`MetaExplorer`: Facilitating Reasoning with Epistemic Uncertainty in Meta-analysis

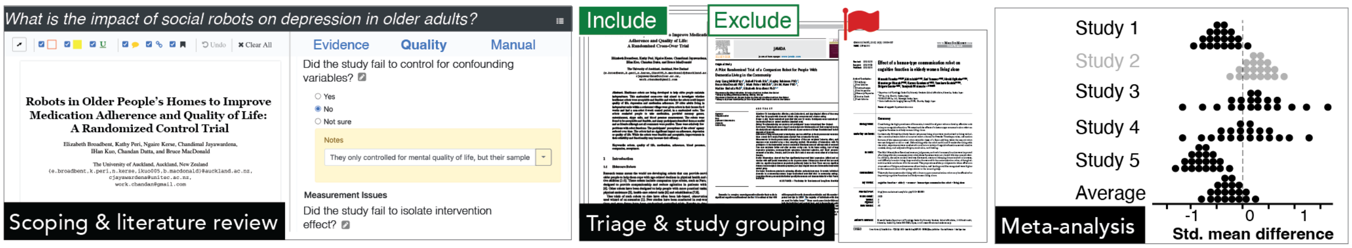
Alex Kale
kalea@uchicago.edu
University of Chicago
Chicago, Illinois, USA

Sarah Lee
sarahlee@stottlerhenke.com
Stottler Henke Assoc.
Seattle, Washington, USA

Terrance Goan
goan@stottlerhenke.com
Stottler Henke Assoc.
Seattle, Washington, USA

Elizabeth Tipton
tipton@northwestern.edu
Northwestern University
Evanston, Illinois, USA

Jessica Hullman
jhullman@northwestern.edu
Northwestern University
Evanston, Illinois, USA



What is the impact of social robots on depression in older adults?

Robots in Older People's Homes to Improve Medication Adherence and Quality of Life: A Randomized Control Trial

Elizabeth Broadbent, Kathy Peri, Ngaire Kerse, Chandimal Jayawardena, IHan Kuo, Chandan Datta, and Bruce MacDonald

The University of Auckland, Auckland, New Zealand
{e.broadbent,k.peri,n.kerse,ikuo005,b.macdonald}@auckland.ac.nz,
cjayawardena@unitec.ac.nz,
work.chandan@gmail.com

Did the study fail to control for confounding variables?

○ Yes
● No
○ Not sure

Notes: They only controlled for mental quality of life, but their sample

Measurement Issues

Did the study fail to isolate intervention effect?

Figure 1: `MetaExplorer` provides a guided process for literature review and meta-analysis with an emphasis on documenting sources of epistemic uncertainty and choosing how to address them during statistical inference. The workflow proceeds in stages from Scoping and literature review where `MetaExplorer` elicits information about each study, to Triage and study grouping where the user resolves sources of epistemic uncertainty, and finally to Meta-analysis where the user views results alongside contextualizing uncertainty.

ABSTRACT

Scientists often use meta-analysis to characterize the impact of an intervention on some outcome of interest across a body of literature. However, threats to the utility and validity of meta-analytic estimates arise when scientists average over potentially important variations in context like different research designs. Uncertainty about quality and commensurability of evidence casts doubt on results from meta-analysis, yet existing software tools for meta-analysis do not provide an explicit software representation of these concerns. We present `MetaExplorer`, a prototype system for meta-analysis that we developed using iterative design with meta-analysis experts to provide a guided process for eliciting assessments of uncertainty and reasoning about how to incorporate them during statistical inference. Our qualitative evaluation of `MetaExplorer` with experienced meta-analysts shows that imposing a structured workflow both elevates the perceived importance of epistemic concerns and presents opportunities for tools to engage users in dialogue around goals and standards for evidence aggregation.

CHI '23, April 23–28, 2023, Hamburg, Germany

ACM ISBN 978-1-4503-9421-5/23/04…$15.00
https://doi.org/10.1145/3544548.3580869

CCS CONCEPTS

• **Information systems → Information systems applications**; *Information retrieval*; • **Human-centered computing → Human computer interaction (HCI)**.

KEYWORDS

Meta-analysis, literature review, epistemic uncertainty

ACM Reference Format:
Alex Kale, Sarah Lee, Terrance Goan, Elizabeth Tipton, and Jessica Hullman. 2023. **`MetaExplorer`: Facilitating Reasoning with Epistemic Uncertainty in Meta-analysis. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23), April 23–28, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3544548.3580869**

1 INTRODUCTION

Summarizing a corpus of scientific literature poses challenges, even for seasoned researchers with domain knowledge. This is especially difficult when the purpose of the review is to be both systematic—including all relevant studies, not just familiar ones—and to inform a decision or practice. The process of extracting and combining data from multiple studies is referred to as *meta-analysis*, and involves searching for and finding relevant papers, ensuring they answer the research question, extracting information (including statistical estimates) from each paper, and combining this information into summary measures. Meta-analyses are common in a variety of fields, including medicine, social welfare, economics, and education, where results from such reviews are perceived as central to "evidence-based" decision making.

To understand why such reviews are difficult, *imagine a scenario* where a human-computer interaction researcher, Kara, consults for a retirement community about whether purchasing social robots would improve the psychological wellbeing of residents. Seeing demonstrations of early social robots (e.g., [47]) excited Kara's clients, but before purchasing anything, they want Kara to verify whether empirical research supports the idea that social robots can improve mental health indicators such as depression. Since social robots are a relatively new invention, Kara anticipates that there are not many controlled studies on them yet, but sets out to review this small literature. She first searches for relevant papers that include "social robots" and "depression," then screens such studies to ensure they answer her research questions. These papers might report measures of depression for groups that use social robots compared to those that do not, or compare depression within individuals before versus after using social robots. Some might report positive effects and others negative effects, with effect sizes varying from negligible to moderate. Kara could use *meta-analysis* to summarize the evidence as an average effect and its variation across studies. However, to do so, Kara must make difficult judgments about the quality of individual study results, whether different measures can be meaningfully combined, and whether evidence from a given study will generalize to the target context one wants to make an inference about.

Known pitfalls when scientists use meta-analysis to estimate intervention effects and inform decisions (e.g., about purchasing social robots) include focusing too strongly on the average effect, and interpreting this average as a 'true' fixed and universal effect [11, 64], despite many systematic reviews (and common-sense expectations about effects in the world [11]) suggesting a range of effects that vary across contexts and study designs. Conceptual frameworks can help by breaking down these variations by sources, such as Methods used, Units studied, Treatment versions, Outcomes measured, and Settings considered (MUTOS) [5, 43], differentiating, for example, issues of internal and external validity [59]. Scientists must weigh these concerns and consider which studies should be included or grouped together in their meta-analysis, and may even decide that meta-analysis is not appropriate for their corpus.

In particular, scientists routinely struggle to account for the impact of *epistemic uncertainty* on results in evidence aggregation [22, 25, 67]. Unfortunately, current software tools for meta-analysis can limit scientists' ability to externalize concerns about these uncertainties in ways that clearly inform statistical inference [33]. This makes scientific review and meta-analysis a useful application area for investigating how software can represent and facilitate reasoning about epistemic uncertainty in evidence aggregation more broadly.

We contribute `MetaExplorer`, a prototype system providing a workflow for eliciting sources of epistemic uncertainty from scientists during literature review and helping them respond to these during meta-analysis. `MetaExplorer` combines several features uncommon among meta-analysis tools: (1) a guided triage process for reasoning about how epistemic uncertainties may threaten inferential validity; (2) an exploratory visualization for reasoning about quantified inferential uncertainty alongside unquantified uncertainty; and (3) visualizations of inferential uncertainty framed as possible outcomes. We created `MetaExplorer` using an iterative user-centered design process with frequent feedback from experienced (n = 5) and expert (n = 3) meta-analysts, and further evaluate it in guided-tour interviews with participants who are experienced and knowledgeable about systematic review and meta-analysis (n = 12). These participants reflect the population of intended users for `MetaExplorer`, scientists who conduct meta-analysis to answer practical questions. Our analysis of these interviews reveals that guided-process tools like `MetaExplorer` seem to derive their benefits and drawbacks in part from challenging users' conceptualizations of analysis tasks. Elevating concerns about epistemic uncertainty, from optional to focal for meta-analysis, requires a structured procedure for handling them. However, in order to flexibly accommodate standards in a variety of scientific domains, these procedures must also adapt in dialogue with the user's analysis goals. Our findings about `MetaExplorer` point toward new ways of designing *analysis tools as partners in deliberation* about ambiguity in data analysis.

2 BACKGROUND

We contextualize our work on `MetaExplorer` in relation to other tools supporting meta-analysis and literature review, interdisciplinary perspectives on reasoning with epistemic uncertainty, and visualization techniques for showing quantified inferential uncertainty.

2.1 Supporting meta-analysis & scientific review

Constructing a meta-analysis from a scientific review entails judging what demarcates populations of studies: meta-analysis assumes that study results are sampled from a statistical population of studies and thus can be meaningfully averaged [15, 37]. A sampled corpus of literature may contain substantial heterogeneity, which is typically accounted for by using a hierarchical (a.k.a. random effects) model to separate variance between and within studies, representing heterogeneity and residual inferential uncertainty, respectively [15]. `MetaExplorer` employs hierarchical models as they are a "gold standard" [37], and emphasizes reasoning about epistemic uncertainty in literature review in terms of what results to aggregate.

Most software tools for scientific review do not support meta-analysis but are flexible in enabling users to document epistemic uncertainty. Tools such as EPPI-Reviewer [66], Distiller SR [53], Covidence [31], and Rayyan [49] provide *document analysis interfaces* focused on annotating and coding articles. Such tools include reference managers like Mendeley [1] and Zotaro [19]. Some document analysis tools focus specifically on helping researchers evaluate quality of evidence—e.g., Newcastle-Ottawa scale [40], GRADE criteria [23], Cochrane Risk of Bias Assessment [26], or Jadad scale [46]. Many of these are structured like checklists, but focus on disjoint or partially overlapping sets of epistemic issues (e.g., risk of bias vs generalizability) and produce different output formats, making it challenging to collate results across scales. Such quality assessment scales are sometimes integrated as options in document annotation tools—e.g., EPPI-Reviewer [66] includes Cochrane Risk of Bias Assessment, and other tools enable such assessments if the user manually implements them [49, 53]. Extending these designs, MetaExplorer's evidence extraction tool interleaves questions typically required for scientific review and meta-analysis with a new quality assessment form. MetaExplorer's quality assessment synthesizes existing scales to cover a union of sources of epistemic uncertainty addressed by other quality assessments and produces a unified output format, where every question receives an answer of 'yes' (there is an issue), 'no', or 'not sure' accompanied by notes documenting the user's rationale. The new scale and unified output format streamline quality assessment during literature review and resolve difficulties around collating results from existing scales.

In current practice, scientists often need to context-switch to *statistical tools* to perform a meta-analysis. For example, CMA [8] and MIX [4] are implemented as add-ons for Microsoft Excel [45]. These tools also include R packages such as metafor [68] and meta [3]. Others have built cross-platform meta-analysis software such as MetaWin [56], OpenMetaAnalyst [70], Annotation Graphs [72], and MetaInsight [50], which are similarly focused on estimating meta-analytic averages. To our knowledge, Cochrane's RevMan [14] is unique in supporting both meta-analysis and document analysis, however, Cochrane's Risk of Bias Assessment [26] is optional and focuses on internal validity only. MetaExplorer explores how to help scientists transition from literature review to meta-analysis in one tool, with an emphasis on quality assessment.

2.2 Reasoning with epistemic uncertainty in data analysis

Epistemic uncertainty is prevalent in data analysis decisions—e.g., how to model a dataset. The necessity of such judgments and problem of how they impact the results of analysis has been dubbed "researcher degrees of freedom" [71]. Pre-registration [48] and multiverse analysis [61, 63] aim to guard against threats to validity by making analysis choices explicit. However, supporting such procedures requires software representation of epistemic uncertainty around analysis choices [33, 38]. Recent work in human-computer interaction addresses this challenge primarily by attempting to guide analysts in selecting among possible models [32, 39, 69, 72] and surfacing provenance about measurements [44, 60].

One major problem in designing software to help scientists reason about epistemic uncertainty that threatens meta-analysis is that scientists conducting research synthesis tend document these sources of uncertainty (e.g., study quality concerns) in ad hoc ways such as by writing them in lab notebooks [33]. Subsequently, they struggle to integrate these uncertainties into their statistical inferences through practices such as sensitivity analysis [33, 38], perhaps because software tools are not designed to help to maintain awareness of uncertainty [57]. Another major challenge for scientists is deciding how to respond to epistemic uncertainty. Prior work [10, 33] characterizes strategies for resolving uncertainty in terms of "suppressing" or ignoring it versus "reducing" or incorporating it into analysis through mechanisms such as statistical modeling. For example, in meta-analysis, if a study result may be biased, scientists should check the impact on results when removing it from their model (i.e., sensitivity analysis) to see if their statistical inference is robust to potential study quality issues. Studies that seem to measure different constructs should be modeled separately for clarity of interpretation. Study results that are not applicable to the target context a scientist wants to make an inference about may still be informative, but including them in meta-analysis leads to estimates that may not generalize. MetaExplorer extends prior work on representing and managing epistemic uncertainty by structuring the meta-analysis workflow around resolving ambiguities of interpretation that influence decisions in evidence aggregation.

2.3 Visualizing inferential uncertainty

Conventional techniques for uncertainty visualization require an approach to uncertainty quantification that produces distributions or boundaries to show, and thus they do not address unquantified epistemic uncertainties (see Section 2.2). For example, among the most common applications of uncertainty visualization are confidence intervals showing *inferential uncertainty* about estimates from a statistical model [17, 42, 65], such as those in forest plots generated by most meta-analysis software (e.g., [3, 4, 8, 14, 50, 56, 68, 70]). Despite their prevalence, previous work on statistical cognition [6, 27, 62] and uncertainty visualization [12, 16, 18, 29, 34–36] finds widespread misinterpretation of interval representations of uncertainty. Drawing on cognitive science suggesting benefits of framing probabilities as frequencies of events [21, 28], alternative techniques, such as *quantile dotplots* [36] show percentiles of the underlying univariate distribution as stacked dots, enabling users to reason about probabilities by counting. Multiple empirical evaluations to date find that quantile dotplots support visual statistical inferences better than interval representations of uncertainty [18, 34, 36]. Following previous work, we adopt quantile dotplots as an alternative to confidence intervals in MetaExplorer's interactive forest plot (see Section 4.1.5).

3 DESIGNING FOR META-ANALYSIS

Our aim in prototyping MetaExplorer was a guided process for meta-analysis that elicits sources of epistemic uncertainty alongside effect size statistics during literature review and propagates these uncertainties, making it easier to conduct meta-analysis with epistemic uncertainty as a primary consideration. Our primary design goals are:

- **Make epistemic uncertainties explicit.** A tool should elicit and explicitly represent sources of epistemic uncertainty about scientific literature.
- **Non-optional quality assessment.** A tool should integrate quality assessment with meta-analysis as a non-optional procedure, without extending the duration of scientific review.
- **Propagating concerns about study quality.** A tool should collate unquantified epistemic uncertainty in ways that can inform statistical modeling, without overwhelming the user.
- **Sensitivity analysis.** A tool should support exploration of possible inferences a user could reasonably make in a meta-analysis (e.g., by including different sets of results).

3.1 Design process

We arrived at the above design guidelines for MetaExplorer through an iterative user-centered design process. We frequently gathered feedback from *potential users*, initially running think-aloud pilot interviews with a paper prototype of the evidence extraction form to investigate what *experienced meta-analysts* might want from a guided processes, and later eliciting informal feedback from *experts in meta-analysis.* The distinction between experienced and expert meta-analysts reflects groups of participants containing some graduate students versus groups strictly comprised of PhDs with decades of experience specialized in research synthesis. We sought input from both experienced meta-analysts and meta-analysis experts because we envisioned `MetaExplorer` as a tool to help scientists conduct quick meta-analyses to answer practical questions, and we wanted to support existing workflows and resolve pain points that scientists see as threats to the validity or feasibility of meta-analysis. Feedback during this process led us to focus on how meta-analysis software can support epistemic uncertainty, rather than a broader set of meta-analysis considerations (e.g., dual review/collaboration features, support for more study designs).

Paper prototype sessions. We created a paper prototype to elicit feedback on what questions belong in an interface for reviewing scientific articles, extracting effect size information, and eliciting epistemic uncertainty. We created the initial paper prototype drawing on best practices for meta-analysis and organizing principles for scientific review [24, 37]. We instructed participants to think aloud while reviewing an article for inclusion in a hypothetical meta-analysis with the prototype. In the second half of these interviews, we prompted an open-ended discussion with participants:

> How can we generalize the process of evidence extraction through a form like the one you just used? What changes would you make to these materials? Are there things you consider when extracting evidence from articles which are not represented in the form? What sort of interface would be ideal for this task?

This protocol first placed the participant in a "work-like situation" [7] which allowed us to observe "reflection in action" [58], enabling us to clarify what makes evidence extraction cumbersome. We conducted pilot interviews with five participants, who were scientists with previous experience with meta-analysis, recruited from our professional network.

Informal feedback from experts. Throughout the development of `MetaExplorer` we met with experts in meta-analysis to share intermediate versions of the tool, so these experts could suggest changes to the tool and raise potentially challenging edge cases. Our general process was to consider new features in a planning document, make a first-pass implementation, and test the interface by coding example articles. This resulted in working examples of complete meta-analyses that we could use to demonstrate the tool for expert meta-analysts. We queried three experts from our professional network, who were all PhDs with extensive experience in evidence synthesis.

Summary of feedback. Feedback from pilot sessions and informal discussions with experts led to improvements in question wording and revealed issues that were especially challenging to reason about. For example, pilot participants struggled to identify which of the many numbers reported in a paper were required for their meta-analysis. The pilot interviews enabled us to try a handful of approaches to orient the user's attention to the information they needed to answer their research question. This design process resulted in `MetaExplorer`'s sequential questionnaire format.

A major theme in pilot interview sessions was the difficulty of judging the quality of evidence presented in a study and its applicability to the participant's research question. Participants pointed out that these kinds of judgments were typically considered optional, corroborating findings from formative work that meta-analysts tend not to formally assess quality of evidence [33]. One participant noted the inadequacy of existing tools for quality assessment: *"these judgments are not black and white" (Pilot05).* This inspired us to synthesize existing quality assessment scales (Section 2.1) into a questionnaire allowing ambiguous 'not sure' responses.

Sharing intermediate versions of `MetaExplorer` with experts helped us choose among possible design strategies for handling epistemic uncertainty. For example, an intermediate version of the tool lacked a triage process for epistemic uncertainty and instead displayed all elicited risks of bias alongside study results in an interactive visualization, explicitly depicting different sources of epistemic uncertainty for the user to explore. We shared this version of `MetaExplorer` with two experts, a scientist in the Navy and a professor at a major research university. Although they agreed that this design made epistemic uncertainty explicit, they also said it undermined confidence in the ability to produce a useful meta-analytic estimate. Experts requested a triage process (see Section 4.1.4) where users can express their *"first gut feel" (Expert01)* about how important potential sources of bias might be and whether they can be resolved before viewing study results. For this reason, we pivoted our design to focus on helping users reduce epistemic uncertainty, rather than finding more elaborate ways to display it.

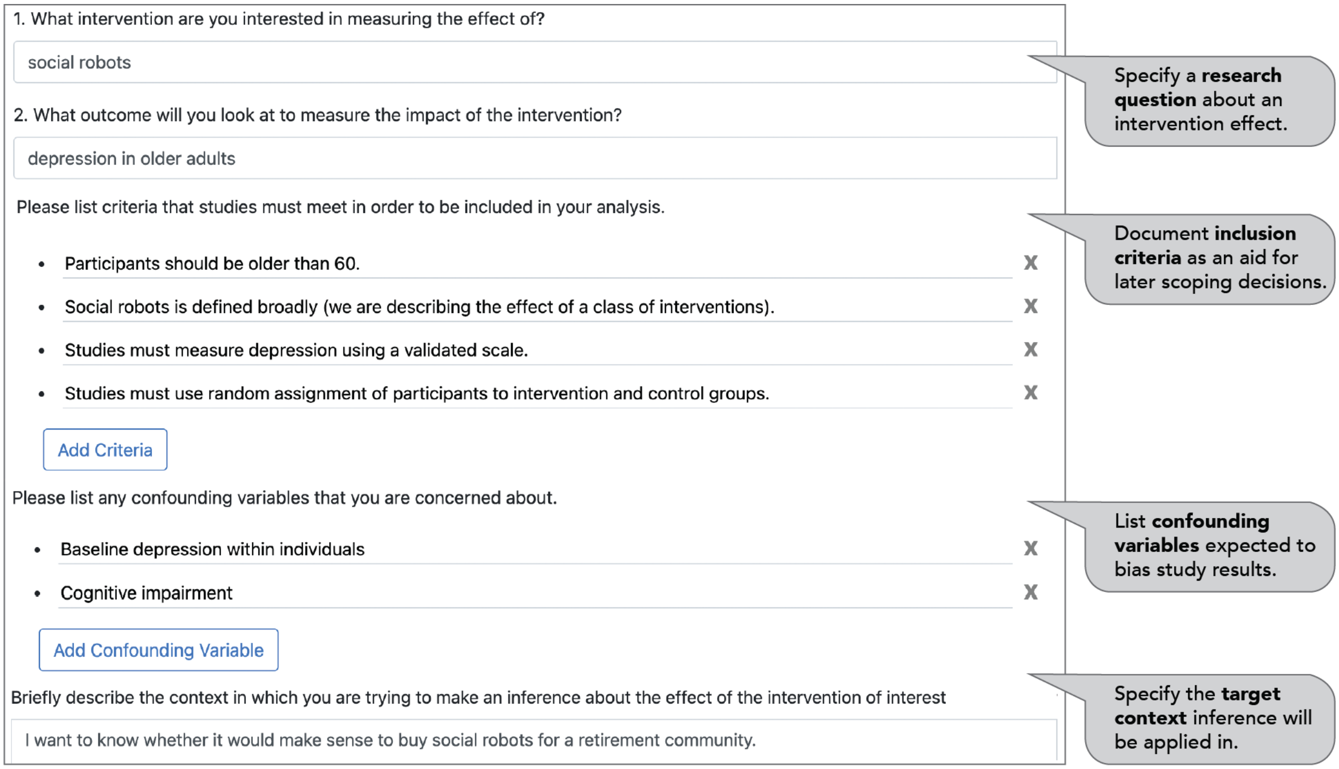
4 SYSTEM

We present `MetaExplorer`, a prototype meta-analysis tool that elicits sources of epistemic uncertainty in literature review and propagates them alongside quantitative study results during meta-analysis.

4.1 Exposition & use case scenario

Scenario: To demonstrate the `MetaExplorer` workflow, we return to Kara, the hypothetical scientist investigating the influence of social robots on older adults' mental health (Section 1).

4.1.1 Scoping. The `MetaExplorer` workflow begins with a view where users express research topics and possible research questions per topic. `MetaExplorer` supports research questions of the form, "What is the impact of `<intervention/>` on `<outcome/>`?" which are typical for meta-analyses (Figure 2, first row). Next, the user describes what will count as evidence by specifying study inclusion criteria, potential confounding variables, and the target context to which the meta-analytic inference will be applied (Figure 2, second-fourth rows). Like pre-registration, answering these questions helps users focus their review and creates a mechanism for personal accountability, something they can check when unsure about how to handle a study. However, unlike pre-registration, users can return to this page and edit the scope of the review as they review the literature.



1. What intervention are you interested in measuring the effect of?

social robots

2. What outcome will you look at to measure the impact of the intervention?

depression in older adults

Please list criteria that studies must meet in order to be included in your analysis.

- Participants should be older than 60.
- Social robots is defined broadly (we are describing the effect of a class of interventions).
- Studies must measure depression using a validated scale.
- Studies must use random assignment of participants to intervention and control groups.

Please list any confounding variables that you are concerned about.

- Baseline depression within individuals
- Cognitive impairment

Briefly describe the context in which you are trying to make an inference about the effect of the intervention of interest

I want to know whether it would make sense to buy social robots for a retirement community.

X
Add Criteria
Add Confounding Variable
Specify a **research question** about an intervention effect.
Document **inclusion criteria** as an aid for later scoping decisions.
List **confounding variables** expected to bias study results.
Specify the **target context** inference will be applied in.

Figure 2: Reviews in MetaExplorer start with the Scoping view, where the user documents choices that will guide what evidence gets included in their summary of the scientific literature.

Scenario: Kara uses MetaExplorer to document the target context for her inference (a retirement community interested in social robots for mental health; Figure 2, fourth row), so she can focus on applicable papers. Kara specifies her research question, "What is the impact of social robots on depression in older adults?" since depression is the most commonly measured outcome in the sparse literature that quantitatively measures the impact of social robots. Kara scopes her review using inclusion criteria—e.g., documenting that 'social robots' refers to a class of interventions rather than a specific robot. She notes concern about study results that fail to control for baseline depression.

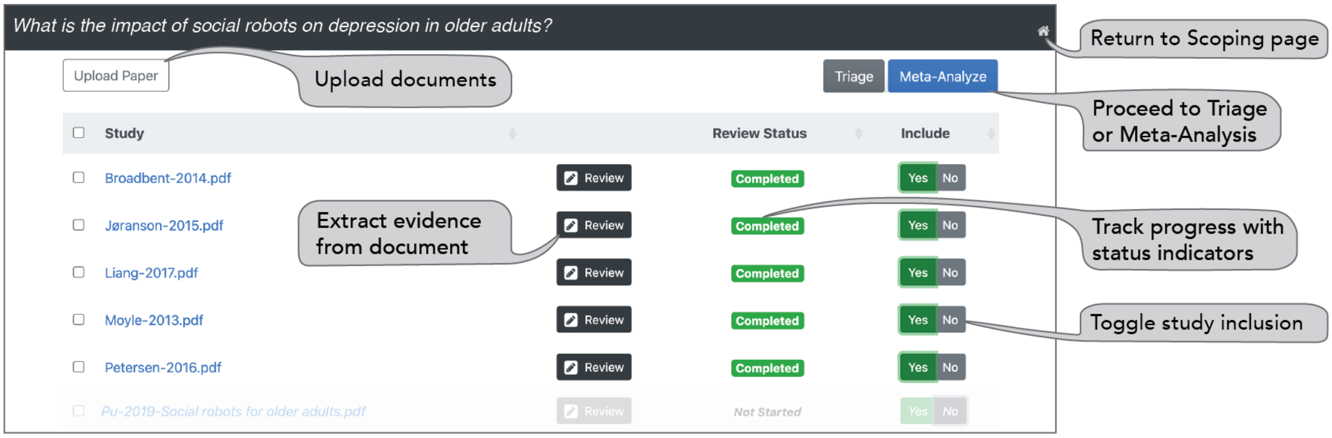
4.1.2 Review management. After scoping, users begin to review scientific articles on their topic. MetaExplorer provides a review management view that enables users to upload articles, toggle provisional study inclusion choices, and navigate between system components for reviewing literature, triaging epistemic uncertainty, and meta-analysis (Figure 3).

Scenario: To save time searching for papers, Kara decides to start by replicating an existing meta-analysis [54], thus she already has documents to upload. For a new meta-analysis, Kara would need to search for articles via online databases and citation networks. Kara uses the review management view to navigate between interfaces for evidence extraction, triage, and meta-analysis.

4.1.3 Evidence extraction. Literature review happens in MetaExplorer's evidence extraction tool. Its major components facilitate (1) annotating documents, (2) recording how a study was run and its results, (3) recording sources of epistemic uncertainty through quality assessment, and (4) checking terminology and coding procedures that come up in scientific review. The interface is a split view with a PDF annotation tool on the left and a dynamic web form on the right (Figure 4) containing three navigation tabs: evidence extraction, quality assessment, and coding manual.

The PDF annotation tool enables users to highlight, draw boxes, underline, and comment on the PDF, and to bookmark and link selected locations in the PDF document (Figure 4, left column). We designed this tool after observing how participants used printed articles during pilot interviews.

The evidence extraction form guides the user through coding each article in a meta-analysis (Figure 4, middle column). The form includes sections about study identity (i.e., authors, year, title), study context (e.g., what was the study design? What were the mechanisms for experimental control?), participants (i.e., who were the participants? How were they recruited?), measurement (e.g., how were variables defined and measured? What comparisons are reported in the article?), and effect size (i.e., what statistics should be used in a meta-analysis?). The form is dynamic: user responses to questions about study design and measurement determine what statistical information the tool asks for. For example, if the user



What is the impact of social robots on depression in older adults?

Upload Paper

Triage Meta-Analyze

☐	Study		Review Status	Include
☐	Broadbent-2014.pdf	Review	Completed	Yes No
☐	Jøranson-2015.pdf	Review	Completed	Yes No
☐	Liang-2017.pdf	Review	Completed	Yes No
☐	Moyle-2013.pdf	Review	Completed	Yes No
☐	Petersen-2016.pdf	Review	Completed	Yes No
☐	Pu-2019-Social robots for older adults.pdf	Review	Not Started	Yes No

Upload documents

Extract evidence from document

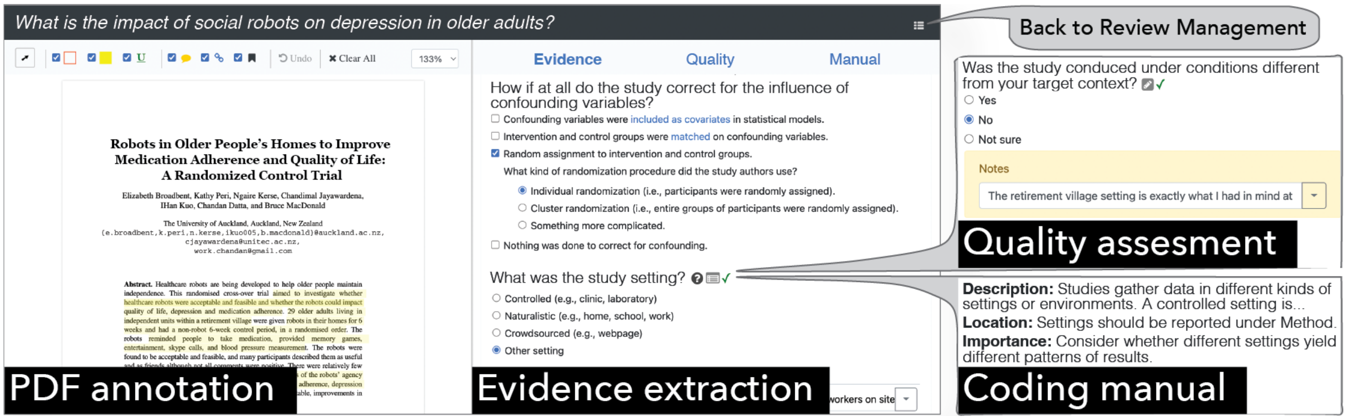
Return to Scoping page

Proceed to Triage or Meta-Analysis

Track progress with status indicators

Toggle study inclusion

Figure 3: The Review management view in `MetaExplorer` is a tabular interface where users can upload PDFs to a database, track their progress in reviewing each document, and toggle the inclusion or exclusion of each article.



What is the impact of social robots on depression in older adults?

Robots in Older People's Homes to Improve Medication Adherence and Quality of Life: A Randomized Control Trial

Elizabeth Broadbent, Kathy Peri, Ngaire Kerse, Chandimal Jayawardena, IHan Kuo, Chandan Datta, and Bruce MacDonald

The University of Auckland, Auckland, New Zealand
{e.broadbent,k.peri,n.kerse,ikuo005,b.macdonald}@auckland.ac.nz,
cjayawardena@unitec.ac.nz,
work.chandan@gmail.com

Abstract. Healthcare robots are being developed to help older people maintain independence. This randomised cross-over trial aimed to investigate whether healthcare robots were acceptable and feasible and whether the robots could impact quality of life, depression and medication adherence. 29 older adults living in independent units within a retirement village were given robots in their homes for 6 weeks and had a non-robot 6-week control period, in a randomised order. The robots reminded people to take medication, provided memory games, entertainment, skype calls, and blood pressure measurement. The robots were found to be acceptable and feasible, and many participants described them as useful and as friends although not all comments were positive. There were relatively few … of the robots' agency … adherence, depression … able, improvements in

PDF annotation

Evidence Quality Manual

How if at all do the study correct for the influence of confounding variables?
☐ Confounding variables were included as covariates in statistical models.
☐ Intervention and control groups were matched on confounding variables.
☑ Random assignment to intervention and control groups.
What kind of randomization procedure did the study authors use?
● Individual randomization (i.e., participants were randomly assigned).
○ Cluster randomization (i.e., entire groups of participants were randomly assigned).
○ Something more complicated.
☐ Nothing was done to correct for confounding.

What was the study setting? ✓
○ Controlled (e.g., clinic, laboratory)
○ Naturalistic (e.g., home, school, work)
○ Crowdsourced (e.g., webpage)
● Other setting

workers on site

Evidence extraction

Back to Review Management

Was the study conduced under conditions different from your target context? ✓
○ Yes
● No
○ Not sure

Notes
The retirement village setting is exactly what I had in mind at

Quality assesment

Description: Studies gather data in different kinds of settings or environments. A controlled setting is...
Location: Settings should be reported under Method.
Importance: Consider whether different settings yield different patterns of results.

Coding manual

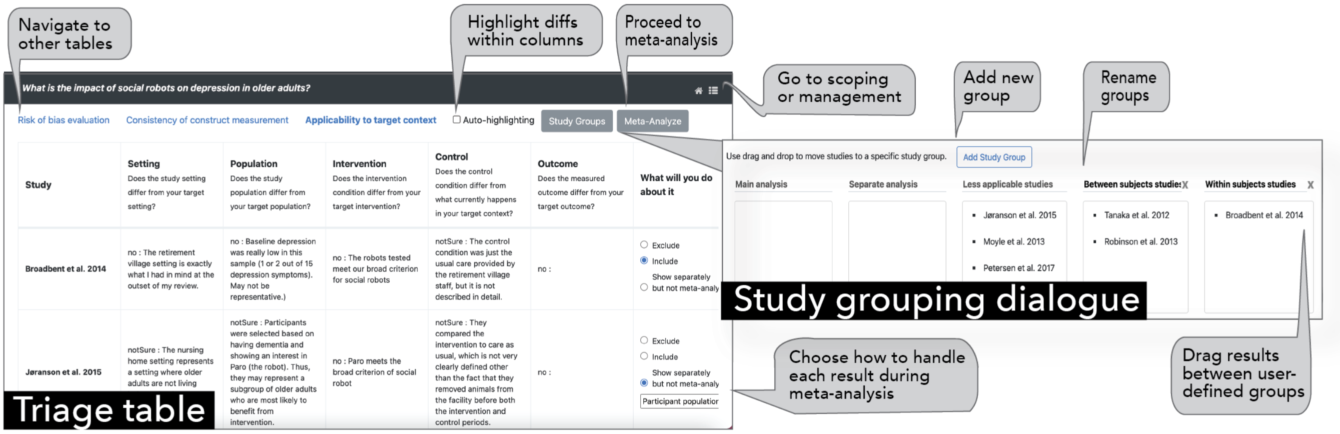
Figure 4: The Evidence extraction tool in `MetaExplorer` is where users pull effect size information from each study in their review and document concerns about epistemic uncertainty. The major components of this tool are PDF annotation where users read and markup documents, Evidence extraction where `MetaExplorer` elicits information about the study design and results, Quality assessment where `MetaExplorer` elicits judgments of epistemic uncertainty, and a Coding manual which guides evidence extraction.

indicates that a study result adjusts for potential confounding variables, the form will later ask which covariates were adjusted for. Evidence extraction culminates in an evidence table used for data input. The evidence table asks for only the statistics presented in the article which are needed for meta-analysis.

The quality assessment form asks the user to judge the quality of evidence presented in a given article (e.g., Figure 4, top of right column) as they move through the evidence extraction form. The form includes sections on risk of selection bias (e.g., did the study fail to control for important confounding variables?), measurement issues (e.g., did the study use a validated measurement scale?), and applicability (e.g., are the participants different from the population the user would like to make an inference about?). Each quality assessment question is linked to a related question in the evidence extraction form, such that users can navigate quickly between related questions across the two forms.

The coding manual provides explanations of questions in the evidence extraction form for new users (e.g., Figure 4, bottom of right column). Upon clicking beside a question, users receive a *description* of the question with links to online definitions for necessary jargon; a *location* in the article where the user might find this information; and a brief explanation of the question's *importance* in a typical review.

Scenario: For each article in her review, Kara follows the guided process for evidence extraction. She fills her database with the necessary quantitative information and develops a sense of where the interpretation of the literature is uncertain. For one study, she notes that the authors omitted information about participants, making it difficult to say whether there is selection bias. Another study failed to control for individual differences in baseline depression, a confounder she is concerned about *a priori*. Kara notes that most studies in her corpus recruited participants with dementia, who


Navigate to other tables
Highlight diffs within columns
Proceed to meta-analysis
Go to scoping or management
Add new group
Rename groups
Triage table
Study grouping dialogue
Choose how to handle each result during meta-analysis
Drag results between user-defined groups
Risk of bias evaluation
Consistency of construct measurement
Applicability to target context
☐ Auto-highlighting
Study Groups
Meta-Analyze
Add Study Group

What is the impact of social robots on depression in older adults?

Study	Setting Does the study setting differ from your target setting?	Population Does the study population differ from your target population?	Intervention Does the intervention condition differ from your target intervention?	Control Does the control condition differ from what currently happens in your target context?	Outcome Does the measured outcome differ from your target outcome?	What will you do about it
Broadbent et al. 2014	no : The retirement village setting is exactly what I had in mind at the outset of my review.	no : Baseline depression was really low in this sample (1 or 2 out of 15 depression symptoms). May not be representative.)	no : The robots tested meet our broad criterion for social robots	notSure : The control condition was just the usual care provided by the retirement village staff, but it is not described in detail.	no :	○ Exclude ● Include ○ Show separately but not meta-analy
Jøranson et al. 2015	notSure : The nursing home setting represents a setting where older adults are not living	notSure : Participants were selected based on having dementia and showing an interest in Paro (the robot). Thus, they may represent a subgroup of older adults who are most likely to benefit from intervention.	no : Paro meets the broad criterion of social robot	notSure : They compared the intervention to care as usual, which is not very clearly defined other than the fact that they removed animals from the facility before both the intervention and control periods.	no :	○ Exclude ○ Include ● Show separately but not meta-analy Participant populatior

Use drag and drop to move studies to a specific study group.

Main analysis	Separate analysis	Less applicable studies	Between subjects studies X	Within subjects studies X
		Jøranson et al. 2015 Moyle et al. 2013 Petersen et al. 2017	Tanaka et al. 2012 Robinson et al. 2013	Broadbent et al. 2014

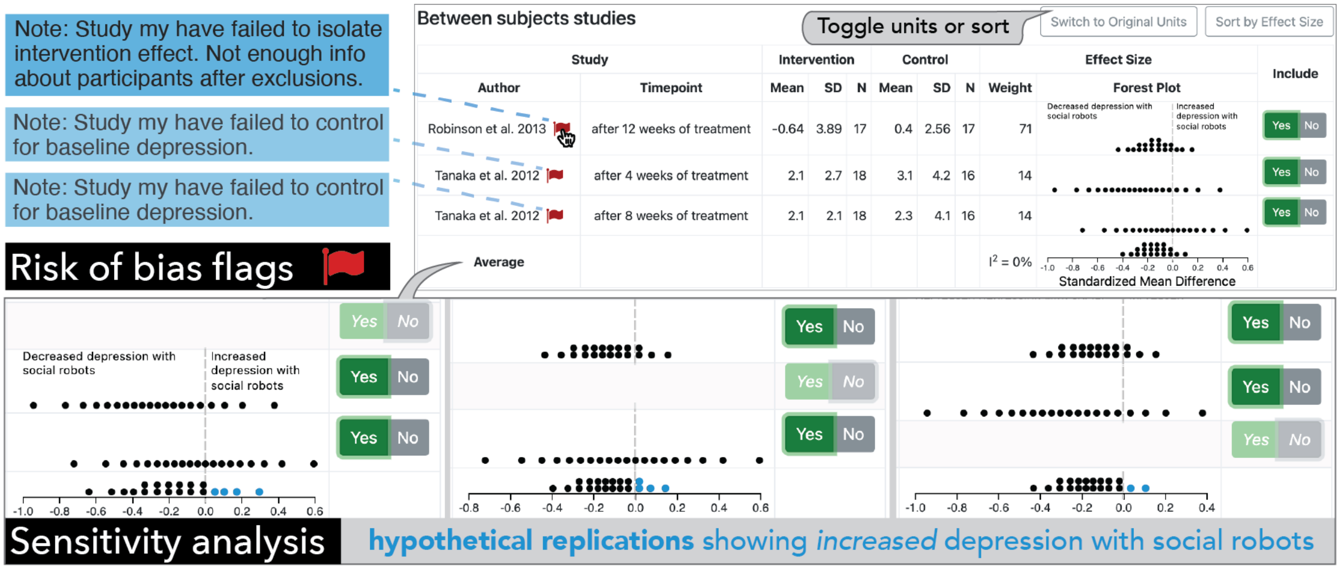
Figure 5: This Triage table presents judgments about the applicability of study results to the target context, which were elicited during quality assessment. MetaExplorer's triage view also generates similar tables for judging risk of bias and consistency of construct measurement. The Study grouping dialogue gives users control over the grouping of study results in meta-analysis. Actions selected by the user in the rightmost column of the triage table generate default groups (main analysis, separate analysis, and less applicable studies). Study grouping allows users to edit these defaults.

live less independently and experience greater cognitive decline than the people she intends to make an inference about. She will need to consider these concerns in making a statistical inference. Without MetaExplorer, Kara might have written these concerns down in her lab notebook and forgotten about them, or found them hard to reconcile, upon meta-analyzing her corpus [33].

4.1.4 Triage & study grouping. MetaExplorer provides a triage process to help users reduce elicited epistemic uncertainty into a set of considerations they believe should guide statistical inference after they complete excluding or reviewing studies in the review management view. The triage view includes three tables: (1) risk of bias, helping the analyst avoid potentially misleading evidence, (2) consistency of construct measurement, helping the analyst interpret estimates arising from different procedures, and (3) applicability, helping the analyst reason about generalizing study results to their target context. Triage tables contain one row per study result and one column for each relevant question from the evidence extraction and quality assessment forms (e.g., Figure 5). Each table corresponds to different challenges that come up in meta-analysis (Section 2.2) and each challenge warrants a different *action*. The rightmost column of each table asks the user, *"What will you do about it?"* with radio buttons that enable the user to include, exclude, or—depending on the triage table—flag results for risk of bias, group results into separate analyses based on what they seem to measure, or show results from less applicable studies without meta-analyzing them. MetaExplorer auto-highlights differences between cells in each column to draw the user's attention to discrepancies between study designs. The triage view also provides a drag-and-drop dialogue (Figure 5, right) for creating and naming study groups for meta-analysis and dragging results between these groups. The outputs of the triage process are study groups to be meta-analyzed separately and flags summarizing concerns about potentially biased results.

Scenario: Kara uses triage to get an overview of her corpus and decide how evidence should be combined in a meta-analysis. In the risk of bias table, she places flags on two studies which may not have controlled for confounding variables. In the consistency of construct measurement table, she sees that her review contains both within- and between-subjects study designs, which she decides to analyze separately because within-subjects effects represent a different construct (i.e., average treatment effect on an individual) than between-subjects effects (i.e., average treatment effect in a population). In the applicability table, Kara sees that many studies in her corpus recruited only participants with dementia, which is not the population she wants to make an inference about. She decides to view the results of these less applicable studies separately without meta-analyzing them. The review that Kara replicated [54] did not separate within- versus between-subjects results, and combined evidence across populations of participants with versus without dementia. Kara realizes that her meta-analysis will yield a set of contextualized estimates rather than a single estimate that averages over many potentially important variations in the corpus. Although this will make her result less concise, she thinks it is a more realistic summary of available evidence.

4.1.5 Meta-analysis & visualization. The interactive visualization summarizes all results included in a review and enables the analyst to perform sensitivity analysis, assessing how the estimates from meta-analysis change depending on the set of study results included in the model. This final step facilitates quick explorations of simple meta-analytic models in light of epistemic uncertainty documented during review. The D3-generated [9] MetaExplorer visualization is modeled after a forest plot (e.g., Figure 6). Each study group defined in triage gets its own table, including the group of less applicable studies that are shown but not meta-analyzed. Each table row contains summary information about a specific study result alongside a quantile dotplot [36] showing the quantitative result in standardized effect size units. At the top of each forest plot



Note: Study my have failed to isolate intervention effect. Not enough info about participants after exclusions.

Note: Study my have failed to control for baseline depression.

Note: Study my have failed to control for baseline depression.

Risk of bias flags 🚩

Between subjects studies

Toggle units or sort

Study		Intervention			Control			Effect Size		Include
Author	Timepoint	Mean	SD	N	Mean	SD	N	Weight	Forest Plot	
Robinson et al. 2013 🚩	after 12 weeks of treatment	-0.64	3.89	17	0.4	2.56	17	71		Yes No
Tanaka et al. 2012 🚩	after 4 weeks of treatment	2.1	2.7	18	3.1	4.2	16	14		Yes No
Tanaka et al. 2012 🚩	after 8 weeks of treatment	2.1	2.1	18	2.3	4.1	16	14		Yes No
Average								I² = 0%		

Sensitivity analysis

hypothetical replications showing *increased* depression with social robots

Figure 6: MetaExplorer's visualization displays summarized epistemic uncertainty alongside quantitative evidence. Mousing over risk of bias flags shows user-generated annotations from triage in a tooltip. Clicking toggle buttons in the table facilitates interactive sensitivity analysis to examine how study inclusion choices impact averages within study groups defined in triage. Quantile dotplots frame each sampling distribution of effect estimates as 20 hypothetical replications drawn from of a given population of studies.

are buttons to "Sort [table rows] by effect size" and to "Convert [study results] to original measurement units", adding independent x-axis scales in each row to show non-standardized effects rather than the standardized effect sizes [13] typically used in meta-analysis, since non-standardized effect size can provide important context and more robust estimates under certain conditions [2]. Effect sizes supported in MetaExplorer include non-standardized and standardized mean differences for continuous measures as well as risk differences and log odds ratios for dichotomous measures. Each table uses a common x-axis scale to facilitate comparisons across rows of the forest plot, and the bottom row in each shows the meta-analytic average effect size within the study group. A flag icon appears on rows where the user flagged the results for risk of bias, which users can mouseover to see a description of their concerns. The rightmost cell in each row contains a toggle button for conducting sensitivity analysis by including or excluding results.

Scenario: MetaExplorer's visualization shows Kara three forest plots corresponding to her three study groups. One group shows between-subjects results (shown in Figure 6) which were flagged for risk of bias. Although these results might suggest that social robots reduce depression in older adults, Kara explores the space of possible inferences, using sensitivity analysis to determine that the meta-analytic average is not robust to inclusion choices. The results of the within-subjects comparisons, shown in a second forest plot, seem inconclusive. Kara inspects her third forest plot of less applicable studies that were not meta-analyzed. She sees mixed results in studies that recruited participants with dementia. It seems like there is little evidence in this literature that social robots reduce depression in older adults. Although this conclusion is in line with the meta-analysis Kara replicated [54], she can now provide better reasons for her clients about *why* investing in social robots would be premature given current scientific evidence. The original meta-analysis averaged all of these study results together, suppressing epistemic uncertainty which gives these results meaning to produce a single estimate, whereas Kara's analysis with MetaExplorer produced groups of results informed by epistemic uncertainty that better describe how the literature is methodologically scattered and empirically inconclusive. While she is disappointed not to have a single straightforward result to report to her client, Kara thinks that revealing the messiness of the literature is an honest result, which both suggests opportunities to improve future research and answers her practical question.

5 QUALITATIVE EVALUATION OF METAEXPLORER WITH META-ANALYSTS

To evaluate MetaExplorer, we conducted a qualitative user study with 12 scientists knowledgeable about using meta-analysis for different ends across a range of disciplines. We structured our evaluation as a guided tour of MetaExplorer during which we interviewed potential users about how they saw the tool supporting their specific meta-analysis workflows. We synthesized the results of these interviews into a set of themes capturing appraisals of MetaExplorer along multiple dimensions (e.g., usability, trust) as well as remaining challenges and opportunities in designing for meta-analysis.

5.1 Participants

We recruited 12 knowledgeable meta-analysts, without overlap with previous participants from the design process (Section 3.1), for our interviews. This was a convenience sample recruited from our professional network via email and Twitter. All participants had sufficient previous experience conducting meta-analyses to inform workflow preferences and other valuable perspectives about scientific review. Participants were academic researchers in eight countries, mostly in Europe and North America. Three participants study technology, four study education, two study biological science, one is a cognitive scientist, and two are quantitative methodologists. The sample composition reflects the intended users of MetaExplorer, scientists across a variety of domains with previous experience conducting meta-analysis.

5.2 Interviews

The interviews were structured as a guided tour of MetaExplorer. We held interviews on Zoom and saved recordings of each interview for subsequent analysis. In the first 40-50 minutes of each interview, the interviewer walked participants through the functionality and workflow of MetaExplorer in detail to get their feedback about workflow and features. We instructed participants to, *"Please speak up if you have impressions about how various software features might be useful to you or how they might create barriers to your work."* In the last 10-20 minutes of each interview, the interviewer asked participants two high-level questions to prompt discussion about MetaExplorer. First, the interviewer asked, *"What merits and drawbacks to you see in a guided process for meta-analysis?"* Second, the interviewer asked, *"Does MetaExplorer change the way you think about epistemic uncertainty in meta-analysis? If so, in what ways?"*

5.3 Qualitative analysis

The first author reviewed and coded video recordings from all 12 interviews. We adopted a lightweight coding scheme to analyze what participants said about MetaExplorer and meta-analysis more broadly, starting with open codes describing what we discussed with participants. We used deductive labels for *affordances (A)* and *drawbacks (D)* of MetaExplorer, as well as *feature requests (FR)* and *pain points in current practice (PP)*. We also labeled open codes inductively based on the topics that participants frequently raised: *epistemic uncertainty (EU)*, *usability (U)*, *collaboration (C)*, and *domain specificity (DS)*. For codes associated with a concise and interesting quote, we transcribed the relevant portion of the recording. We iteratively grouped open codes and quotes into themes and tensions following an affinity diagramming procedure.

5.4 Results

Overview of results. Our qualitative analysis surfaced two primary themes: (1) resolving versus propagating epistemic uncertainty, and (2) imposing structure on scientific workflows for which no normative process is available. Participants' comments generally supported our design hypothesis that providing a guided process to elicit and resolve human judgments of epistemic uncertainty should contribute to more trustworthy meta-analyses. Participants envisioned the role of MetaExplorer as building confidence in the review by assisting humans in finding *"what to compare with what and which data to extract" (P03)* and *identifying the difficult spots in the coding sheet (P11)*, which reflect epistemic uncertainty that may need to be resolved by a team of coders. Guiding human judgments and making them explicit in MetaExplorer facilitates open analyses that can be shared with colleagues for audit *(P01, P12)*, which aids in socially distributed construction of knowledge and trust across networks of scientists [41]. However, different standards and procedures are meaningful in different scientific disciplines. As a consequence, participants disagree about whether MetaExplorer expects a review procedure that is too rigid *(e.g., P02, P08)* or not rigid enough *(e.g., P06)*. This speaks to challenges and opportunities—both specific to meta-analysis and broadly applicable—in creating software that encourages researchers to adopt new practices as part of statistical reform.

5.4.1 *Resolving versus propagating uncertainty.*

Participants often questioned whether to resolve or propagate epistemic uncertainty, by either dismissing sources of uncertainty as negligible or summarizing concerns for further consideration later in analysis. We summarize observations on when participants face this choice, how MetaExplorer might help, and challenges that make this choice difficult to design for.

Shades of gray in scoping decisions. Deciding whether to resolve or propagate epistemic uncertainty surfaced primarily as participants considered MetaExplorer's support for scoping decisions. Participants described how they draw boundaries around their corpus to meta-analyze enough evidence to be informative without including so much variety as to obfuscate their inference. This balance becomes difficult when the literature is sparse or heterogeneous in construct definitions or measurements *(P01)*, which only becomes clear *"once you've looked at a dozen on more studies" (P12)*. Participants comments supported our hypothesis that the scope of a meta-analysis cannot be fixed or 'preregistered' from the outset of a review, but instead must be reconsidered throughout the review.

Multiple participants commented that MetaExplorer helps to track the evolving scope of a review. One participant *(P03)* typically keeps a notebook of scoping decisions and remarked that MetaExplorer's scoping page would be more systematic. Another participant *(P05)* bemoaned how ad hoc workflows for handling epistemic uncertainty can erode sense of scope and introduce mission drift about the goals of meta-analysis. Without a tool like MetaExplorer to document scope and guide study inclusion decisions, participants *(e.g., P04, P05)* must invent their own systems of organization and accountability.

Participants saw MetaExplorer assisting with scoping decisions most directly by asking users about shades of gray in study inclusion—i.e., reasoning about *"if the causal inference [supported by a result] is strong or not." (P08)*. Participants clarified that they typically resolve concerns about applicability by narrowing scope rather than by considering what can be learned from different groups of studies.

> *We don't look at interventions, for example, in students with disabilities if our population of interest is English language learners... The target context becomes part of the inclusion criteria. (P10).*

However, what it means to generalize can become ambiguous. One participant described how in academic research *"There's not always an applied target context." (P03)*. Participants *(P05, P10)* told us that decisions about how to parse the literature are informed by norms, which can feel arbitrary. MetaExplorer makes these considerations explicit.

> *'Does this study fit the context I want to generalize about?' It's something that I've vaguely heard people think about, but it's not something that I've seen anybody put into a tool like this. I think that's great because a lot of meta-analyses are: find everything you can, throw it into a big pot, and stir, and out comes something that is of dubious usefulness for particular purposes, like when you are trying to make decisions. (P12).*

However, sometimes participants reported there is no satisfying way to scope a review. For example, *"We actually shelved this meta-analysis on data literacy tools because... the way that people operationalize data literacy is so varied and diverse that it actually doesn't make sense to compare." (P09).* With a larger scale of about 60-100 studies, several participants *(e.g., P04)* said grouping results for meta-analysis becomes more difficult, even with MetaExplorer's triage process, because the number of possible groupings grows with the number of results.

Preference for statistical approaches to uncertainty. Participants reported preferring to use statistical tools to resolve questions about how and whether measurements should be combined. For example, some participants valued quantitative feedback for inclusion decisions: *"I like having the ability to run sensitivity analysis. Like, if something looks off, how much does it change the results?" (P08).* Many participants *(e.g., P03, P12)* wanted to use hierarchical models to account for how sources of variation are clustered depending on study designs. By default MetaExplorer applies a separate hierarchical model to each user-defined study group, however, it does not handle special cases where measurements are inherently correlated—e.g., when combining multiple measurements of the same sample. Because MetaExplorer doesn't enable such complexity in modeling, one participant *(P03)* worried it may not encourage users to be ambitious enough about incorporating a wide variety of evidence into meta-analysis.

Participants disagreed about adding more complex modeling features, but some wanted the reassurance of verifying what models MetaExplorer runs. *"It may not estimate or run the models the way that I would need to to publish my papers, but I'm not totally sure." (P07).* Some participants *(P03, P06, P11)* wanted to manipulate the underlying R code. In contrast, one methodologist and tool builder *(P12)* recommended not revealing model specifications, acknowledging that this would be inaccessible to less experienced users. Future tools like MetaExplorer could strike a better balance in model exposure by having an optional view that makes code available for more expert users. However, we question whether novice users should rely on meta-analytic models without understanding them.

Collaboration. Often sources of epistemic uncertainty cannot be resolved through statistical approaches—e.g., when methodological variations do not form clear clusters—and deliberations among colleagues play a crucial role in deciding how to handle a concern. Participants viewed the tool as a skeptical collaborator in such deliberations.

> *I would model this tool to be a grouchy reviewer that constantly convincing me not to publish the study because I don't have a corpus that is good enough, or I don't have enough certainty. (P09).*

This participant valued MetaExplorer as a way of organizing knowledge to promote reflection. Another participant expanded on this, remarking that the tool pushes users to discuss what would count as a generalizable inference in the target context.

> *Now that I've seen this, I really think that needs to be an integral part of a meta-analysis. I have to admit that in meta-analyses I've been involved in, these conversations didn't come up that much. I don't remember having deep, long conversations about how studies contribute to making policy decisions for particular situations in a particular context. (P12).*

Participants frequently commented that MetaExplorer would make an excellent collaboration platform. Three participants *(P05, P06, P07)* bemoaned the difficulty of finding free literature review tools that support synchronous collaboration. One participant described how collaborating through reference managers can lead to epistemic uncertainty getting lost in communication.

> *Mendeley did a whole lot of heavy lifting for one of the meta-analyses I completed years ago. We just couldn't find anything... It just wasn't streamlined, and it would get really frustrating because inevitably someone would say, 'Oh, I left you a note about that three months ago.' (P05).*

MetaExplorer facilitates progress tracking through indicators in the review management view of what has (not) been coded. We envision extending this interface to include action items for collaborators, e.g., assigning people to documents, requesting clarification on codes, or resolving disagreements through *dual review*—independent coding by multiple scientists, which was the most common feature request.

Multiple participants *(P04, P07, P11)* remarked on the affordances of MetaExplorer's guided process for training teams of coders with mixed levels of experience, and they said that this sort of coordination usually takes a lot of time and energy. We observe that much of what teams need to train and coordinate about involves the handling epistemic uncertainty (e.g., what needs to be coded to differentiate study groups). MetaExplorer provides workflows dedicated to handling these concerns and in doing so makes it less likely teams of coders will lose important contextualizing information.

5.4.2 Imposing structure without a clear normative procedure. Our analysis of interviews suggests that the primary tension around designing for meta-analysis is how much structure to impose on the process. We find a striking contrast between consensus around the need for standardization in research synthesis and participants' idiosyncratic preferences about what standards are meaningful in their domain.

Need for structure. All participants highlighted the benefits of MetaExplorer's streamlined process. Scaffolding document analysis helps users think through coding decisions *(P07)*, structures resulting knowledge *(P09)*, prevents decision paralysis regarding *"what to worry about" (P10)*, and could reduce variance in results across research teams *(P11)*.

Participants contrasted MetaExplorer's guided process with their typical, more ad hoc approach. *"I typically think about [epistemic uncertainty] more manually, less systematically. It comes up all the time, but the tool allows you to have a very strict, very formal way of dealing with it." (P02)*. Another participant echoed, *"It helps to find weaknesses or blind spots that you hadn't thought about, moreso than if you were to do it more chaotically." (P03)*. Participants *(P04, P11)* mentioned often adding risk of bias items to coding spreadsheets midway through a review, and then re-coding articles for previously *"hidden"* information. Beyond structuring their thinking, multiple participants *(P05, P08)* appreciated how MetaExplorer backed their work with a relational database, which reduces the time required for data cleaning in meta-analysis, e.g., from months to minutes. MetaExplorer generates triage tables from this relational database, another data management automation that one participant particularly appreciated. *"When I was describing the spreadsheet we did, it looked pretty much like this. The fact that this spreadsheet gets generated as I'm doing each review—it's very helpful not to have to do this by hand." (P09)*.

Need for customization. Scientific fields have different ways of designing and reporting studies, so participants frequently requested to tailor MetaExplorer to their domain, similar to customizing codebooks in spreadsheets. For example, *"Would you make this more flexible for people who are in engineering or ecology and evolution? Because our experiments or studies are very much different than social psychology, like a lot of ecology and evolution is observational." (P05)*. This echos concerns from other participants, e.g., that MetaExplorer is geared toward a *"specific type of research design" (P09)* in ways that rule out qualitative evidence, and that MetaExplorer doesn't support certain standards like PRISMA [52], MUTOS [5, 43], or PICOTS [55] *(P06, P08)*, which are popular in medicine. On the other hand, some participants *(e.g., P08, P11)* found MetaExplorer sufficiently aligned with the spirit of these standards in encouraging documentation of and reflection about review scope.

One form of document analysis that requires considerable codebook customization is qualitative evidence synthesis. A common grievance with MetaExplorer *(P04, P05, P10)* was prioritizing quantitative meta-analysis over qualitative systematic review, especially eliciting research questions in terms of causal effects of interventions. One participant envisioned how MetaExplorer could be extended to support evidence from mixed methods:

> *Is there some sort of mapping that I could have between this [forest] plot and the qualitative description of results? What that would show me is why—because here I can see with the forest plot some effect sizes, but I don't know why I am seeing those. If I could click to say, 'Show me the thematic analysis for people who were in this group,' or 'What was any sort of summary of qualitative coding of interviews with people in this group?' That's something I've never seen. (P09).*

We discuss ways to realize this vision in Section 6.2.

One proposed consequence of MetaExplorer not offering enough customizability is that users won't adopt a tool that doesn't cover the same use cases as their hand-rolled procedures, however ad hoc they are *(P04)*. In developing MetaExplorer, implementing a streamlined process required opinionated choices about supported procedures. However, the preference among participants to work in spreadsheets despite their problems implies that users will incur substantial time and labor costs to maintain entrenched workflows and practices. Interoperability with Excel and more support for on-the-fly procedural modifications might promote widespread adoption of tools like MetaExplorer.

Structure as a representation of mental models. We interpret the lack of agreement among participants about standards as evidence that scientists' mental models of research synthesis are highly divergent. MetaExplorer was hit-and-miss in matching these mental models. For some participants, MetaExplorer's evidence extraction process seemed to mirror their preferred perspective—e.g., *"This is how it looks in my brain." (P05)*. For other participants *(e.g., P05, P06, P08, P09)*, the guided tour of MetaExplorer elicited requests for different standards. At the same time, some comments we observed imply that mismatch is often an opportunity. One participant described how many benefits of MetaExplorer come from users updating their mental models to match the tool.

> *A guided tool like this imposes that structure which maybe a person doing a meta-analysis is not thinking about it this way. Maybe they have a different structure in their head which can lead to some discrepancy or tension. But having a tool like this imposes a structure that can be very useful to people doing a meta-analysis if they have not fully set up a structure themselves or just have a vague notion. (P12).*

Our results point overall to the need to provide users with ways to express their mental models so that tools like MetaExplorer can update reciprocally. This fits with participants' *(e.g., P09, P12)* conceptualization of MetaExplorer as a partner in collaboration.

6 DISCUSSION

Through developing and evaluating MetaExplorer with experienced meta-analysts and meta-analysis experts (25 people total), our work advances the design of software for promoting awareness of sources of epistemic uncertainty in meta-analysis that get dredged up during literature review but are seldom propagated to resulting inferences [33]. In particular, MetaExplorer's features for structuring scoping and triage decisions and conducting sensitivity analysis through interactive forest plots were successful design strategies. Our interviews with potential users suggest that MetaExplorer's emphasis on epistemic uncertainty might result in meta-analyses that better characterize heterogeneity in scientific literature, rather than averaging over disparate results.

Our research on MetaExplorer also points to design implications beyond meta-analysis. Data analysis tools writ broadly might benefit users by **guiding documentation of and direct consideration of how to address epistemic uncertainty**, e.g., by systematically resolving or propagating descriptive concerns about data quality or meaning alongside statistical information. Similarly, other interactive systems for data analysis should provide **data management and workflow automations**, since participants claim these are instrumental for accelerating and systematizing documentation and triage of data quality concerns. However, these automations must be configurable (see Section 6.2). A particularly important lesson from designing MetaExplorer is that **epistemic uncertainty should be summarized according to a predefined workflow rather than explored in an open ended fashion** because (1) scientists tend to have principled ways of handling specific sources of epistemic uncertainty *a priori* and (2) open exploration of epistemic uncertainty can promote a form of decision paralysis where the breadth of reasonable interpretations of data is exaggerated by a non-reduced overview. In particular, we expect these principles to be useful for data analysis and communication settings that involve aggregating evidence under hard-to-quantify epistemic uncertainty, such as forecasting applications (e.g., [20, 51]) or combining different forms of evaluative information to assess models (e.g., [30]).

6.1 Limitations

Developing a sufficiently flexible document analysis interface with appropriate scope for a prototype required us to make opinionated decisions, such as tailoring MetaExplorer to handle controlled experiments rather than a wider variety of study designs. While necessary, these scoping decisions limit MetaExplorer to reviews that terminate in meta-analysis, which is not appropriate when available evidence does not support causal inferences or the user wants another form of evidence summary.

Additionally, a more formal evaluation, where users conduct a meta-analysis with MetaExplorer and the quality of their inferences is assessed rigorously, would allow us to say whether design patterns in MetaExplorer will improve the quality of inferences in actual use compared to current practices. However, our experiences suggest this may be hard to realize in practice due to (1) the difficulty of benchmarking user performance when the core tasks in MetaExplorer involve seemingly "subjective" contextually-dependent judgments, and (2) the challenges of recruiting meta-analysis experts to use a tool for the extended time period that meta-analysis tends to require. We opt for guided tour interviews because we seek *holistic* feedback on MetaExplorer, and given the many tasks involved in scientific review and meta-analysis, other methods we considered (e.g., think-aloud, case studies) would have required more time than our participants could offer.

6.2 Future work

Our interviews surfaced opportunities for future work extending a system like MetaExplorer for **supporting collaborative document analysis**, such as by adding functionality for assigning individual users to review specific documents and resolving disagreements between reviewers. Review assignment could be handled in MetaExplorer's review management view using a tagging system to request an individual's attention on a document, and using personalized progress indicators and to-do lists to guide each user's attention. Disagreements between reviewers could be resolved in a tabular interface similar to the triage view, generated automatically from a database but customizable to subsets of questions, that would show disagreements across independent reviews of the same document. These refinements would make the social aspects of analysis decisions explicit in MetaExplorer, enabling users to calculate disagreement statistics and to better resolve ambiguity about the specific statistics they need to extract for meta-analysis.

Future work might also **add qualitative results to MetaExplorer** (e.g., thematic analysis), relaxing the assumptions that MetaExplorer makes about what should be considered evidence and giving users more flexibility to define appropriate standards for their review. Doing so requires changes to the evidence extraction interface, the triage tables, and the forest plot. During evidence extraction, users need a way to select which questions are mandatory to answer in order for the review to be marked complete. Users also need to be able to add custom questions that are tailored for the specific research design of the study. During triage, users need ways to search, sort, and filter study results in order to more easily cluster studies according to what and how they measure. Participants also suggested an overview visualization of current study groupings and the ability to add or remove columns from the default triage table layout. These changes would make it more feasible to organize qualitative evidence across studies, and could also improve the triage process for quantitative evidence at the scale of 100 studies.

Summarizing qualitative evidence would also require incorporating additional contextualizing information into the MetaExplorer visualization. For example, we might add word clouds or additional annotations to summarize coding schemes from qualitative analysis, perhaps highlighting common codes or themes across analyses. This would improve the affordances of the MetaExplorer visualization for propagating sources of epistemic uncertainty, providing a more flexible mapping between sampling distributions and qualitative claims.

Progressive form customization, or just-in-time form branching, is a promising way to support greater flexibility in the evidence extraction and quality assessment forms. MetaExplorer already does some of this—e.g., using questions about study design to filter subsequent questions about what was reported, which in turn determine the layout of the evidence table. With MetaExplorer, we demonstrate how this design pattern can be used to cover substantial methodological variation within interventional experiments. However, we could extend this design pattern by using *templates* to capture important considerations and contingencies under different kinds of research designs. These templates would represent questions and contingencies among them, which users could select from on the fly during evidence extraction. We could also expose the template formalism to users through an editing interface, enabling them to author templates by composing new questions or combining questions from existing templates. This would support different standards in quality assessment—e.g., users could opt for the Cochrane Risk of Bias Assessment [26] if this is meaningful to their research community. It would also more formally demarcate the roles of expert users (authoring templates) and novice users (following templates) in collaboration, a use case that was not a design goal for MetaExplorer but which participants described as a pain point in current practice.

Finally, a handful of participants *(P06, P07, P08, P11)* requested ways to **flexibly specify models** of their choice within MetaExplorer. Adding a model editing dialogue to the MetaExplorer visualization where users could modify the default meta-analytic model in a code block would support this.

7 CONCLUSION

We present MetaExplorer, a software prototype for representing and reasoning with uncertainty about scientific literature during meta-analysis. MetaExplorer is a proof-of-concept exploring new design patterns for propagating unquantified epistemic uncertainty in end-to-end quantitative data analysis. By prototyping these design patterns in MetaExplorer and eliciting feedback from knowledgeable meta-analysts, we find challenges and opportunities around (1) supporting the documentation, collaboration, and modeling efforts required to resolve sources of epistemic uncertainty and (2) developing widely-applicable yet sufficiently flexible standards around data quality. MetaExplorer opens the door to new ways to design data analysis software with an emphasis on how unquantified uncertainty informs statistical inference.

ACKNOWLEDGMENTS

This work was funded by US Navy STTR Contract N6833517-C-0410 in partnership with Stottler Henke Assoc, and NSF award #1749266.

REFERENCES

[1] Elsevier 2022. *Mendeley.* Elsevier. https://www.mendeley.com/reference-management/reference-manager

[2] Thom Baguley. 2009. Standardized or simple effect size: What should be reported? *British journal of psychology* 100, 3 (2009), 603–617.

[3] Sara Balduzzi, Gerta Rücker, and Guido Schwarzer. 2019. How to perform a meta-analysis with R: a practical tutorial. *Evidence-Based Mental Health* 22 (2019), 153–160.

[4] Leon Bax, Ly Mee Yu, Noriaki Ikeda, Harukazu Tsuruta, and Karel G.M. Moons. 2006. Development and validation of MIX: Comprehensive free software for meta-analysis of causal research data. *BMC Medical Research Methodology* 6, 50 (2006), 1–11. https://doi.org/10.1186/1471-2288-6-50

[5] Betsy Jane Becker. 2017. Improving the design and use of meta-analyses of career interventions. *Integrating theory, research, and practice in vocational psychology: Current status and future directions* (2017), 95.

[6] Sarah Belia, Fiona Fidler, Jennifer Williams, and Geoff Cumming. 2005. Researchers Misunderstand Confidence Intervals and Standard Error Bars. *Psychological Methods* 10, 4 (2005), 389–396. https://doi.org/10.1037/1082-989X.10.4.389

[7] Susanne Bodker and Kaj Gronbaek. 1991. Cooperative prototyping: Users and designers in mutual activity. *International Journal of Man-Machine Studies* 34 (1991), 453–478. http://www.sciencedirect.com/science/article/pii/002073739190030B

[8] M Borenstein, L Hedges, J Higgins, and H Rothstein. 2005. Comprehensive Meta-Analysis 2. Engelwood, NJ, Biostat.

[9] Mike Bostock. 2012. D3.js - Data-Driven Documents. http://d3js.org/

[10] Nadia Boukhelifa, Marc-Emmanuel Perrin, Samuel Huron, and James Eagan. 2017. How Data Workers Cope with Uncertainty : A Task Characterisation Study. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2017). https://doi.org/10.1145/3025453.3025738

[11] Christopher J Bryan, Elizabeth Tipton, and David S Yeager. 2021. Behavioural science is unlikely to change the world without a heterogeneity revolution. *Nature human behaviour* 5, 8 (2021), 980–989.

[12] Spencer C. Castro, P. Samuel Quinan, Helia Hosseinpour, and Lace Padilla. 2022. Examining Effort in 1D Uncertainty Communication Using Individual Differences in Working Memory and NASA-TLX. *IEEE Transactions on Visualization and Computer Graphics* 28, 1 (2022), 411–421. https://doi.org/10.1109/TVCG.2021.3114803

[13] Robert Coe. 2002. It's the Effect Size, Stupid. *British Educational Research Association Annual Conference*, 1–18.

[14] The Cochrane Collaboration. 2014. Review Manager (RevMan) 5.1.0. Copenhagen: The Nordic Cochrane Centre.

[15] Harris Cooper, Larry V Hedges, and Jeffrey C Valentine. 2019. *The handbook of research synthesis and meta-analysis.* Russell Sage Foundation.

[16] Michael Correll and Michael Gleicher. 2014. Error bars considered harmful: Exploring alternate encodings for mean and error. *IEEE Transactions on Visualization and Computer Graphics* 20, 12 (2014), 2142–2151. https://doi.org/10.1109/TVCG.2014.2346298

[17] Geoff Cumming. 2014. The New Statistics: Why and How. *Psychological Science* 25, 1 (2014), 7–29. https://doi.org/10.1177/0956797613504966

[18] Michael Fernandes, Logan Walls, Sean A Munson, Jessica Hullman, and Matthew Kay. 2018. Uncertainty Displays Using Quantile Dotplots or CDFs Improve Transit Decision-Making. In *ACM Transactions on Computer-Human Interaction.* Montreal. https://doi.org/10.1145/3173574.3173718

[19] Roy Rosenzweig Center for History and New Media. 2016. *Zotero.* https://www.zotero.org/download

[20] Andrew Gelman, Jessica Hullman, Christopher Wlezien, and George Elliott Morris. 2020. Information, incentives, and goals in election forecasts. *Judgment and Decision Making* 15, 5 (2020), 863.

[21] Gerd Gigerenzer and Ulrich Hoffrage. 1995. How to Improve Bayesian Reasoning Without Instruction: Frequency Formats. *Psychological review* 102, 4 (1995), 684–704. https://doi.org/10.1037/0033-295X.102.4.684

[22] S. Greenland. 2001. On the bias produced by quality scores in meta-analysis, and a hierarchical view of proposed solutions. *Biostatistics* 2, 4 (2001), 463–471. https://doi.org/10.1093/biostatistics/2.4.463

[23] Gordon H Guyatt, Andrew D Oxman, Gunn E Vist, Regina Kunz, Yngve Falck-Ytter, Pablo Alonso-Coello, and Holger J Schünemann. 2008. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ* 336, 7650 (2008), 924–926. https://doi.org/10.1136/bmj.39489.470347.AD arXiv:https://www.bmj.com/content/336/7650/924.full.pdf

[24] JPT Higgins and S Green (Eds.). 2011. *Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0.* The Cochrane Collaboration. www.cochrane-handbook.org.

[25] Julian P.T. Higgins and Simon G. Thompson. 2002. Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine* 21, 11 (2002), 1539–1558. https://doi.org/10.1002/sim.1186

[26] Julian P T Higgins, Douglas G Altman, Peter C Gøtzsche, Peter Jüni, David Moher, Andrew D Oxman, Jelena Savović, Kenneth F Schulz, Laura Weeks, and Jonathan A C Sterne. 2011. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *BMJ* 343 (2011). https://doi.org/10.1136/bmj.d5928 arXiv:https://www.bmj.com/content/343/bmj.d5928.full.pdf

[27] Rink Hoekstra, Richard D. Morey, Jeffrey N. Rouder, and Eric-Jan Wagenmakers. 2014. Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review* 21, 5 (2014), 1157–1164. https://doi.org/10.3758/s13423-013-0572-3

[28] U Hoffrage and G Gigerenzer. 1998. Using natural frequencies to improve diagnostic inferences. *Academic medicine: Journal of the Association of American Medical Colleges* 73, 5 (1998), 538–540. https://doi.org/10.1097/00001888-199805000-00024

[29] Jessica Hullman, Paul Resnick, and Eytan Adar. 2015. Hypothetical Outcome Plots Outperform Error Bars and Violin Plots for Inferences about Reliability of Variable Ordering. *PloS one* 10, 11 (2015), e0142444. https://doi.org/10.1371/journal.pone.0142444

[30] Ben Hutchinson, Negar Rostamzadeh, Christina Greer, Katherine Heller, and Vinodkumar Prabhakaran. 2022. Evaluation Gaps in Machine Learning Practice. *arXiv preprint arXiv:2205.05256* (2022).

[31] Veritas Health Innovation. 2022. *Covidence systematic review software.* https://www.covidence.org

[32] Eunice Jun, Audrey Seo, Jeffrey Heer, and René Just. 2022. Tisane: Authoring Statistical Models via Formal Reasoning from Conceptual and Data Relationships. In *ACM Human Factors in Computing Systems (CHI).* http://idl.cs.washington.edu/papers/tisane

[33] Alex Kale, Matthew Kay, and Jessica Hullman. 2019. Decision-Making Under Uncertainty in Research Synthesis: Designing for the Garden of Forking Paths. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems.* ACM.

[34] Alex Kale, Matthew Kay, and Jessica Hullman. 2021. Visual Reasoning Strategies for Effect Size Judgments and Decisions. *IEEE Trans. Vis. Comput. Graph.* 27, 2 (2021), 272–282. https://doi.org/10.1109/TVCG.2020.3030335

[35] Alex Kale, Francis Nguyen, Matthew Kay, and Jessica Hullman. 2019. Hypothetical Outcome Plots Help Untrained Observers Judge Trends in Ambiguous Data. *IEEE Trans. Visualization & Comp. Graphics (Proc. InfoVis)* (2019). http://idl.cs.washington.edu/papers/hops-trends

[36] Matthew Kay, Tara Kola, Jessica R Hullman, and Sean A Munson. 2016. When (ish) is My Bus? User-centered Visualizations of Uncertainty in Everyday, Mobile Predictive Systems. In *Proceedings of the 2016 ACM annual conference on Human Factors in Computing Systems.*

[37] M.W. Lipsey and D.B. Wilson. 2001. *Practical Meta-Analysis.* SAGE Publications.

[38] Yang Liu, Tim Althoff, and Jeffrey Heer. 2020. Paths Explored, Paths Omitted, Paths Obscured: Decision Points & Selective Reporting in End-to-End Data Analysis. In *ACM Human Factors in Computing Systems (CHI).* http://idl.cs.washington.edu/papers/analysis-decision-points

[39] Yang Liu, Alex Kale, Tim Althoff, and Jeffrey Heer. 2021. Boba: Authoring and Visualizing Multiverse Analyses. *IEEE Trans. Visualization & Comp. Graphics (Proc. VAST)* (2021). http://idl.cs.washington.edu/papers/boba

[40] C.KL. Lo, D. Mertz, and M. Loeb. 2014. Newcastle-Ottawa Scale: comparing reviewers' to authors' assessments. *BMC Med Res Methodol* 45 (2014). Issue 14. https://doi.org/10.1186/1471-2288-14-45

[41] Helen E. Longino. 1990. *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry.* Princeton University Press.

[42] Charles F Manski. 2018. Communicating uncertainty in policy analysis. *Proceedings of the National Academy of Sciences of the United States of America* 2018 (2018), 201722389. https://doi.org/10.1073/pnas.1722389115

[43] Charles F Manski. 2019. *Meta-Analysis for Medical Decisions.* Working Paper 25504. National Bureau of Economic Research. https://doi.org/10.3386/w25504

[44] Nina McCurdy, Julie Gerdes, and Miriah Meyer. 2019. A Framework for Externalizing Implicit Error Using Visualization. *IEEE Transactions on Visualization and Computer Graphics (InfoVis)* 25, 1 (2019), 925–935. https://doi.org/10.1109/TVCG.2018.2864913

[45] Microsoft Corporation. 2019. *Microsoft Excel.* https://office.microsoft.com/excel

[46] D Moher, DJ Cook, Alejandro R Jadad, P Tugwell, M Moher, A Jones, B Pham, and TP Klassen. 1999. Assessing the quality of reports of randomised trials: implications for the conduct of meta-analyses. *Health Technology Assessment (Winchester, England)* 3, 12 (1999), i–98.

[47] Wendy Moyle, Marie Cooke, Elizabeth Beattie, Cindy Jones, Barbara Klein, Glenda Cook, and Chrystal Gray. 2013. Exploring the Effect of Companion Robots on Emotional Expressions in Older Adults with Dementia: A Pilot Randomized Controlled Trial. *Journal of Gerontological Nursing* 39 (2013). Issue 5.

[48] Brian A. Nosek, Charles R. Ebersole, Alexander C. DeHaven, and David T. Mellor. 2018. The preregistration revolution. *Proceedings of the National Academy of Sciences* 115, 11 (2018), 2600–2606. https://doi.org/10.1073/pnas.1708274114 arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.1708274114

[49] Mourad Ouzzani, Hossam Hammady, Zbys Fedorowicz, and Ahmed Elmagarmid. 2016. Rayyan - a web and mobile app for systematic reviews. https://doi.org/10.1186/s13643-016-0384-4

[50] R. K. Owen, N. Bradbury, Y. Xin, N. Cooper, and A. Sutton. 2019. MetaInsight: An interactive web-based tool for analyzing, interrogating, and visualizing network meta-analyses using R-shiny and netmeta. *Research synthesis methods* 10, 4 (2019), 569–581. https://doi.org/10.1002/jrsm.1373

[51] Lace MK Padilla, Maia Powell, Matthew Kay, and Jessica Hullman. 2021. Uncertain about uncertainty: How qualitative expressions of forecaster confidence impact decision-making with uncertainty visualizations. *Frontiers in Psychology* 11 (2021), 579267.

[52] Matthew J Page, David Moher, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, Roger Chou, Julie Glanville, Jeremy M Grimshaw, Asbjørn Hróbjartsson, Manoj M Lalu, Tianjing Li, Elizabeth W Loder, Evan Mayo-Wilson, Steve McDonald, Luke A McGuinness, Lesley A Stewart, James Thomas, Andrea C Tricco, Vivian A Welch, Penny Whiting, and Joanne E McKenzie. 2021. PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. *BMJ* 372 (2021). https://doi.org/10.1136/bmj.n160 arXiv:https://www.bmj.com/content/372/bmj.n160.full.pdf

[53] Evidence Partners. 2021. *DistillerSR.* https://www.evidencepartners.com

[54] Lihui Pu, Wendy Moyle, Cindy Jones, and Michael Todorovic. 2018. The Effectiveness of Social Robots for Older Adults: A Systematic Review and Meta-Analysis of Randomized Controlled Studies. *The Gerontologist* 59, 1 (06 2018), e37–e51. https://doi.org/10.1093/geront/gny046 arXiv:https://academic.oup.com/gerontologist/article-pdf/59/1/e37/27456572/gny046.pdf

[55] John J. Riva, Keshena Malik, Stephen J. Burnie, Andrea R Endicott, and Jason Walter Busse. 2012. What is your research question? An introduction to the PICOT format for clinicians. *The Journal of the Canadian Chiropractic Association* 56 3 (2012), 167–171.

[56] M.S. Rosenberg, D.C. Adams, and J. Gurevitch. 1997. MetaWin: Statistical Software for Meta-Analysis with Resampling Tests. (1997).

[57] Dominik Sacha, Hansi Senaratne, Bum Chul Kwon, Geoffrey Ellis, and Daniel A. Keim. 2016. The Role of Uncertainty, Awareness, and Trust in Visual Analytics. *IEEE Transactions on Visualization and Computer Graphics* 22 (2016). Issue 1. https://doi.org/10.1109/TVCG.2015.2467591

[58] Donald A. Schon. 1987. *Educating the Reflective Practitioner.* John Wiley & Sons Inc., San Francisco, CA.

[59] William R. Shadish, Thomas D. Cook, and Donald T. Campbell. 2001. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference.* Houghton Mifflin.

[60] Cláudio T. Silva, Erik Anderson, Emanuele Santos, and Juliana Freire. 2011. Using VisTrails and provenance for teaching scientific visualization. *Computer Graphics Forum* 30 (2011), 75–84. Issue 1. https://doi.org/10.1111/j.1467-8659.2010.01830.x

[61] Uri Simonsohn, Joseph P. Simmons, and Leif D. Nelson. 2015. Specification Curve: Descriptive and Inferential Statistics on All Reasonable Specifications. *SSRN* (Nov 2015). https://doi.org/10.2139/ssrn.2694998

[62] Emre Soyer and Robin M. Hogarth. 2012. The illusion of predictability: How regression statistics mislead experts. *International Journal of Forecasting* 28, 3 (2012), 712–714. https://doi.org/10.1016/j.ijforecast.2012.02.004

[63] Sara Steegen, Francis Tuerlinckx, Andrew Gelman, and Wolf Vanpaemel. 2016. Increasing Transparency Through a Multiverse Analysis. *Perspectives on Psychological Science* 11, 5 (2016), 702–712. https://doi.org/10.1177/1745691616658637

[64] Barnabas Szaszi, Anthony Higney, Aaron Charlton, Andrew Gelman, Ignazio Ziano, Balazs Aczel, Daniel G Goldstein, David S Yeager, and Elizabeth Tipton. 2022. No reason to expect large and consistent effects of nudge interventions. *Proceedings of the National Academy of Sciences* 119, 31 (2022), e2200732119.

[65] Barry N. Taylor and Chris E. Kuyatt. 1994. *Guidelines for evaluating and expressing the uncertainty of NIST measurement results.* Technical Report.

[66] James Thomas, Sergio Graziosi, Jeff Brunton, Z. Ghouze, Patrick O'Driscoll, M. Bond, and A. Koryakina. 2022. EPPI-Reviewer: advanced software for systematic reviews, maps and evidence synthesis. (2022).

[67] Rebecca M. Turner, David J. Spiegelhalter, Gordon C.S. Smith, and Simon G. Thompson. 2009. Bias modelling in evidence synthesis. *Journal of the Royal Statistical Society. Series A: Statistics in Society* 172, 1 (2009), 21–47. https://doi.org/10.1111/j.1467-985X.2008.00547.x

[68] Wolfgang Viechtbauer. 2010. Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software* 36, 3 (2010), 1–48. https://doi.org/10.18637/jss.v036.i03

[69] Chat Wacharamanotham, Krishna Subramanian, Sarah Theres Völkel, and Jan Borchers. 2015. Statsplorer: Guiding Novices in Statistical Analysis. In *CHI '15: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea). New York, NY, USA, 2693–2702. https://doi.org/10.1145/2702123.2702347

[70] Byron C. Wallace, Issa J. Dahabreh, Thomas A. Trikalinos, Joseph Lau, Paul Trow, and Christopher H. Schmid. 2012. Closing the Gap between Methodologists and End-Users: R as a Computational Back-End. *Journal of Statistical Software* 5 (2012). Issue 49.

[71] Jelte M. Wicherts, Coosje L.S. Veldkamp, Hilde E.M. Augusteijn, Marjan Bakker, Robbie C.M. van Aert, and Marcel A.L.M. van Assen. 2016. Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid P-hacking. *Frontiers in Psychology* 7, Nov (2016). https://doi.org/10.3389/fpsyg.2016.01832

[72] J. Zhou, M. Glueck, S. Breslav, F. Chevalier, and A. Khan. 2017. Annotation graphs: a graph-based visualization for meta-analysis of data based on user-authored annotations. *IEEE transactions on visualization and computer graphics* 23, 1 (2017), 261–270.