American Journal of Epidemiology

Vol. 193, No. 2
https://doi.org/10.1093/aje/kwad184
Advance Access publication:
September 8, 2023

## Practice of Epidemiology

# Evidence Synthesis for Complex Interventions Using Meta-Regression Models

**Kristin J. Konnyu*, Jeremy M. Grimshaw, Thomas A. Trikalinos, Noah M. Ivers, David Moher, and Issa J. Dahabreh**

* Correspondence to Dr. Kristin J. Konnyu, Health Services Research Unit, University of Aberdeen, 3rd Floor, Health Sciences Building, Foresterhill, Aberdeen AB25 2ZD, United Kingdom (e-mail: kristin.konnyu@abdn.ac.uk).

*Initially submitted January 6, 2023; accepted for publication August 22, 2023.*

A goal of evidence synthesis for trials of complex interventions is to inform the design or implementation of novel versions of complex interventions by predicting expected outcomes with each intervention version. Conventional aggregate data meta-analyses of studies comparing complex interventions have limited ability to provide such information. We argue that evidence synthesis for trials of complex interventions should forgo aspirations of estimating causal effects and instead model the response surface of study results to 1) summarize the available evidence and 2) predict the average outcomes of future studies or in new settings. We illustrate this modeling approach using data from a systematic review of diabetes quality improvement (QI) interventions involving at least 1 of 12 QI strategy components. We specify a series of meta-regression models to assess the association of specific components with the posttreatment outcome mean and compare the results to conventional meta-analysis approaches. Compared with conventional approaches, modeling the response surface of study results can better reflect the associations between intervention components and study characteristics with the posttreatment outcome mean. Modeling study results using a response surface approach offers a useful and feasible goal for evidence synthesis of complex interventions that rely on aggregate data.

complex interventions; hierarchical models; meta-analysis; meta-regression; multicomponent interventions

Abbreviations: CrI, credible interval; HbA1c, hemoglobin A1c; ICC, intraclass correlation coefficient; QI, quality improvement.

Interventions designed to change health-care practice and health policy are often “complex” in that they involve multiple components delivered to multiple levels of health care (e.g., patients, providers, clinics) and across multiple settings (1, 2). Although researchers continue to refine the definition of a complex intervention (1–5), most definitions require 3 features: 1) complex interventions involve combinations of several more-elemental intervention “components”; 2) there can be intricate interactions between the components and/or any ancillary cointerventions; and 3) the components may have effects that vary with study-specific characteristics, such as health-care settings and patient populations—that is, the components may exhibit different treatment effect heterogeneity patterns (6, 7).

The above way of thinking about complex interventions is intuitive and not particularly restrictive. It can describe interventions considered for a wide range of policy, health-care delivery, and operations management problems, from what behavioral measures to roll out in the beginning of the coronavirus disease 2019 epidemic to managing patients with diabetes (8), developing decision aids for patients facing difficult decisions (9), or proposing combination chemotherapy regimens to be evaluated in future trials (10). It also suggests an obvious organizing scheme for describing how a complex intervention works and theorizing about which modifications might be improvements. Behavioral-intervention theorists and implementation scientists often employ similar modular approaches to designing their interventions (11, 12).

A naturally arising need is to predict average outcomes with different versions of a complex intervention in a target setting. Prediction of average outcomes sets a lower bar than estimation of causal effects. Except for narrow questions, results of evidence synthesis of aggregate data do not have a clear causal interpretation. Causally interpretable meta-analyses involve estimating the effects of well-defined interventions in a well-defined target population—a challenging task that requires strong structural assumptions and rich individual participant data (13, 14).

Our emphasis on prediction of posttreatment average outcomes with complex interventions differs from the stated goals of other methodological (15, 16) and applied (17–23) works that aim to estimate causal effects for each component of a complex intervention by modeling the between-arm differences in each trial. Other authors have proposed the use of multivariable meta-regression models as an alternative synthesis approach for complex interventions (24, 25), reframing the goal of synthesis away from estimation of any single causal effect to the estimation of a *response surface* (25) conditional on the individual intervention components and on population and setting characteristics of interest. In this paper, we adopt the latter view: We argue that in most cases, particularly when using aggregate data from very diverse studies, the goal of evidence synthesis should be not to estimate a single causal effect but to estimate a function of the different components and, potentially, their interactions with population characteristics and contextual factors, that can be used to predict the average outcomes of a complex intervention of interest in a future setting or study.

We illustrate an application of the above in a systematic review and meta-analysis of complex interventions for quality improvement (QI) of diabetes management. We first introduce the application and discuss limitations of conventional meta-analysis. We then develop and apply response-surface meta-regression models and discuss estimation, prediction, and ranking. We examine alternative model specifications that allow interactions among intervention components and with study-level covariates. We also discuss handling of discrete outcomes and missing data, including missing estimates of the intraclass correlation coefficient (ICC), to adjust variance data in cluster-randomized trials. Last, we discuss the strengths and limitations of our approach.

## MOTIVATING EXAMPLE: QI 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 (26). People living with diabetes are at increased risk for serious complications, such as cardiovascular events and blindness. Despite evidence that clinical interventions such as monitoring glycemic control, monitoring microvascular complications, and managing vascular risk factors improve patient outcomes and reduce costs (27, 28), many patients with diabetes do not receive evidence-based care and have suboptimal control of risk factors (29, 30). Diabetes QI interventions seek to address these evidence-to-practice gaps by targeting system, provider, or patient factors influencing diabetes care (8). These QI interventions fit our definition of complex interventions—their different versions are combinations of a subset of 12 more-elemental components (8).

As an example, we use data from a systematic review by Tricco et al. (8) that examined evidence from 142 randomized trials of the effects of QI interventions (comprised of component QI strategies) on a range of procedural (e.g., foot screening) and intermediate patient (e.g., glycemic control) outcomes. The review codes QI interventions using a taxonomy of 12 component strategies, adapted from Cochrane's Effective Practice and Organization of Care 2002 taxonomy (Table 1) (8, 31, 32). Most included trials evaluated intervention versions with a median of 3 QI components (range, 1–8). For the main outcome, levels of hemoglobin A1c (HbA1c; a measure of glycemic control), the authors found that in 120 trials that compared using a QI intervention with not using one, the average mean reduction in HbA1c levels was 0.37% (95% confidence interval: 0.28, 0.45) in random-effects meta-analysis, but with evidence of substantial heterogeneity ($I^2$ statistic = 73%). Analyses assessing the efficacy of QI interventions containing a specific component of interest (e.g., case management), as compared with QI interventions not containing that component (e.g., no case management), found improvements associated with most QI components but could not disentangle the effects of co-occurring components. Finally, the meta-analysis did not assess nonadditivity in the relationship between intervention components and the posttreatment outcome mean or examine modification of the association between each intervention component and the posttreatment outcome mean (also referred to as "moderation") by study-level covariates. Thus, despite the ostensibly large number of studies, the authors of the review could not explain the observed heterogeneity or predict the outcomes of novel combinations of components in a new population or setting.

## PROBLEMS WITH CONVENTIONAL PAIRWISE META-ANALYSIS

One important problem in meta-analyses of complex interventions is that the empirical evidence is often sparse. The number of possible versions of complex interventions grows exponentially with the number of components. Assuming we can meaningfully combine $m$ components without constraints, we can create $2^m - 1$ nonempty versions of a complex intervention. This number is 4,095 in our QI example—an order of magnitude more than the total number of available studies. Moreover, only a handful of the possible versions are observed in the empirical data. For the HbA1c outcome, only 83 unique interventions were observed (about 2% of the possible ones), and 59 of these were assessed in only 1 trial (Figure 1).

Comparing any version of a QI intervention with doing nothing in a pairwise meta-analysis may have some descriptive value but has little practical usefulness (8). Such a meta-analysis involves an ill-defined comparison: The experimental arm involves doing *something* (from a rather mixed bag of somethings) versus nothing—and the target population is not precisely defined. This question reduces a complicated problem to a simple one, at the cost of obtaining uninformative results: A difference between doing something and doing nothing tells us nothing about which version of the complex intervention a policy-maker should choose. Conversely, finding no difference on average does not imply that all versions of the complex intervention are ineffective.

| | Single Strategy | | | | | | | 2-Strategy Combinations | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| AF | | | | | | ■ | | | ■ | | | | | | | | ■ | ■ | | | |
| CM | | | | | | | | ■ | | | | ■ | ■ | | | | | | | | |
| TC | | | | | ■ | | | | | | | ■ | | ■ | ■ | | | | | | |
| EPR | | | | | | | | | | | | | | | | | ■ | | ■ | | |
| CE | | | ■ | | | | | | ■ | | | | | ■ | | ■ | | | | ■ | |
| CR | | | | | | | ■ | | | | | | | | | | | ■ | | ■ | |
| FR | | | | ■ | | | | | | ■ | | | | | | | | | | | |
| PE | ■ | | | | | | | ■ | | | ■ | | | | ■ | ■ | | | | | ■ |
| PSM | | ■ | | | | | | | | ■ | ■ | | ■ | | | | | | | | |
| PR | | | | | | | | | | | | | | | | | | | ■ | | ■ |
| CQI | | | | | | | | | | | | | | | | | | | | | |
| FI | | | | | | | | | | | | | | | | | | | | | |
| No. of Occurrences | 15 | 7 | 5 | 3 | 2 | 1 | 1 | 5 | 4 | 4 | 3 | 2 | 2 | 2 | 2 | 2 | 1 | 1 | 1 | 1 | 1 |

| | 3-Strategy Combinations | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| AF | | | | | | ■ | ■ | ■ | | | | | | | | | | | | |
| CM | ■ | | | ■ | | | | | ■ | ■ | ■ | ■ | ■ | ■ | | | | | | |
| TC | | ■ | | ■ | | | | | ■ | ■ | | | | | ■ | ■ | | | | |
| EPR | | | ■ | | | | | | | | ■ | | | | | | ■ | ■ | | |
| CE | | | | | | ■ | ■ | ■ | | | | ■ | | | | | ■ | | ■ | |
| CR | | | | | | ■ | | | | | | | | | | | ■ | | | ■ |
| FR | | | ■ | | ■ | | | | ■ | | | | ■ | | | | | ■ | | |
| PE | ■ | ■ | | ■ | ■ | | ■ | | | | | ■ | | ■ | ■ | ■ | | | ■ | ■ |
| PSM | ■ | ■ | ■ | | ■ | | | | | ■ | ■ | | ■ | | | | | | ■ | ■ |
| PR | | | | | | | | | | | | | | ■ | ■ | | | ■ | | |
| CQI | | | | | | | | ■ | | | | | | | | | | | | |
| FI | | | | | | | | | | | | | | | | ■ | | | | |
| No. of Occurrences | 8 | 7 | 4 | 3 | 3 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |

| | 4-Strategy Combinations | | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| AF | | | ■ | | ■ | ■ | ■ | ■ | ■ | ■ | | | | | | | | | | | | | | | | |
| CM | ■ | | | ■ | ■ | | | | | | ■ | ■ | ■ | ■ | ■ | ■ | ■ | | | | | | | | | |
| TC | ■ | | | ■ | | | | | | | ■ | | | | | | | ■ | ■ | ■ | ■ | ■ | | | | |
| EPR | | ■ | | | | ■ | ■ | | | | | ■ | ■ | ■ | | | | ■ | ■ | | | | ■ | ■ | | |
| CE | | | ■ | | | ■ | | ■ | ■ | | ■ | | | | ■ | | | | | ■ | | | ■ | | | |
| CR | | | ■ | | ■ | | ■ | | | ■ | | | | | | | | | | | ■ | | ■ | ■ | ■ | ■ |
| FR | | ■ | | ■ | | | | ■ | | | | ■ | | | | ■ | | ■ | | | | ■ | | ■ | ■ | ■ |
| PE | ■ | ■ | ■ | | | | | ■ | ■ | ■ | ■ | | ■ | ■ | ■ | ■ | ■ | | ■ | ■ | ■ | ■ | | | ■ | |
| PSM | ■ | ■ | | ■ | | | | | ■ | | | ■ | ■ | ■ | ■ | ■ | ■ | | ■ | | | ■ | | ■ | ■ | ■ |
| PR | | | | | ■ | ■ | ■ | | | ■ | | | | | | | ■ | ■ | | ■ | ■ | | ■ | | | ■ |
| CQI | | | | | | | | | | | | | | | | | | | | | | | | | | |
| FI | | | | | | | | | | | | | | | | | | | | | | | | | | |
| No. of Occurrences | 15 | 4 | 2 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |

| | 5-Strategy Combinations | | | | | | | | | | | 6 | | | 7 | 8 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| AF | ■ | ■ | | | | | | | | | | | ■ | | | |
| CM | | | ■ | ■ | ■ | ■ | ■ | ■ | ■ | | | ■ | | ■ | ■ | ■ |
| TC | | | ■ | ■ | ■ | | | | | | | ■ | ■ | ■ | ■ | ■ |
| EPR | ■ | | | | | ■ | | | | ■ | | | | ■ | ■ | ■ |
| CE | ■ | ■ | | | | | ■ | | | | | | ■ | ■ | | |
| CR | ■ | ■ | ■ | | | | | ■ | | | ■ | | ■ | | | ■ |
| FR | | | | ■ | | ■ | | ■ | ■ | ■ | ■ | ■ | | | ■ | ■ |
| PE | | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ |
| PSM | | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | | ■ | ■ |
| PR | ■ | | | | ■ | | | | ■ | ■ | ■ | ■ | | ■ | ■ | ■ |
| CQI | | | | | | | | | | | | | | | | |
| FI | | | | | | | ■ | | | | | | | | | |
| No. of Occurrences | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 4 | 1 | 1 | 1 | 1 |

**Figure 1.** Frequency of evaluations of quality improvement (QI) interventions comprised of component QI strategies. AF, audit and feedback; CE, clinician education; CM, case management; CR, clinician reminder; EPR, electronic patient registry; FR, facilitated relay; PE, patient education; PR, patient reminders; PSM, promotion of self-management; TC, team changes.

**Table 1.** Taxonomy of Quality Improvement Strategies Adapted From Cochrane's Effective Practice and Organization of Care 2002 Taxonomy That Were Used to Code Quality Improvement Interventions

| QI Strategy | 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 the 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 reminders | Any effort to remind patients about upcoming appointments or important aspects of 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 or patients or systemwide changes in reimbursement |

Abbreviation: QI, quality improvement.

An additional problem with commonly used meta-analytical approaches is that they do not use all available information. Some trials of complex interventions have 3 or more arms (3, 33), of which 2 are typically selected for a meta-analysis. For example, when comparing any active intervention with nothing, analysts often use only 2 arms from each study (i.e., complex interventions with the greatest number of components vs. the least number of components). This further complicates the interpretability of the findings, because the most intensive complex intervention in one study can be the least intensive intervention in another. Additional problems involve missing estimates of the sampling variance or missing estimates of the ICC in cluster-randomized trials that are needed to adjust unadjusted estimates of the sampling variance.

To obtain useful information about any version of the complex intervention, we must *extrapolate*, through statistical modeling, from the observed versions to the unobserved ones, using *all available information*. From the point of view of learning a response surface described above (i.e., that maps combinations of components and settings to average outcomes), using all available information amounts to *modeling outcomes of all arms in all studies*, as described below.

## SPECIFICATION OF THE BASIC RESPONSE SURFACE META-REGRESSION MODEL AND OF PRIOR DISTRIBUTIONS AND INFERENCE

### Specification of the response surface model

We model the associations of intervention components and study- or arm-level modifiers with the posttreatment outcome means using a random-effects regression with heteroskedastic errors (a hierarchical meta-regression model) (24, 25, 34, 35).

Let $Y_{ij}$ be the posttreatment mean in the $j$th arm of the $i$th study, distributed as

$$Y_{ij} \sim N\left(\mu_{ij}, \theta_{ij}^2\right), i = 1, \ldots, n; j = 1, \ldots n_i, \qquad (1)$$

where $\mu_{ij}$ is the arm-specific true mean and $\theta_{ij}^2$ is the conditional (sampling) variance. Given the large number of components and potential interactions between them, we will usually have to assume a parsimonious model for $\mu_{ij}$. We begin by considering a linear additive model in terms of the $m$ components:

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

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 $\beta_{ki}$ as "mean differences" because they express the difference in the posttreatment mean when component $X_{ijk}$ is present but make no causality claims. The intercept, $\beta_{0i}$, represents the posttreatment mean in the absence of intervention (36). $\beta_{0i}$ and $\beta_{ki}$ are treated as study-specific nuisance parameters in the estimation. We assume that they are realizations of underlying random variables, each with a normal distribution

$$\beta_{ki} \sim N\left(\beta_k, \tau_k^2\right), k = 1, \ldots, m \tag{3}$$

whose mean $\beta_k$ and variance $\tau_k^2$ will be estimated from the model. Study intercepts, $\beta_{0i}$, are also assumed to be random variables with a normal distribution with mean $\beta_0$ and variance $\tau_0^2$,

$$\beta_{0i} \sim N\left(\beta_0, \tau_0^2\right). \tag{4}$$

Equation 3 assumes exchangeable arm-specific parameters. We believe that this is justified in meta-analyses of trials of complex interventions where studies commonly evaluate active arms, the distinction between intervention and control arms is often unclear, and it is common for some intervention components to be assessed in each arm. In addition, for simplicity, equation 3 uses independent normal distributions. We believe that this is justified for our application because intervention components often operate at different levels (e.g., health system, provider, patient), and it is likely that their distribution would not involve strong correlation. If such correlations were in evidence in the data, they would be revealed in the estimated joint distribution of the β's. Alternatively, note that equation 3 is equivalent to an $m$-variate normal distribution with a diagonal covariance matrix, and, if desired, one could assume correlated β's with various structures for the covariance matrix (e.g., unstructured, compound-symmetrical, or other, depending on topic-specific information).

### Specification of prior distributions

Some features of our approach, such as incorporating external information and handling of missing data (see "Application to the Diabetes QI Example"), are most naturally achieved in the Bayesian framework (37, 38). Furthermore, Bayesian hierarchical modeling more fully reflects parameter uncertainty and is appealing for evidence synthesis because the study similarity judgments that systematic reviewers make are conceptually related to Bayesian exchangeability assumptions (39). For these reasons, we opted to use a Bayesian approach to estimate model parameters. Specifically, we used the minimally informative normal distributions for $\beta_0$ and $\beta_k$ and uniform distributions for $\tau_k$ and $\tau_0$ (see the Discussion section). Systematic reviewers can often rely on other meta-analyses or expert opinion to specify prior distributions to improve estimation with sparse data (37). Use of informative prior distributions is particularly useful for the heterogeneity parameters, $\tau_k^2$, which are often poorly estimated (40, 41).

### Inference

We can use the model in "Specification of the response surface model" to 1) estimate the posterior distribution of the mean difference for each intervention component and make inferences from that posterior distribution; 2) rank the components by the magnitude of the mean differences—that is, estimate the probability that the mean difference for a component has the greatest mean difference, the second greatest, and so on among the components included in the model (42); and 3) predict the posttreatment mean in future studies, possibly using combinations of components that have not been previously assessed in trials (43). For example, predictive inference can be obtained by examining the posterior predictive distribution for the posttreatment mean for any combination of components

$$\mu_{\text{new}} = \beta_{0,\text{new}} + \sum_{k=1}^{m} \beta_{k,\text{new}} X_{k,\text{new}} \tag{5}$$

and

$$\beta_{k,\text{new}} \sim N\left(\beta_k, \tau_{\beta k}^2\right), k = 0, \ldots, m, \tag{6}$$

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

$$Y_{\text{new}} \sim N\left(\mu_{\text{new}}, \theta_{\text{new}}^2\right), \tag{7}$$

where $\theta_{\text{new}}^2$ denotes the sampling variance of the new study, which depends on the planned sample size.

### Model extensions

Web Appendix 1 (available at https://doi.org/10.1093/aje/kwad184) extends the model to include pairwise product terms among components and between components and study-level covariates, handle discrete outcomes and missing data, and impute missing estimates of the ICC to adjust variance data in cluster-randomized trials.

## APPLICATION TO THE DIABETES QI EXAMPLE

We implemented a series of response surface models to describe the associations of each QI component with average outcomes and compared them to estimates from the conventional meta-analysis model. We used data from 114

**Table 2.** Summary of Results Comparing Analysis of Diabetes Quality Improvement Components Using Conventional and Response Surface Meta-Regression Models

| QI Strategy | Analysis and Model | | | | | |
|---|---|---|---|---|---|---|
| | Analysis I Conventional MA[a] | | Analysis II Meta-Regression[a,b] | | Analysis III Meta-Regression[b,c] | |
| | MD, % HbA1c | 95% CI | MD, % HbA1c | 95% CrI | MD, % HbA1c | 95% CrI |
| CM | −0.42 | −0.55, −0.29 | 0.03 | −0.13, 0.18 | 0.03 | −0.12, 0.17 |
| TC | −0.53 | −0.69, −0.37 | −0.37 | −0.55, −0.19 | −0.36 | −0.54, −0.18 |
| EPR | −0.37 | −0.53, −0.22 | −0.16 | −0.40, 0.08 | −0.15 | −0.38, 0.07 |
| CE | −0.23 | −0.37, −0.09 | −0.18 | −0.47, 0.09 | −0.17 | −0.44, 0.08 |
| FR | −0.40 | −0.54, −0.26 | −0.23 | −0.42, −0.03 | −0.24 | −0.43, −0.06 |
| PE | −0.44 | −0.56, −0.32 | −0.05 | −0.24, 0.15 | −0.10 | −0.28, 0.08 |
| PSM | −0.41 | −0.52, −0.30 | −0.21 | −0.41, −0.01 | −0.17 | −0.37, 0.01 |
| PR | −0.33 | −0.53, −0.14 | 0.03 | −0.21, 0.28 | −0.00 | −0.23, 0.22 |
| Other[d] | −0.19 | −0.31, −0.06 | −0.01 | −0.25, 0.21 | 0.02 | −0.20, 0.18 |

Abbreviations: CE, clinician education; CI, confidence interval; CM, case management; CrI, credible interval; EPR, electronic patient registry; FR, facilitated relay; HbA1c, hemoglobin A1c; MA, meta-analysis; MD, mean difference; PE, patient education; PR, patient reminders; PSM, promotion of self-management; QI, quality improvement; TC, team changes.

[a] Analyses used only the most intensive arms of multiple-arm trials versus the least intensive arms.

[b] The following priors were used in the Bayesian analyses to estimate parameters in analyses I and II: $\beta_0 \sim N(8, 100)$; $\tilde{\tau}_0 = U(0, 2)$; $\beta_k \sim N(0, 4)$; and $\tau_{\beta_k} = U(0, 2)$.

[c] Analyses used all arms from all trials.

[d] "Other" represents a combined category for infrequently evaluated components, including audit and feedback, clinician reminders, continuous quality improvement, and financial incentives.

trials (241 arms, 48,969 patients) that reported mean HbA1c levels at baseline and postintervention obtained from the 2012 version of the review (8), which is in the process of being updated (44). While updating the original review, we revised some data extraction algorithms, which resulted in small changes to the data (44). Because of these changes, our results are similar but not identical to those reported in 2012. We imputed missing information on standard errors of posttreatment arm means in individually randomized trials and missing intracluster correlation coefficients in cluster-randomized trials as per Web Appendix 2. To facilitate modeling, and with content-expert input, we combined QI components that were observed infrequently (present in less than 10% of arms) into an "other" category. This category included the QI components for clinician reminders (9.5%), audit and feedback (8.7% of arms), financial incentives (0.8%), and continuous QI (0.4%).

The conventional meta-analysis model was fitted in R (overall and subgroup analyses for each QI component) using the "meta" package in R (45); estimates are reported as mean changes and 95% confidence intervals. Hierarchical models were fitted using Markov chain Monte Carlo methods with the software JAGS (46) called from R, with 100,000 iterations for burn-in and 100,000 iterations to obtain the posterior distribution of parameters of interest. We assigned normal prior distributions for the coefficients $\beta_k \sim N(0, 4)$ and the baseline intercept $\beta_0 \sim N(8, 100)$ and uniform prior distributions for the between-studies standard deviations $\tau_0, \tau_k \sim U(0, 2)$. Estimates are reported as median changes and 95% credible intervals (CrIs), representing the mean difference associated with the presence of the component relative to the component's not being present. We used the Brooks-Gelman-Rubin diagnostic to assess parameter convergence (47, 48).

### Comparison of meta-regression and conventional synthesis models

We compared parameter estimates from 3 analyses:

- *Analysis I* imitated commonly used analyses in reviews of complex interventions in using only 2 arms from each trial, selecting the most and least intensive ones in multiarm trials (228 arms; 44,375 individuals). It comprised 1 random-effects meta-regression per QI component, where the sole predictor was the presence or absence of the QI component in the experimental arm. The model for the *k*th component did not adjust for the remaining components.
- *Analysis II* used the same 2 arms from each trial as analysis I in a single random-effects meta-regression that fitted the response surface model in equations 1–4.
- *Analysis III* used all arms in all trials in a meta-regression according to the model in equations 1–4 (241 arms; 48,969 individuals).

In analysis I, we imputed missing standard deviations and ICCs with fixed values (2.22 for standard deviations, the 99th percentile in observed data; and 0.027 for ICCs, obtained from a single study), as was done in the original review (49). We used ICCs to correct standard errors from cluster-randomized trials in which results were not appropriately adjusted for the clustering effect (49). Sensitivity analyses using a less conservative standard deviation (median, 1.34) and the higher ICC used in the 2012 review (8) (ICC = 0.07) did not change the overall mean and precision of the random-effects meta-analysis appreciably. Missing data in analyses II and III were imputed in the hierarchical meta-regression model as described in Web Appendices 1 and 2. Web Appendix 3 (including Web Tables 1 and 2) describes the different missing data patterns observed in the data set. Software code for all analyses is presented in Web Appendices 4 and 5 and on the authors' GitHub page (https://github.com/kkonnyu/evsynthmetaregression).

Table 2 summarizes the results. Compared with analysis I, the estimated coefficients for each component were smaller in analysis II and had a wider variation in magnitude. If the models in analyses II and III are approximately correctly specified, the smaller (and more varied) magnitude of the estimates from analyses II and III as compared with the estimates from analysis I may indicate better isolation of the expected posttreatment mean reduction associated with unique QI components, *accounting for co-occurring components*. In other words, analysis I, and thus conventional meta-analysis, would overestimate the associations of a single QI component with differences in outcomes. Point estimates of parameters from analysis II, which used only the most and least intensive arms from each trial, were similar to those of analysis III, which used all arms in all trials. However, estimates were more precise in analysis III.

## Ranking of coefficients for intervention components

An example rankogram that can be produced from estimates of a meta-regression model is presented in Figure 2. Using the output of analysis III above, the rankogram indicates the probability of each component's being the best, second best, and so on among the modeled components with respect to postintervention mean difference. For example, based on this example, the "team changes" component had a higher probability of ranking as one of the top 3 QI components, while "case management" appeared to rank in the bottom 3.

## Assessing nonadditivity

Web Appendix 1 (equation 8) extends the model in equations 1 and 2 to include pairwise product terms (i.e., allowing for nonadditive associations) among QI components. A series of models were fitted that each included product terms for a single QI strategy, $r$, with all remaining QI strategies, $l \in \{1, \ldots, m\} \backslash r$, corresponding to 8 additional parameters estimated in each of the 9 additional models (Web Appendix 6). We present results from these models in Figure 3. The CrIs of the coefficients of the product terms all crossed 0, consistent with the absence of nonadditive associations between the QI strategies. We therefore present results only from the more parsimonious models in the main paper.

## Assessing modification of associations by covariates

The original QI review identified the study average of HbA1c levels at baseline as a potential effect modifier. Thus, we explored the addition of product terms between baseline HbA1c (coded as both a binary and a continuous covariate) and each of the QI components as per equation 10 (Web Appendix 1). 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 parameters (i.e., the coefficient of baseline HbA1c plus 9 coefficients for product terms between baseline HbA1c and QI strategies; see Web Appendices 7 and 8). Although estimates were not grossly incompatible with the absence of modification by HbA1c, there was some evidence that baseline HbA1c level modified the association between some QI strategies and the posttreatment mean outcome (Figure 4) and improved model predictions, particularly when treated as a continuous variable (see "Assessment of convergence, model evaluation, and robustness" section below). For example, the association between case management and the posttreatment mean outcome appeared to be greater when delivered in populations with higher baseline risk (Table 3). 12 However, the coefficients of the product terms were imprecisely estimated. In the end, because our results did not suggest systematic differences in posttreatment mean outcomes for different QI components over baseline HbA1c control, we continued to prefer our base model.

## Predicting outcomes for a novel combination of QI components

Assumingthat our response surface models are approximately correctly specified, we predict the distribution of the posttreatment mean for a specific combination of QI components in a new setting or trial (43). Table 4 presents the posterior predictive distribution of the outcome mean for novel combinations of components that were not observed in the included trials. For example, the QI components of "team changes," "facilitated relay," and "electronic patient registry" had relatively strong negative associations with the posttreatment outcome mean in analysis III and would be reasonable to combine in a novel intervention according to our content experts. The posterior and predictive distributions of the posttreatment mean (% HbA1c) in a new study with no QI components were estimated to be 8.14 (95% CrI: 7.96, 8.32) and 8.14 (95% CrI: 6.46, 9.81), respectively. We estimated that the new complex intervention would be associated with a substantially lower posttreatment mean. Using the posterior distribution, the posttreatment mean was 7.38% (95% CrI: 7.04, 7.72), and the mean difference as compared with no intervention was −0.75 (95% CrI: –1.05, −0.45). Using the posterior predictive distribution, the posttreatment mean was 7.38% (95% CrI: 5.56, 9.20), and the mean difference compared with no intervention was −0.75 (95% CrI: –1.51, −0.03).

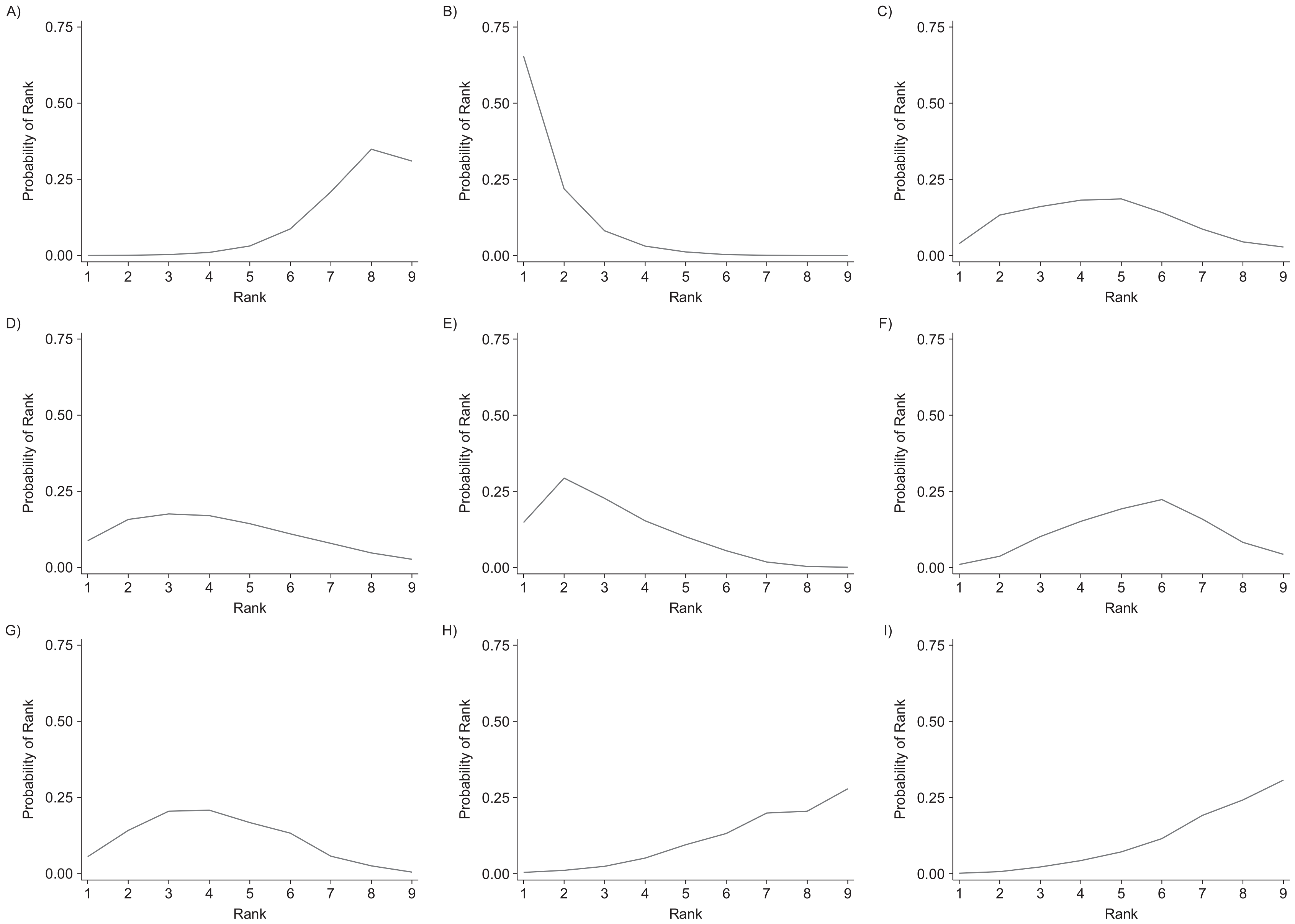

**Figure 2.** Ranking of quality improvement strategy components. A) Case management; B) team changes; C) electronic patient registry; D) clinician education; E) facilitated relay; F) patient education; G) promotion of self-management; H) patient reminders; I) other. “Other” represents a combined category for infrequently evaluated components, including audit and feedback, clinician reminders, continuous quality improvement, and financial incentives.

**Table 3.** Average Outcomes of Quality Improvement Strategies in Study Arms With a Controlled Hemoglobin A1c Level at Baseline Versus an Uncontrolled Level at Baseline

| QI Strategy | Uncontrolled HbA1c[a,b] + QI Strategy $(\beta_0 + \beta_k + \phi + \psi_k)$ | | Difference in Uncontrolled HbA1c[a] $(\beta_k + \psi_k)$ | | Controlled HbA1c + QI Strategy[a, c] $(\beta_0 + \beta_k)$ | | Difference in Controlled HbA1c[a] $(\beta_k)$ | | Difference Between Differences[a] $[\beta_k - (\beta_k + \psi_k)]$ | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Median, % HbA1c | 95% CrI | Median, % HbA1c | 95% CrI | Median, % HbA1c | 95% CrI | Median, % HbA1c | 95% CrI | Median, % HbA1c | 95% CrI |
| CM | 8.45 | 8.08, 8.82 | −0.18 | −0.48, 0.11 | 7.52 | 7.26, 7.80 | 0.12 | −0.07, 0.33 | 0.30 | −0.05, 0.67 |
| TC | 8.35 | 8.00, 8.72 | −0.29 | −0.56, 0.01 | 7.13 | 6.79, 7.46 | −0.27 | −0.56, 0.02 | 0.02 | −0.41, 0.41 |
| EPR | 8.65 | 7.90, 9.34 | 0.01 | −0.70, 0.66 | 7.32 | 6.98, 7.69 | −0.08 | −0.39, 0.26 | −0.09 | −0.78, 0.69 |
| CE | 8.53 | 7.84, 9.25 | −0.11 | −0.78, 0.59 | 7.30 | 7.00, 7.61 | −0.10 | −0.38, 0.20 | 0.01 | −0.75, 0.75 |
| FR | 8.42 | 8.07, 8.79 | −0.22 | −0.51, 0.09 | 7.11 | 6.74, 7.50 | −0.29 | −0.64, 0.07 | −0.08 | −0.51, 0.40 |
| PE | 8.55 | 8.21, 8.89 | −0.08 | −0.40, 0.23 | 7.49 | 7.21, 7.76 | 0.09 | −0.19, 0.36 | 0.18 | −0.23, 0.58 |
| PSM | 8.45 | 8.09, 8.81 | −0.19 | −0.49, 0.11 | 7.13 | 6.78, 7.45 | −0.27 | −0.59, 0.01 | −0.08 | −0.51, 0.32 |
| PR | 8.75 | 8.12, 9.37 | 0.12 | −0.48, 0.69 | 7.20 | 6.76, 7.61 | −0.19 | −0.62, 0.16 | −0.31 | −1.03, 0.36 |
| Other[d] | 8.48 | 7.66, 9.18 | −0.15 | −0.93, 0.49 | 7.42 | 7.11, 7.68 | 0.02 | −0.25, 0.22 | 0.17 | −0.52, 0.99 |

Abbreviations: CE, clinician education; CM, case management; CrI, credible interval; EPR, electronic patient registry; FR, facilitated relay; HbA1c, hemoglobin A1c; PE, patient education; PR, patient reminders; PSM, promotion of self-management; QI, quality improvement; TC, team changes.

[a] The following priors were used in the Bayesian analyses to estimate parameters in the models allowing for modification of the association by HbA1c: $\beta_0 \sim N(8, 100)$; $\beta_k \sim N(0, 4)$; $\phi \sim N(0, 4)$; $\psi_k \sim N(0, 4)$; and all $\tau \sim U(0, 2)$.

[b] Postintervention mean change in patients with an uncontrolled HbA1c level who did not receive the QI strategy: $\beta_0 + \phi$ = 8.63 (95% CrI: 8.42, 8.86).

[c] Postintervention mean change in patients with a controlled baseline HbA1c level who did not receive the QI strategy: $\beta_0$ = 7.40 (95% CrI: 7.24, 7.56).

[d] “Other” represents a combined category for infrequently evaluated components, including audit and feedback, clinician reminders, continuous quality improvement, and financial incentives.

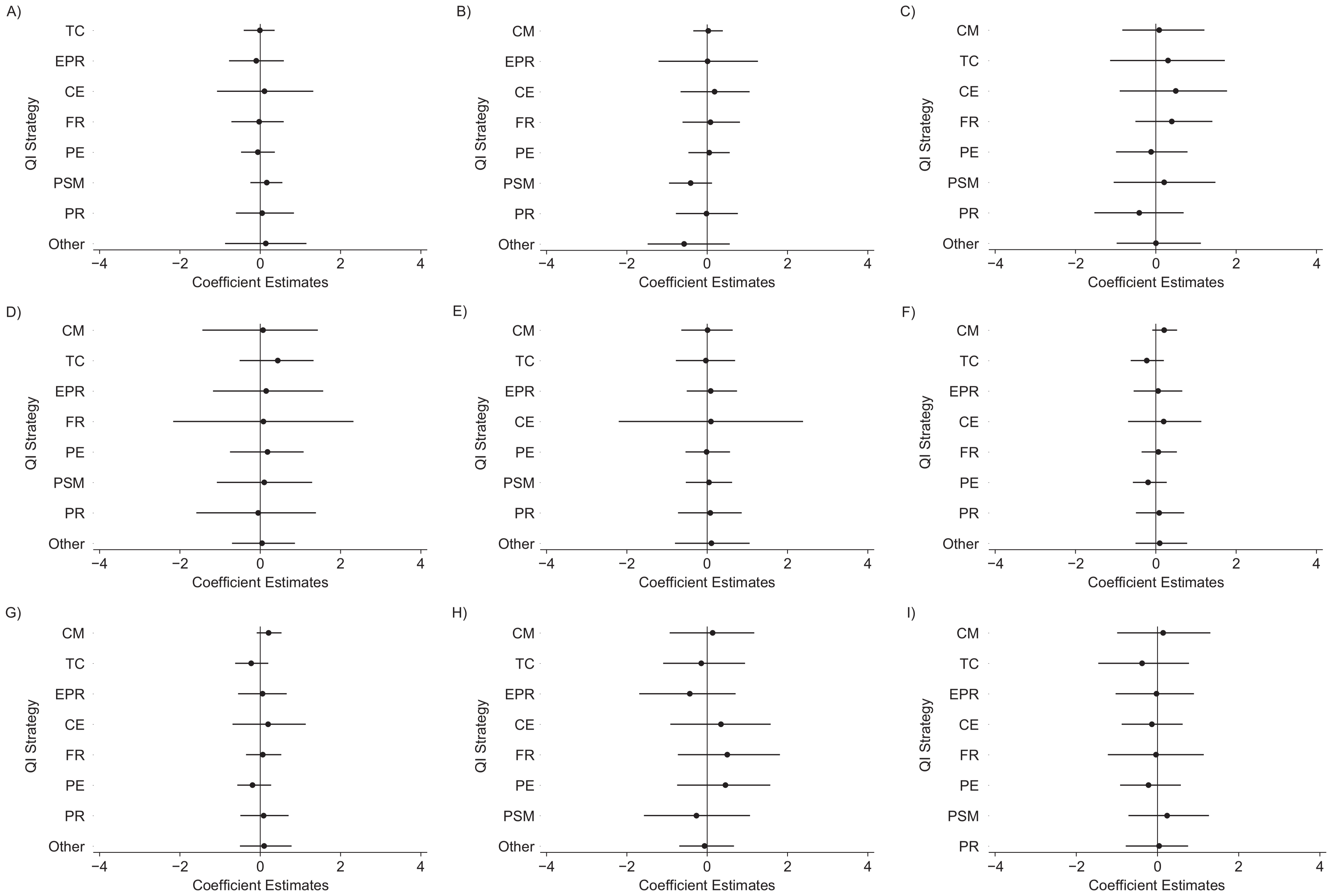

**Figure 3.** Coefficient estimates for pairwise product terms between quality improvement (QI) strategy components. The figure shows the results for pairwise product terms with A) case management (CM) (*n* = 65); B) team changes (TC) (*n* = 56); C) electronic patient registry (EPR) (*n* = 28); D) clinician education (CE) (*n* = 34); E) facilitated relay (FR) (*n* = 43); F) patient education (PE) (*n* = 109); G) promotion of self-management (PSM) (*n* = 93); H) patient reminders (PR) (*n* = 27); and I) “other” (*n* = 36). “Other” represents a combined category for infrequently evaluated components, including audit and feedback, clinician reminders, continuous quality improvement, and financial incentives. The following priors were used in the Bayesian analyses to estimate parameters in the models that included product terms: $\beta_0 \sim N(8, 100)$; $\beta_k \sim N(0, 4)$; (the prior distributions for the mean of the product term $\gamma$) $\gamma_l \sim N(0, 2)$; and all $\tau \sim U(0, 2)$. For each model, the reported *n* indicates the number of observations (arms) for the component of interest (i.e., 65 arms in which CM was present).

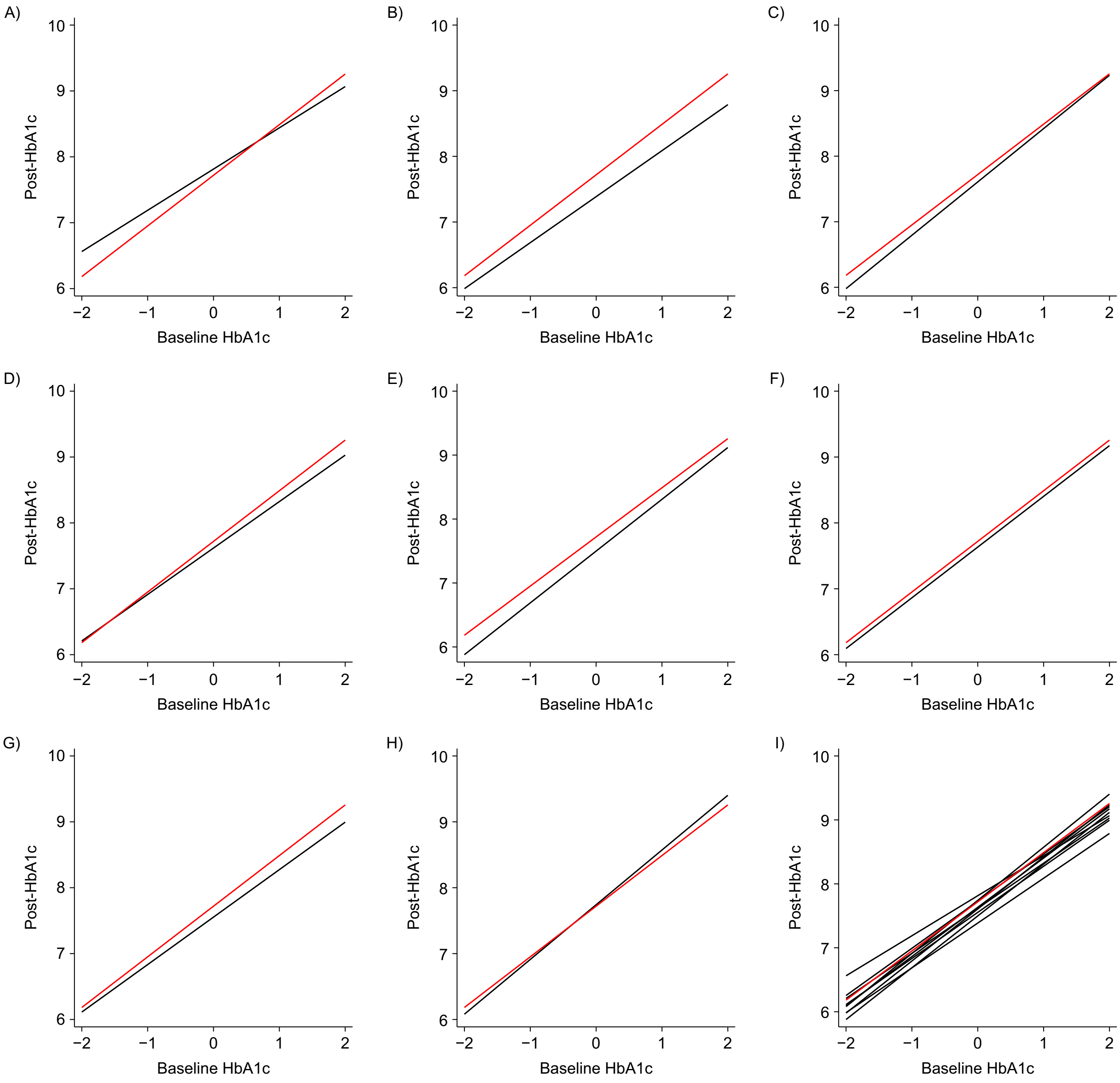

**Figure 4.** Modification of the association between quality improvement (QI) components and the posttreatment outcome mean by baseline hemoglobin A1c (HbA1c) level. A) Case management; B) team changes; C) electronic patient registry; D) clinician education; E) facilitated relay; F) patient education; G) promotion of self-management; H) patient reminders; I) all components combined. Baseline HbA1c values were centered. The red line shows the estimated posttreatment outcome mean without the QI strategy. The black line shows the estimated posttreatment outcome mean with the QI strategy. The following priors were used in the Bayesian analyses to estimate parameters in the association modification models: $\beta_0 \sim N(8, 100)$; $\beta_k \sim N(0, 4)$; $\phi \sim N(0, 4)$; $\psi_k \sim N(0, 4)$; and all $\tau = U(0, 2)$.

## Assessment of convergence, model evaluation, and robustness

The upper limit of the 95% CrI for the Brooks-Gelman-Rubin statistic was less than 1.1 for 96% of all parameters monitored (and less than 1.2 for all parameters monitored).

We used mixed posterior predictive checks (50) to compare the predicted outcome means from the models of new studies with the same combination of components as those observed in our sample against the observed outcome means. As illustrated in Figure 5, the meta-regression model performs reasonably well in terms of its probability of returning the observed means from parameter estimates. Other standard methods, such as the deviance information criterion, can also be used to evaluate model performance (50, 51).

**Table 4.** Estimated Median Values of the Posterior and Posterior Predictive Distributions of Average Outcomes for Untested Combinations of Quality Improvement Strategies

| Untested Combination | Posterior Distribution of the Posttreatment Mean | | Difference From Baseline | | Posterior Predictive Distribution of the Posttreatment Mean | | Difference From Baseline | |
|---|---|---|---|---|---|---|---|---|
| | Median, % HbA1c | 95% CrI | Median, % HbA1c | 95% CrI | Median, % HbA1c | 95% CrI | Median, % HbA1c | 95% CrI |
| CM + EPR + FR | 7.77 | 7.45, 8.09 | −0.36 | −0.64, −0.09 | 7.77 | 6.01, 9.53 | −0.36 | −0.94, 0.19 |
| TC + FR + PSM | 7.36 | 7.06, 7.66 | −0.77 | −1.03, −0.52 | 7.36 | 5.54, 9.18 | −0.78 | −1.50, −0.04 |
| TC + FR + EPR | 7.38 | 7.04, 7.72 | −0.75 | −1.05, −0.45 | 7.38 | 5.56, 9.20 | −0.75 | −1.51, −0.03 |

Abbreviations: CM, case management; CrI, credible interval; EPR, electronic patient registry; FR, facilitated relay; HbA1c, glycated hemoglobin; PSM, promotion of self-management; QI, quality improvement; TC, team changes.

In general, the assessment approach should be chosen to reflect the goals of the modeling. We favor posterior predictive checks in our application because our primary goal is prediction. Because Bayesian posterior probabilities can be influenced by the specified priors (52), we performed sensitivity tests on our chosen priors and found that our findings were robust to these alterations.

Finally, we compared our analyses with meta-regressions that model differences in outcomes between study arms (53) rather than outcomes for each arm, and found similar estimates of differences in posttreatment mean outcomes for different QI components. Data for these analyses were sparser because the coefficients of components that are common in 2 or more arms of the same study "cancel" out, and thus the analyses took longer to converge (Web Table 3).

## DISCUSSION

We believe that *predicting* the average outcomes of novel complex interventions in new settings or future studies is a key goal when synthesizing evidence from trials of complex interventions. While any prediction model is unlikely to be correctly specified, examining a collection of models may offer useful insights for decision-making or planning future research: Patterns across evidence may be unearthed, and the diversity of the large data sets can be used as a strength rather than a limitation. Our experience is that decision-makers find these insights more useful than those afforded by conventional meta-analysis methods. The latter ask very abstract questions (e.g., doing something vs. nothing; including a component vs. not including it), do not adjust for combinations of components or study-level characteristics, and do not address data complications (e.g., missing data, differences in study designs).

Our approach includes all data from all studies and estimates the average posttreatment mean associated with each component. Our response surface models extrapolate to unobserved combinations of components and different settings by employing simplifying additivity assumptions. The assumptions are explicit and can be debated by substantive experts, examined statistically, and relaxed by using more flexible models if the data permit. Natural model extensions can handle missing data or other complications that arise in evidence synthesis of trials of complex interventions. We focused on models that use arm-level data, though models based on differences are also possible and led to similar results in our example.

We have no hope of observing empirical data on all 4,095 nonempty combinations of 12 components; we must extrapolate from the few observed combinations, about 2% of the total, to the remaining 98% of unobserved ones through a model which we argue lends itself to a response surface approach. The models in equations 1–4 enable these extrapolations while reducing the parameter space from order $2^m$, in a fully saturated nonparametric model, to an order $m$, in a model of main effects only. This 3-orders-of-magnitude reduction in the number of parameters assumes that the information modeled by the omitted parameters is negligible. However, making such assumptions and testing some of them in alternative models (e.g., models that add pairwise interactions between components) allows us to acquire *working predictions* that can be used by policy-makers and researchers who plan future research. Working predictions are useful: Policy-makers must make decisions even without data—and in many cases, their initial decisions are tentative, dynamically monitored, and subject to revision. Researchers who plan novel interventions will eventually put them to the test in randomized experiments or through observation. This viewpoint builds on a conceptual proposal that meta-analysis is best viewed as response surface estimation (25), and is most similar to the application of evidence synthesis in social science (24).

All predictions are, of course, conditional on the model, and the "true" model is ultimately unknown. Substantial heterogeneity in results will often remain in applications of our approach, because the representation of a complex intervention as a vector of components does not capture everything about an intervention (54–56); the measured study level attributes do not capture everything about populations and settings; and important covariates are poorly measured or missing. These difficulties limit the potential usefulness of predictive modeling based on aggregate data (57, 58) and are perhaps even more limiting for analyses aspiring to produce results that have causal interpretations.

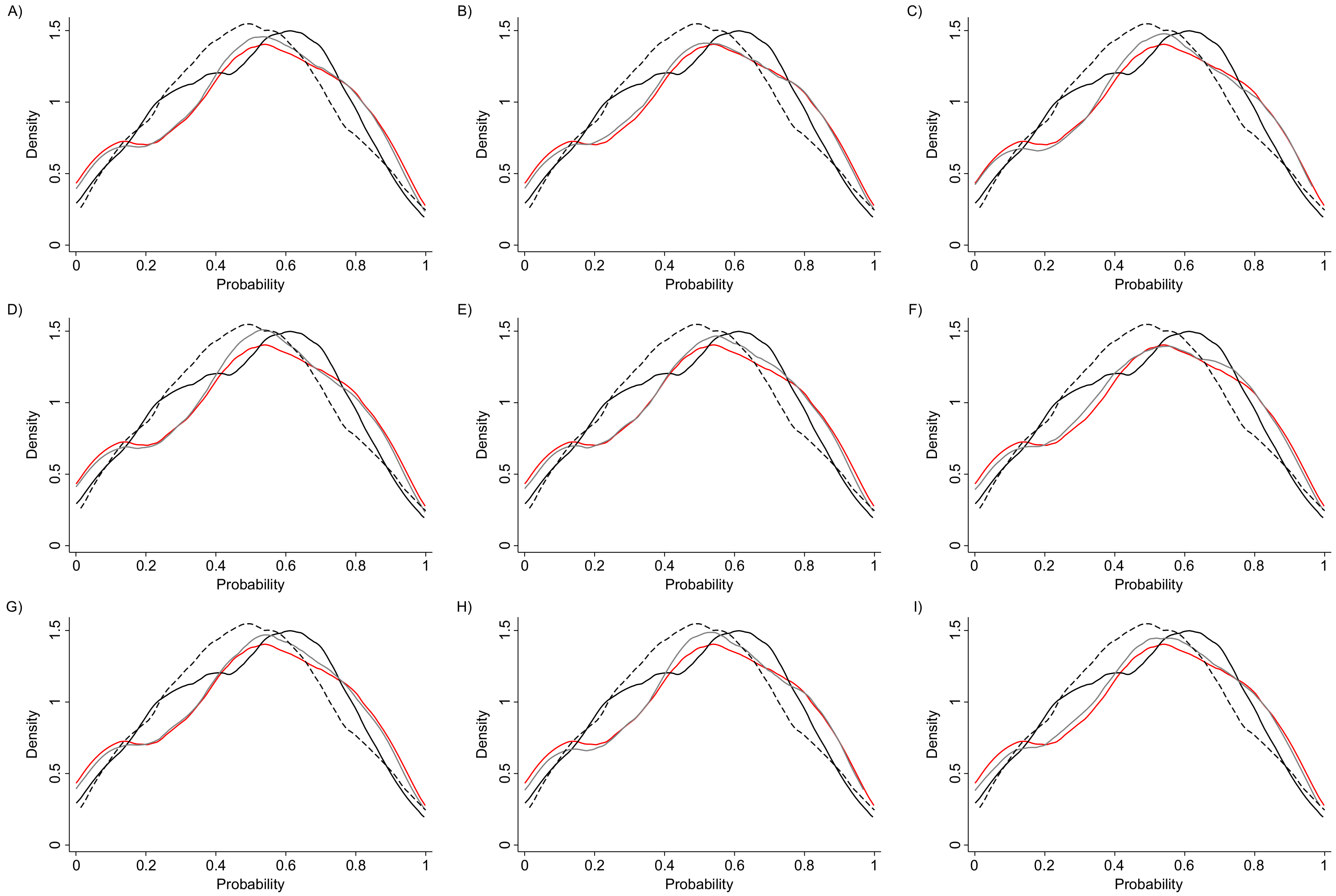

**Figure 5.** Posterior predictive checks. A) Quality improvement (QI) × case management; B) QI × team changes; C) QI × electronic patient registry; D) QI × clinician education; E) QI × facilitated relay; F) QI × patient education; G) QI × promotion of self-management; H) QI × patient reminders; I) QI × other. “Other” represents a combined category for infrequently evaluated components, including audit and feedback, clinician reminders, continuous quality improvement, and financial incentives. Each panel represents the posterior predictive check of the main model (red line), the model including product terms between QI components (9 models including a product term between one QI strategy and the other 8 QI strategies; gray line), the effect modification model when baseline hemoglobin A1c (HbA1c) level is treated as binary (black solid line), and the association modification model when baseline Hba1c is treated as continuous (black dashed line). These are densities of Bayesian 1-sided probabilities comparing observed and predicted mean outcomes. When the models fit well, one expects to see a mode at 0.5 (agreement of observed and predicted) with small mass at either extreme.

In conclusion, predicting the average outcomes of complex interventions in a new setting or study is a key goal for evidence synthesis of trials comparing complex interventions. Collections of meta-regression models can be used to estimate the response surface relating study outcomes to intervention components and study characteristics, to isolate component-specific associations with outcomes, and to predict the outcomes of complex interventions (including those not previously evaluated) in new populations or settings, while addressing common data complications.

## ACKNOWLEDGMENTS

Author affiliations: Center for Evidence Synthesis in Health, Brown University, Providence, Rhode Island, United States (Kristin J. Konnyu, Thomas A. Trikalinos); Department of Health Services, Policy & Practice, School of Public Health, Brown University, Providence, Rhode Island, United States (Kristin J. Konnyu, Thomas A. Trikalinos); School of Epidemiology and Public Health, University of Ottawa, Ottawa, Ontario, Canada (Jeremy M. Grimshaw); Department of Medicine, Faculty of Medicine, University of Ottawa, Ottawa, Ontario, Canada (Jeremy M. Grimshaw); Clinical Epidemiology Program, Ottawa Hospital Research Institute, Ottawa, Ontario, Canada (Jeremy M. Grimshaw, David Moher); Family Practice Health Centre, Women's College Research Institute, and Institute for Health Systems Solutions and Virtual Care, Women's College Hospital, Toronto, Ontario, Canada (Noah M. Ivers); Department of Family and Community Medicine, Temerty Faculty of Medicine, University of Toronto, Toronto, Ontario, Canada (Noah M. Ivers); Institute of Health Policy Management and Evaluation, University of Toronto, Toronto, Ontario, Canada (Noah M. Ivers); CAUSALab, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, United States (Issa. J. Dahabreh); Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, United States (Issa J. Dahabreh); and Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, United States (Issa J. Dahabreh). Dr. Kristin J. Konnyu is currently affiliated with the Health Services Research Unit and the Institute of Applied Health Sciences at the School of Medicine, Medical Sciences and Nutrition, University of Aberdeen (Aberdeen, Scotland, United Kingdom).

This study was funded by the Canadian Institutes of Health Research (grants FDN-143269 and FRN-123345) and a research fellowship held by K.J.K. (Frederick Banting and Charles Best Canada Graduate Scholarship GSD-134936). N.M.I. holds a Canada Research Chair (Tier 2) in Implementation of Evidence Based Practice and a Clinician Scientist Award from the Department of Family and Community Medicine at the University of Toronto (Toronto, Ontario, Canada). J.M.G. held a Canada Research Chair in Health Knowledge Transfer and Uptake during the time of the study's conduct and was supported by a Foundation Grant from the Canadian Institutes of Health Research. D.M. was supported by a University of Ottawa Research Chair during the time of study conduct.

Software code and data are available in the authors' GitHub repository (https://github.com/kkonnyu/evsynthmetaregression).

We acknowledge Dr. Samir Nazarali (University of Alberta, Edmonton, Alberta, Canada), who assisted in data-checking and reextraction of the data used in this paper.

This work was presented at the following conferences: 10th Annual Conference on the Science of Dissemination and Implementation in Health, Arlington, Virginia, United States, December 4–6, 2017; Global Evidence Summit, Cape Town, South Africa, September 13–16, 2017; Cochrane Canada Symposium, Hamilton, Ontario, Canada, May 11 and 12, 2017; 11th Annual Meeting of the Society for Research Synthesis Methodology, Florence, Italy, July 11–13, 2016; and 21st Cochrane Colloquium, Quebec City, Quebec, Canada, September 19–23, 2013.

The Canadian Institutes of Health Research and other funders played no role in the study design, plans for data collection and analysis, the decision to publish, or preparation of the manuscript.

Conflict of interest: none declared.

## REFERENCES

1. Craig P, Dieppe P, Macintyre S, et al. Developing and evaluating complex interventions: the new Medical Research Council guidance. *BMJ.* 2008;337:a1655.
2. Craig P, Petticrew M. Developing and evaluating complex interventions: reflections on the 2008 MRC guidance. *Int J Nurs Stud.* 2013;50(5):585–587.
3. Anderson LM, Petticrew M, Chandler J, et al. Introducing a series of methodological articles on considering complexity in systematic reviews of interventions. *J Clin Epidemiol.* 2013;66(11):1205–1208.
4. Guise JM, Chang C, Butler M, et al. 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.
5. Wong G. Is complexity just too complex? *J Clin Epidemiol.* 2013;66(11):1199–1201.
6. Anderson LM, Oliver SR, Michie S, et al. Investigating complexity in systematic reviews of interventions by using a spectrum of methods. *J Clin Epidemiol.* 2013;66(11): 1223–1229.
7. Kühne F, Ehmcke R, Härter M, et al. Conceptual decomposition of complex health care interventions for evidence synthesis: a literature review. *J Eval Clin Pract.* 2015;21(5):817–823.
8. Tricco AC, Ivers NM, Grimshaw JM, et al. Effectiveness of quality improvement strategies on the management of diabetes: a systematic review and meta-analysis. *Lancet.* 2012;379(9833):2252–2261.
9. Trikalinos TA, Wieland LS, Adam GP, et al. *Decision Aids for Cancer Screening and Treatment*. (Comparative Effectiveness Reviews, no. 145. (Report no. 15-EHC002-EF)Rockville, MD: Agency for Healthcare Research and Quality, US Department of Health and Human Services; 2014.

10. Silberholz J, Bertsimas D, Vahdat L. Clinical benefit, toxicity and cost of metastatic breast cancer therapies: systematic review and meta-analysis. *Breast Cancer Res Treat.* 2019; 176(3):535–543.
11. Michie S, Richardson M, Johnston M, 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(1):81–95.
12. Michie S, van Stralen MM, West R. The behaviour change wheel: a new method for characterising and designing behaviour change interventions. *Implement Sci.* 2011;6:42.
13. Dahabreh IJ, Petito LC, Robertson SE, et al. Toward causally interpretable meta-analysis: transporting inferences from multiple randomized trials to a new target population. *Epidemiology.* 2020;31(3):334–344.
14. Dahabreh IJ, Robertson SE, Petito LC, et al. Efficient and robust methods for causally interpretable meta-analysis: transporting inferences from multiple randomized trials to a target population. *Biometrics.* 2022;79(2):1057–1072.
15. Rücker G, Petropoulou M, Schwarzer G. Network meta-analysis of multicomponent interventions. *Biom J.* 2020;62(3):808–821.
16. Welton NJ, Caldwell DM, Adamopoulos E, et al. Mixed treatment comparison meta-analysis of complex interventions: psychological interventions in coronary heart disease. *Am J Epidemiol.* 2009;169(9):1158–1165.
17. Caldwell DM, Davies SR, Hetrick SE, et al. School-based interventions to prevent anxiety and depression in children and young people: a systematic review and network meta-analysis. *Lancet Psychiatry.* 2019;6(12):1011–1020.
18. Freeman SC, Scott NW, Powell R, et al. Component network meta-analysis identifies the most effective components of psychological preparation for adults undergoing surgery under general anesthesia. *J Clin Epidemiol.* 2018;98: 105–116.
19. López-López JA, Davies SR, Caldwell DM, et al. The process and delivery of CBT for depression in adults: a systematic review and network meta-analysis. *Psychol Med.* 2019; 49(12):1937–1947.
20. Mills EJ, Druyts E, Ghement I, et al. Pharmacotherapies for chronic obstructive pulmonary disease: a multiple treatment comparison meta-analysis. *Clin Epidemiol.* 2011;3:107–129.
21. Mills EJ, Thorlund K, Ioannidis JP. Calculating additive treatment effects from multiple randomized trials provides useful estimates of combination therapies. *J Clin Epidemiol.* 2012;65(12):1282–1288.
22. Pompoli A, Furukawa TA, Efthimiou O, et al. Dismantling cognitive-behaviour therapy for panic disorder: a systematic review and component network meta-analysis. *Psychol Med.* 2018;48(12):1945–1953.
23. Rücker G, Schmitz S, Schwarzer G. Component network meta-analysis compared to a matching method in a disconnected network: a case study. *Biom J.* 2021;63(2): 447–461.
24. Gelman A, Stevens M, Chan V. Regression modeling and meta-analysis for decision making. *J Bus Econ Stat.* 2003; 21(2):213–225.
25. Rubin DB. Meta-analysis: literature synthesis or effect-size surface estimation? *J Educ Stat.* 1992;17(4):363–374.
26. International Diabetes Federation. *Diabetes Atlas*. Brussels, Belgium: International Diabetes Federation; 2015. http://www.idf.org/diabetesatlas. Accessed November 12, 2017.
27. Diabetes Canada. *Clinical Practice Guidelines—Full Guidelines*. Toronto, ON, Canada: Diabetes Canada; 2018. http://guidelines.diabetes.ca/fullguidelines. Accessed August 22, 2017.
28. American Diabetes Association. American diabetes association standards of medical care in diabetes—2017. *Diabetes Care.* 2017;40(suppl 1):S1–S135.
29. Hawthorne G, Hrisos S, Stamp E, et al. Diabetes care provision in UK primary care practices. *PloS One.* 2012; 7(7):e41562.
30. Presseau J, Mackintosh J, Hawthorne G, 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(1):65.
31. Shojania KG, Ranji SR, Shaw LK, et al. *Closing the Quality Gap: A Critical Analysis of Quality Improvement Strategies (Vol. 2: Diabetes Care)*. (AHRQ Technical Reviews, no. 9.2. (Report no. 04-0051-2)Rockville, MD: Agency for Healthcare Research and Quality, US Department of Health and Human Services; 2004.
32. Cochrane Effective Practice and Organisation of Care (EPOC) Review Group. EPOC Taxonomy. http://epoc.cochrane.org/epoc-taxonomy. Published 2002. Accessed November 25, 2014.
33. Pigott T, Noyes J, Umscheid CA, et al. AHRQ series on complex intervention systematic reviews—paper 5: advanced analytic methods. *J Clin Epidemiol.* 2017;90: 37–42.
34. Gelman A, Hill J. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. 1st ed (Analytical Methods for Social Research. New York, NY: Cambridge University Press; 2009.
35. Senn S. Hans van Houwelingen and the art of summing up. *Biom J.* 2010;52(1):85–94.
36. Viechtbauer W. Conducting meta-analyses in R with the metafor package. *J Stat Softw.* 2010;36(3):1–48.
37. Spiegelhalter DJ, Abrams KR, Myles JP. *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. 1st ed. Chichester, United Kingdom: John Wiley & Sons Ltd.; 2004.
38. Higgins JP, Thompson SG, Spiegelhalter DJ. A re-evaluation of random-effects meta-analysis. *J R Stat Soc Ser A Stat Soc.* 2009;172(1):137–159.
39. Gelman A, Carlin JB, Stern HS, et al. *Bayesian Data Analysis*. 1st ed (Chapman & Hall/CRC Texts in Statistical Science. Boca Raton, FL: Chapman & Hall/CRC Press; 1995.
40. Pullenayegum EM. An informed reference prior for between-study heterogeneity in meta-analyses of binary outcomes. *Stat Med.* 2011;30(26):3082–3094.
41. Rhodes KM, Turner RM, Higgins JP. Predictive distributions were developed for the extent of heterogeneity in meta-analyses of continuous outcome data. *J Clin Epidemiol.* 2015;68(1):52–60.
42. Cipriani A, Furukawa TA, Salanti G, et al. Comparative efficacy and acceptability of 12 new-generation antidepressants: a multiple-treatments meta-analysis. *Lancet.* 2009;373(9665):746–758.
43. Nikolakopoulou A, Mavridis D, Salanti G. Planning future studies based on the precision of network meta-analysis results. *Stat Med.* 2016;35(7):978–1000.
44. Ivers N, Tricco AC, Trikalinos TA, 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.

45. Schwarzer G. Meta: an R package for meta-analysis. *R News.* 2007;7(3):40–45.
46. Plummer M. JAGS: A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling. https://www.r-project.org/conferences/DSC-2003/Proceedings/Plummer.pdf. Published 2003. Accessed December 16, 2022.
47. Brooks SP, Gelman A. General methods for monitoring convergence of iterative simulations. *J Comput Graph Stat.* 1998;7(4):434–455.
48. Gelman A, Rubin DB. Inference from iterative simulation using multiple sequences. *Stat Sci.* 1992;7(4):457–472.
49. Higgins JPT, Deeks JJ, Altman DG. Chapter 16: special topics in statistics. In: Higgins JPT, Green S, eds. *Cochrane Handbook for Systematic Reviews of Interventions*. Chichester, United Kingdom: The Cochrane Collaboration; 2011.
50. Spiegelhalter DJ, Best NG, Carlin BP, et al. Bayesian measures of model complexity and fit. *J R Stat Soc Series B Stat Methodology.* 2002;64(4):583–639.
51. Marshall EC, Spiegelhalter DJ. Approximate cross-validatory predictive checks in disease mapping models. *Stat Med.* 2003;22(10):1649–1660.
52. Sutton AJ, Abrams KR. Bayesian methods in meta-analysis and evidence synthesis. *Stat Methods Med Res.* 2001;10(4): 277–303.
53. Dias S, Ades AE. Absolute or relative effects? Arm-based synthesis of trial data. *Res Synth Methods.* 2016;7(1):23–28.
54. Bragge P, Grimshaw JM, Lokker C, et al. AIMD—a validated, simplified framework of interventions to promote and integrate evidence into health practices, systems, and policies. *BMC Med Res Methodol.* 2017;17(1):38.
55. Lokker C, McKibbon KA, Colquhoun H, et al. A scoping review of classification schemes of interventions to promote and integrate evidence into practice in healthcare. *Implement Sci.* 2015;10:27.
56. Michie S, Fixsen D, Grimshaw JM, et al. Specifying and reporting complex behaviour change interventions: the need for a scientific method. *Implement Sci.* 2009;4:40.
57. Hoffmann TC, Glasziou PP, Boutron I, et al. Better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide. *BMJ.* 2014; 348(3):g1687.
58. Pinnock H, Epiphaniou E, Sheikh A, et al. Developing standards for reporting implementation studies of complex interventions (StaRI): a systematic review and e-Delphi. *Implement Sci.* 2015;10:42.