1:18) {
  prec[m] = (1/tau[m]) * (1/tau[m])
}

```

```

tau[1] ~ dunif(0,2)
for(m in 2:18) {      tau[m] ~ dunif(0,2) }

for(m in 1:18) {
  mu_bar_new[m] ~ dnorm(mu_bar[m] , prec[m])
}

#predictive distribution
for(i in 1:114) {
  for(j in 1:n_arms[i]) {
    y_new [i,j] ~ dnorm(mu_new[i,j] , prec.resp[i,j])
    mu_new[i,j] <- mu_bar_new[1] + mu_bar_new[2] * CM[i,j]
                                     + mu_bar_new[3] * TC[i,j]
                                     + mu_bar_new[4] * EPR[i,j]
                                     + mu_bar_new[5] * CE[i,j]
                                     + mu_bar_new[6] * FR[i,j]
                                     + mu_bar_new[7] * PE[i,j]
                                     + mu_bar_new[8] * PSM[i,j]
                                     + mu_bar_new[9] * PR[i,j]
                                     + mu_bar_new[10] * Other[i,j]
                                     + mu_bar_new[11] * TC[i,j] * CM[i,j]
                                     + mu_bar_new[12] * EPR[i,j] * CM[i,j]
                                     + mu_bar_new[13] * CE[i,j] * CM[i,j]
                                     + mu_bar_new[14] * FR[i,j] * CM[i,j]
                                     + mu_bar_new[15] * PE[i,j] * CM[i,j]
                                     + mu_bar_new[16] * PSM[i,j] * CM[i,j]
                                     + mu_bar_new[17] * PR[i,j] * CM[i,j]
                                     + mu_bar_new[18] * Other[i,j] * CM[i,j]
  }

}

#probability model predictions exceed observed value
for(i in 1:114) {
  for(j in 1:n_arms[i]){
    p.crossval[i,j]<-step(y_new [i,j]-y[i,j])
  }

}

```

Supplemental Text E-5 Extension to hierarchical model: Effect modification – binary HbA1c

```

model {
  for(i in 1:59) {
    for(j in 1:n_arms[i]) {
      y[i,j] ~ dnorm(mu[i,j] , prec.resp[i,j])
      mu[i,j] = beta[i,1] + beta[i,2] * CM[i,j]
                                     + beta[i,3] * TC[i,j]
                                     + beta[i,4] * EPR[i,j]
                                     + beta[i,5] * CE[i,j]
                                     + beta[i,6] * FR[i,j]
                                     + beta[i,7] * PE[i,j]
                                     + beta[i,8] * PSM[i,j]
                                     + beta[i,9] * PR[i,j]
                                     + beta[i,10] * Other[i,j]
                                     + beta [i,11] * Baseline_un [i]
                                     + beta[i,12] * CM[i,j] * Baseline_un [i]
                                     + beta[i,13] * TC[i,j] * Baseline_un [i]
                                     + beta[i,14] * EPR[i,j] * Baseline_un [i]
                                     + beta[i,15] * CE[i,j] * Baseline_un [i]
                                     + beta[i,16] * FR[i,j] * Baseline_un [i]
                                     + beta[i,17] * PE[i,j] * Baseline_un [i]
                                     + beta[i,18] * PSM[i,j] * Baseline_un [i]
                                     + beta[i,19] * PR[i,j] * Baseline_un [i]
                                     + beta[i,20] * Other[i,j] * Baseline_un
    }
  }

  [i]

      prec.resp[i,j] <- (1/se[i,j])*(1/se[i,j])
    }
    for(m in 1:20) {
      beta[i,m] ~ dnorm(mu_bar[m], prec[m])
    }
  }

  for(i in 60:68) {
    for(j in 1:n_arms[i]) {
      y[i,j] ~ dnorm(mu[i,j] , prec.resp[i,j])
      mu[i,j] = beta[i,1] + beta[i,2] * CM[i,j]
                                     + beta[i,3] * TC[i,j]
                                     + beta[i,4] * EPR[i,j]
                                     + beta[i,5] * CE[i,j]
                                     + beta[i,6] * FR[i,j]
                                     + beta[i,7] * PE[i,j]
                                     + beta[i,8] * PSM[i,j]
                                     + beta[i,9] * PR[i,j]
                                     + beta[i,10] * Other[i,j]
                                     + beta [i,11] * Baseline_un [i]
                                     + beta[i,12] * CM[i,j] * Baseline_un [i]
                                     + beta[i,13] * TC[i,j] * Baseline_un [i]

```

```

+ beta[i,14] * EPR[i,j] * Baseline_un [i]
+ beta[i,15] * CE[i,j] * Baseline_un [i]
+ beta[i,16] * FR[i,j] * Baseline_un [i]
+ beta[i,17] * PE[i,j] * Baseline_un [i]
+ beta[i,18] * PSM[i,j] * Baseline_un [i]
+ beta[i,19] * PR[i,j] * Baseline_un [i]
+ beta[i,20] * Other[i,j] * Baseline_un

[i]

prec.resp[i,j] <- (1/(se[i,j]*sqrt(corr[i,j])))*(1/(se[i,j]*sqrt(corr[i,j])))
corr[i,j] <- 1 + (avg_cluster_size[i,j]-1) * ICC[i,j]
ICC[i,j] <- exp(logit_ICC[i,j]) / ( 1 + exp(logit_ICC[i,j]))
logit_ICC[i,j] ~ dnorm(-4.16224, (1/(1.8812*1.8812)))
}
for(m in 1:20) {
beta[i,m] ~ dnorm(mu_bar[m], prec[m])
}
}

for(i in 69:100) {
for(j in 1:n_arms[i]) {
y[i,j] ~ dnorm(mu[i,j] , prec.resp[i,j])
mu[i,j] = beta[i,1] + beta[i,2] * CM[i,j]
+ beta[i,3] * TC[i,j]
+ beta[i,4] * EPR[i,j]
+ beta[i,5] * CE[i,j]
+ beta[i,6] * FR[i,j]
+ beta[i,7] * PE[i,j]
+ beta[i,8] * PSM[i,j]
+ beta[i,9] * PR[i,j]
+ beta[i,10] * Other[i,j]
+ beta [i,11] * Baseline_un [i]
+ beta[i,12] * CM[i,j] * Baseline_un [i]
+ beta[i,13] * TC[i,j] * Baseline_un [i]
+ beta[i,14] * EPR[i,j] * Baseline_un [i]
+ beta[i,15] * CE[i,j] * Baseline_un [i]
+ beta[i,16] * FR[i,j] * Baseline_un [i]
+ beta[i,17] * PE[i,j] * Baseline_un [i]
+ beta[i,18] * PSM[i,j] * Baseline_un [i]
+ beta[i,19] * PR[i,j] * Baseline_un [i]
+ beta[i,20] * Other[i,j] * Baseline_un

[i]

prec.resp[i,j] <- (1/se[i,j])*(1/se[i,j])
se[i,j] ~ dunif(0,2)
}
for(m in 1:20) {
beta[i,m] ~ dnorm(mu_bar[m], prec[m])
}
}

```

```

for(i in 101:101) {
  for(j in 1:n_arms[i]) {
    y[i,j] ~ dnorm(mu[i,j] , prec.resp[i,j])
    mu[i,j] = beta[i,1] + beta[i,2] * CM[i,j]
              + beta[i,3] * TC[i,j]
              + beta[i,4] * EPR[i,j]
              + beta[i,5] * CE[i,j]
              + beta[i,6] * FR[i,j]
              + beta[i,7] * PE[i,j]
              + beta[i,8] * PSM[i,j]
              + beta[i,9] * PR[i,j]
              + beta[i,10] * Other[i,j]
              + beta [i,11] * Baseline_un [i]
              + beta[i,12] * CM[i,j] * Baseline_un [i]
              + beta[i,13] * TC[i,j] * Baseline_un [i]
              + beta[i,14] * EPR[i,j] * Baseline_un [i]
              + beta[i,15] * CE[i,j] * Baseline_un [i]
              + beta[i,16] * FR[i,j] * Baseline_un [i]
              + beta[i,17] * PE[i,j] * Baseline_un [i]
              + beta[i,18] * PSM[i,j] * Baseline_un [i]
              + beta[i,19] * PR[i,j] * Baseline_un [i]
              + beta[i,20] * Other[i,j] * Baseline_un

[i]

    se[i,j] ~ dunif(0, 2)
    prec.resp[i,j] <- (1/(se[i,j]*sqrt(corr[i,j])))*(1/(se[i,j]*sqrt(corr[i,j])))
    corrf[i,j] <- 1 + (avg_cluster_size[i,j]-1) * ICC[i,j]
  }
  for(m in 1:20) {
    beta[i,m] ~ dnorm(mu_bar[m], prec[m])
  }
}

for(i in 102:114) {
  for(j in 1:n_arms[i]) {
    y[i,j] ~ dnorm(mu[i,j] , prec.resp[i,j])
    mu[i,j] = beta[i,1] + beta[i,2] * CM[i,j]
              + beta[i,3] * TC[i,j]
              + beta[i,4] * EPR[i,j]
              + beta[i,5] * CE[i,j]
              + beta[i,6] * FR[i,j]
              + beta[i,7] * PE[i,j]
              + beta[i,8] * PSM[i,j]
              + beta[i,9] * PR[i,j]
              + beta[i,10] * Other[i,j]
              + beta [i,11] * Baseline_un [i]
              + beta[i,12] * CM[i,j] * Baseline_un [i]
              + beta[i,13] * TC[i,j] * Baseline_un [i]
              + beta[i,14] * EPR[i,j] * Baseline_un [i]
              + beta[i,15] * CE[i,j] * Baseline_un [i]

```

```

+ beta[i,16] * FR[i,j] * Baseline_un [i]
+ beta[i,17] * PE[i,j] * Baseline_un [i]
+ beta[i,18] * PSM[i,j] * Baseline_un [i]
+ beta[i,19] * PR[i,j] * Baseline_un [i]
+ beta[i,20] * Other[i,j] * Baseline_un

[i]
  se[i,j] ~ dunif(0,2)
  prec.resp[i,j] <- (1/(se[i,j]*sqrt(corr[i,j])))*(1/(se[i,j]*sqrt(corr[i,j]))))
  corr[i,j] <- 1 + (avg_cluster_size[i,j]-1) * ICC[i,j]
  ICC[i,j] <- exp(logit_ICC[i,j]) / ( 1 + exp(logit_ICC[i,j]))
  logit_ICC[i,j] ~ dnorm(-4.16224, (1/(1.8812*1.8812)))
}
for(m in 1:20) {
  beta[i,m] ~ dnorm(mu_bar[m], prec[m])
}
}

mu_bar[1] ~ dnorm(8,0.01)
for(m in 2:20) {      mu_bar[m] ~ dnorm(0,0.25) }
for(m in 1:20) {
  prec[m] = (1/tau[m]) * (1/tau[m])
}
tau[1] ~ dunif(0,2)
for(m in 2:20) {      tau[m] ~ dunif(0,2) }

for(m in 1:20) {
  mu_bar_new[m] ~ dnorm(mu_bar[m] , prec[m])
}

#predictive distribution
for(i in 1:114) {
  for(j in 1:n_arms[i]) {
    y_new [i,j] ~ dnorm(mu_new[i,j] , prec.resp[i,j])
    mu_new[i,j] <- mu_bar_new[1] + mu_bar_new[2] * CM[i,j]
+ mu_bar_new[3] * TC[i,j]
+ mu_bar_new[4] * EPR[i,j]
+ mu_bar_new[5] * CE[i,j]
+ mu_bar_new[6] * FR[i,j]
+ mu_bar_new[7] * PE[i,j]
+ mu_bar_new[8] * PSM[i,j]
+ mu_bar_new[9] * PR[i,j]
+ mu_bar_new[10] * Other[i,j]
+ mu_bar_new [11] * Baseline_un [i]
+ mu_bar_new[12] * CM[i,j] *

Baseline_un [i]
+ mu_bar_new[13] * TC[i,j] *

Baseline_un [i]
+ mu_bar_new[14] * EPR[i,j] *

Baseline_un [i]

```

```

Baseline_un [i] + mu_bar_new[15] * CE[i,j] *
Baseline_un [i] + mu_bar_new[16] * FR[i,j] *
Baseline_un [i] + mu_bar_new[17] * PE[i,j] *
Baseline_un [i] + mu_bar_new[18] * PSM[i,j] *
Baseline_un [i] + mu_bar_new[19] * PR[i,j] *
Baseline_un [i] + mu_bar_new[20] * Other[i,j] *
Baseline_un [i]
}
}
#probability model predictions exceed observed value
for(i in 1:114) {
  for(j in 1:n_arms[i]){
    p.crossval[i,j]<-step(y_new [i,j]-y[i,j])
  }}

#CM
#uncontrolled+CM
z2_CM_U_Int<-mu_bar[1]+mu_bar[2]+mu_bar[11]+mu_bar[12]

#uncontrolled+NOCM
z2_CM_U_NOInt<-mu_bar[1]+mu_bar[11]

#differenceinuncontrolled
z2_CM_Dif_U<-mu_bar[2]+mu_bar[12]

#controlled+CM
z2_CM_C_Int<-mu_bar[1]+mu_bar[2]

#controlled+NOCM
z2_CM_C<-mu_bar[1]

#differenceincontrolled
z2_CM_Dif_C<-mu_bar[2]

#differenceofdifference
z2_Diff_of_diff_CM<-z2_CM_Dif_C-z2_CM_Dif_U

#TC
#uncontrolled+TC
z3_TC_U_Int<-mu_bar[1]+mu_bar[3]+mu_bar[11]+mu_bar[13]

#uncontrolled+NOTC
z3_TC_U_NOInt<-mu_bar[1]+mu_bar[11]

```

```

#differenceinuncontrolled
z3_TC_Dif_U<-mu_bar[3]+mu_bar[13]

#controlled+TC
z3_TC_C_Int<-mu_bar[1]+mu_bar[3]

#controlled+NOTC
z3_TC_C<-mu_bar[1]

#differenceincontrolled
z3_TC_Dif_C<-mu_bar[3]

#differenceofdifference
z3_Diff_of_diff_TC<-z3_TC_Dif_C-z3_TC_Dif_U

#EPR
#uncontrolled+EPR
z4_EPR_U_Int<-mu_bar[1]+mu_bar[4]+mu_bar[11]+mu_bar[14]

#uncontrolled+NOEPR
z4_EPR_U_NOInt<-mu_bar[1]+mu_bar[11]

#differenceinuncontrolled
z4_EPR_Dif_U<-mu_bar[4]+mu_bar[14]

#controlled+EPR
z4_EPR_C_Int<-mu_bar[1]+mu_bar[4]

#controlled+NOEPR
z4_EPR_C<-mu_bar[1]

#differenceincontrolled
z4_EPR_Dif_C<-mu_bar[4]

#differenceofdifference
z4_Diff_of_diff_EPR<-z4_EPR_Dif_C- z4_EPR_Dif_U

#CE
#uncontrolled+CE
z5_CE_U_Int<-mu_bar[1]+mu_bar[5]+mu_bar[11]+mu_bar[15]

#uncontrolled+NOCE
z5_CE_U_NOInt<-mu_bar[1]+mu_bar[11]

#differenceinuncontrolled
z5_CE_Dif_U<-mu_bar[5]+mu_bar[15]

#controlled+CE
z5_CE_C_Int<-mu_bar[1]+mu_bar[5]

```

```

#controlled+NOCE
z5_CE_C<-mu_bar[1]

#differenceincontrolled
z5_CE_Dif_C<-mu_bar[5]

#differenceofdifference
z5_Diff_of_diff_CE<-z5_CE_Dif_C-z5_CE_Dif_U

#FR
#uncontrolled+FR
z6_FR_U_Int<-mu_bar[1]+mu_bar[6]+mu_bar[11]+mu_bar[16]

#uncontrolled+NOFR
z6_FR_U_NOInt<-mu_bar[1]+mu_bar[11]

#differenceinuncontrolled
z6_FR_Dif_U<-mu_bar[6]+mu_bar[16]

#controlled+FR
z6_FR_C_Int<-mu_bar[1]+mu_bar[6]

#controlled+NOFR
z6_FR_C<-mu_bar[1]

#differenceincontrolled
z6_FR_Dif_C<-mu_bar[6]

#differenceofdifference
z6_Diff_of_diff_FR<-z6_FR_Dif_C-z6_FR_Dif_U

#PE
#uncontrolled+PE
z7_PE_U_Int<-mu_bar[1]+mu_bar[7]+mu_bar[11]+mu_bar[17]

#uncontrolled+NOPE
z7_PEU_NOInt<-mu_bar[1]+mu_bar[11]

#differenceinuncontrolled
z7_PE_Dif_U<-mu_bar[7]+mu_bar[17]

#controlled+PE
z7_PE_C_Int<-mu_bar[1]+mu_bar[7]

#controlled+NOPE
z7_PE_C<-mu_bar[1]

#differenceincontrolled

```

```

z7_PE_Dif_C<-mu_bar[7]

#differenceofdifference
z7_Diff_of_diff_PE<-z7_PE_Dif_C-z7_PE_Dif_U

#PSM
#uncontrolled+PSM
z8_PSM_U_Int<-mu_bar[1]+mu_bar[8]+mu_bar[11]+mu_bar[18]

#uncontrolled+NOPSM
z8_PSM_U_NOInt<-mu_bar[1]+mu_bar[11]

#differenceinuncontrolled
z8_PSM_Dif_U<-mu_bar[8]+mu_bar[18]

#controlled+PSM
z8_PSM_C_Int<-mu_bar[1]+mu_bar[8]

#controlled+NOPSM
z8_PSM_C<-mu_bar[1]

#differenceincontrolled
z8_PSM_Dif_C<-mu_bar[8]

#differenceofdifference
z8_Diff_of_diff_PSM<-z8_PSM_Dif_C-z8_PSM_Dif_U

#PR
#uncontrolled+PR
z9_PR_U_Int<-mu_bar[1]+mu_bar[9]+mu_bar[11]+mu_bar[19]

#uncontrolled+NOPR
z9_PR_U_NOInt<-mu_bar[1]+mu_bar[11]

#differenceinuncontrolled
z9_PR_Dif_U<-mu_bar[9]+mu_bar[19]

#controlled+PR
z9_PR_C_Int<-mu_bar[1]+mu_bar[9]

#controlled+NOPR
z9_PR_C<-mu_bar[1]

#differenceincontrolled
z9_PR_Dif_C<-mu_bar[9]

#differenceofdifference
z9_Diff_of_diff_PR<-z9_PR_Dif_C-z9_PR_Dif_U

```

```

#Other
#uncontrolled+Other
z10_Other_U_Int<-mu_bar[1]+mu_bar[10]+mu_bar[11]+mu_bar[20]

#uncontrolled+NOOther
z10_Other_U_NOInt<-mu_bar[1]+mu_bar[11]

#differenceinuncontrolled
z10_Other_Dif_U<-mu_bar[10]+mu_bar[20]

#controlled+Other
z10_Other_C_Int<-mu_bar[1]+mu_bar[10]

#controlled+NOOther
z10_Other_C<-mu_bar[1]

#differenceincontrolled
z10_Other_Dif_C<-mu_bar[10]

#differenceofdifference
z10_Diff_of_diff_Other<-z10_Other_Dif_C-z10_Other_Dif_U
}

```

Supplemental Text E-6 Extension to hierarchical model: Effect modification – continuous HbA1c

```

model {
  for(i in 1:59) {
    for(j in 1:n_arms[i]) {
      y[i,j] ~ dnorm(mu[i,j] , prec.resp[i,j])
      mu[i,j] = beta[i,1] + beta[i,2] * CM[i,j]
                                     + beta[i,3] * TC[i,j]
                                     + beta[i,4] * EPR[i,j]
                                     + beta[i,5] * CE[i,j]
                                     + beta[i,6] * FR[i,j]
                                     + beta[i,7] * PE[i,j]
                                     + beta[i,8] * PSM[i,j]
                                     + beta[i,9] * PR[i,j]
                                     + beta[i,10] * Other[i,j]
                                     + beta[i,11] * (Baseline_risk_study[i] -
Baseline_risk_sample[i])
                                     + beta[i,12] * CM[i,j] *
(Baseline_risk_study[i] - Baseline_risk_sample[i])
                                     + beta[i,13] * TC[i,j] *
(Baseline_risk_study[i] - Baseline_risk_sample[i])
                                     + beta[i,14] * EPR[i,j] *
(Baseline_risk_study[i] - Baseline_risk_sample[i])
                                     + beta[i,15] * CE[i,j] *
(Baseline_risk_study[i] - Baseline_risk_sample[i])
                                     + beta[i,16] * FR[i,j] *
(Baseline_risk_study[i] - Baseline_risk_sample[i])
                                     + beta[i,17] * PE[i,j] *
(Baseline_risk_study[i] - Baseline_risk_sample[i])
                                     + beta[i,18] * PSM[i,j] *
(Baseline_risk_study[i] - Baseline_risk_sample[i])
                                     + beta[i,19] * PR[i,j] *
(Baseline_risk_study[i] - Baseline_risk_sample[i])
                                     + beta[i,20] * Other[i,j] *
(Baseline_risk_study[i] - Baseline_risk_sample[i])

      prec.resp[i,j] <- (1/se[i,j])*(1/se[i,j])
    }
    for(m in 1:20) {
      beta[i,m] ~ dnorm(mu_bar[m], prec[m])
    }
  }
}

for(i in 60:68) {
  for(j in 1:n_arms[i]) {
    y[i,j] ~ dnorm(mu[i,j] , prec.resp[i,j])
    mu[i,j] = beta[i,1] + beta[i,2] * CM[i,j]
                                     + beta[i,3] * TC[i,j]
                                     + beta[i,4] * EPR[i,j]

```

```

+ beta[i,5] * CE[i,j]
+ beta[i,6] * FR[i,j]
+ beta[i,7] * PE[i,j]
+ beta[i,8] * PSM[i,j]
+ beta[i,9] * PR[i,j]
+ beta[i,10] * Other[i,j]
+ beta [i,11] * (Baseline_risk_study[i] -
Baseline_risk_sample[i])
+ beta[i,12] * CM[i,j] *
(Baseline_risk_study[i] - Baseline_risk_sample[i])
+ beta[i,13] * TC[i,j] *
(Baseline_risk_study[i] - Baseline_risk_sample[i])
+ beta[i,14] * EPR[i,j] *
(Baseline_risk_study[i] - Baseline_risk_sample[i])
+ beta[i,15] * CE[i,j] *
(Baseline_risk_study[i] - Baseline_risk_sample[i])
+ beta[i,16] * FR[i,j] *
(Baseline_risk_study[i] - Baseline_risk_sample[i])
+ beta[i,17] * PE[i,j] *
(Baseline_risk_study[i] - Baseline_risk_sample[i])
+ beta[i,18] * PSM[i,j] *
(Baseline_risk_study[i] - Baseline_risk_sample[i])
+ beta[i,19] * PR[i,j] *
(Baseline_risk_study[i] - Baseline_risk_sample[i])
+ beta[i,20] * Other[i,j] *

prec.resp[i,j] <- (1/(se[i,j]*sqrt(corr[i,j])))*(1/(se[i,j]*sqrt(corr[i,j]))))
corr[i,j] <- 1 + (avg_cluster_size[i,j]-1) * ICC[i,j]
ICC[i,j] <- exp(logit_ICC[i,j]) / ( 1 + exp(logit_ICC[i,j]))
logit_ICC[i,j] ~ dnorm(-4.16224, (1/(1.8812*1.8812)))
}
for(m in 1:20) {
beta[i,m] ~ dnorm(mu_bar[m], prec[m])
}
}

for(i in 69:100) {
for(j in 1:n_arms[i]) {
y[i,j] ~ dnorm(mu[i,j] , prec.resp[i,j])
mu[i,j] = beta[i,1] + beta[i,2] * CM[i,j]
+ beta[i,3] * TC[i,j]
+ beta[i,4] * EPR[i,j]
+ beta[i,5] * CE[i,j]
+ beta[i,6] * FR[i,j]
+ beta[i,7] * PE[i,j]
+ beta[i,8] * PSM[i,j]
+ beta[i,9] * PR[i,j]
+ beta[i,10] * Other[i,j]

```

```

Baseline_risk_sample[i])
(Baseline_risk_study[i] - Baseline_risk_sample[i])
(Baseline_risk_study[i] - Baseline_risk_sample[i])
(Baseline_risk_study[i] - Baseline_risk_sample[i])
(Baseline_risk_study[i] - Baseline_risk_sample[i])
(Baseline_risk_study[i] - Baseline_risk_sample[i])
(Baseline_risk_study[i] - Baseline_risk_sample[i])
(Baseline_risk_study[i] - Baseline_risk_sample[i])
(Baseline_risk_study[i] - Baseline_risk_sample[i])
(Baseline_risk_study[i] - Baseline_risk_sample[i])

      prec.resp[i,j] <- (1/se[i,j])*(1/se[i,j])
      se[i,j] ~ dunif(0,2)
    }
    for(m in 1:20) {
      beta[i,m] ~ dnorm(mu_bar[m], prec[m])
    }
  }

for(i in 101:101) {
  for(j in 1:n_arms[i]) {
    y[i,j] ~ dnorm(mu[i,j] , prec.resp[i,j])
    mu[i,j] = beta[i,1] + beta[i,2] * CM[i,j]
              + beta[i,3] * TC[i,j]
              + beta[i,4] * EPR[i,j]
              + beta[i,5] * CE[i,j]
              + beta[i,6] * FR[i,j]
              + beta[i,7] * PE[i,j]
              + beta[i,8] * PSM[i,j]
              + beta[i,9] * PR[i,j]
              + beta[i,10] * Other[i,j]
              + beta[i,11] * (Baseline_risk_study[i] -
              Baseline_risk_sample[i])
              + beta[i,12] * CM[i,j] *
              + beta[i,13] * TC[i,j] *
              + beta[i,14] * EPR[i,j] *

```

```

(Baseline_risk_study[i] - Baseline_risk_sample[i])
(Baseline_risk_study[i] - Baseline_risk_sample[i])
(Baseline_risk_study[i] - Baseline_risk_sample[i])
(Baseline_risk_study[i] - Baseline_risk_sample[i])
(Baseline_risk_study[i] - Baseline_risk_sample[i])
(Baseline_risk_study[i] - Baseline_risk_sample[i])

+ beta[i,15] * CE[i,j] *
+ beta[i,16] * FR[i,j] *
+ beta[i,17] * PE[i,j] *
+ beta[i,18] * PSM[i,j] *
+ beta[i,19] * PR[i,j] *
+ beta[i,20] * Other[i,j] *

se[i,j] ~ dunif(0, 2)
prec.resp[i,j] <- (1/(se[i,j]*sqrt(corr[i,j])))*(1/(se[i,j]*sqrt(corr[i,j]))))
corr[i,j] <- 1 + (avg_cluster_size[i,j]-1) * ICC[i,j]
}
for(m in 1:20) {
  beta[i,m] ~ dnorm(mu_bar[m], prec[m])
}
}

for(i in 102:114) {
  for(j in 1:n_arms[i]) {
    y[i,j] ~ dnorm(mu[i,j], prec.resp[i,j])
    mu[i,j] = beta[i,1] + beta[i,2] * CM[i,j]
+ beta[i,3] * TC[i,j]
+ beta[i,4] * EPR[i,j]
+ beta[i,5] * CE[i,j]
+ beta[i,6] * FR[i,j]
+ beta[i,7] * PE[i,j]
+ beta[i,8] * PSM[i,j]
+ beta[i,9] * PR[i,j]
+ beta[i,10] * Other[i,j]
+ beta[i,11] * (Baseline_risk_study[i] -
Baseline_risk_sample[i])
+ beta[i,12] * CM[i,j] *
+ beta[i,13] * TC[i,j] *
+ beta[i,14] * EPR[i,j] *
+ beta[i,15] * CE[i,j] *
+ beta[i,16] * FR[i,j] *
+ beta[i,17] * PE[i,j] *
(Baseline_risk_study[i] - Baseline_risk_sample[i])

```

```

+ beta[i,18] * PSM[i,j] *
(Baseline_risk_study[i] - Baseline_risk_sample[i])
+ beta[i,19] * PR[i,j] *
(Baseline_risk_study[i] - Baseline_risk_sample[i])
+ beta[i,20] * Other[i,j] *
(Baseline_risk_study[i] - Baseline_risk_sample[i])

se[i,j] ~ dunif(0,2)
prec.resp[i,j] <- (1/(se[i,j]*sqrt(corr[i,j])))*(1/(se[i,j]*sqrt(corr[i,j]))))
corr[i,j] <- 1 + (avg_cluster_size[i,j]-1) * ICC[i,j]
ICC[i,j] <- exp(logit_ICC[i,j]) / ( 1 + exp(logit_ICC[i,j]))
logit_ICC[i,j] ~ dnorm(-4.16224, (1/(1.8812*1.8812)))
}
for(m in 1:20) {
beta[i,m] ~ dnorm(mu_bar[m], prec[m])
}
}

mu_bar[1] ~ dnorm(8,0.01)
for(m in 2:20) { mu_bar[m] ~ dnorm(0,0.25) }
for(m in 1:20) {
prec[m] = (1/tau[m]) * (1/tau[m])
}
tau[1] ~ dunif(0,2)
for(m in 2:20) { tau[m] ~ dunif(0,2) }

rk <-rank(mu_bar[2:20])

for(m in 1:20) {
mu_bar_new[m] ~ dnorm(mu_bar[m] , prec[m])
}

#predictive distribution
for(i in 1:114) {
for(j in 1:n_arms[i]) {
y_new [i,j] ~ dnorm(mu_new[i,j] , prec.resp[i,j])
mu_new[i,j] <- mu_bar_new[1] + mu_bar_new[2] * CM[i,j]
+ mu_bar_new[3] * TC[i,j]
+ mu_bar_new[4] * EPR[i,j]
+ mu_bar_new[5] * CE[i,j]
+ mu_bar_new[6] * FR[i,j]
+ mu_bar_new[7] * PE[i,j]
+ mu_bar_new[8] * PSM[i,j]
+ mu_bar_new[9] * PR[i,j]

```

```

+ mu_bar_new[10] * Other[i,j]
+ mu_bar_new[11] * (Baseline_risk_study[i] -
Baseline_risk_sample[i])
+ mu_bar_new[12] * CM[i,j] * (Baseline_risk_study[i] -
Baseline_risk_sample[i])
+ mu_bar_new[13] * TC[i,j] * (Baseline_risk_study[i] -
Baseline_risk_sample[i])
+ mu_bar_new[14] * EPR[i,j] * (Baseline_risk_study[i]
- Baseline_risk_sample[i])
+ mu_bar_new[15] * CE[i,j] * (Baseline_risk_study[i] -
Baseline_risk_sample[i])
+ mu_bar_new[16] * FR[i,j] * (Baseline_risk_study[i] -
Baseline_risk_sample[i])
+ mu_bar_new[17] * PE[i,j] * (Baseline_risk_study[i] -
Baseline_risk_sample[i])
+ mu_bar_new[18] * PSM[i,j] * (Baseline_risk_study[i]
- Baseline_risk_sample[i])
+ mu_bar_new[19] * PR[i,j] * (Baseline_risk_study[i] -
Baseline_risk_sample[i])
+ mu_bar_new[20] * Other[i,j] *
(Baseline_risk_study[i] - Baseline_risk_sample[i])
}
}

#probability model predictions exceed observed value
for(i in 1:114) {
  for(j in 1:n_arms[i]){
    p.crossval[i,j]<-step(y_new [i,j]-y[i,j])
  }
}
}

```