Evaluation of Reproducible Research Practices in Oncology Systematic Reviews With Meta-analyses Referenced by National Comprehensive Cancer Network Guidelines

Cole Wayant; Matthew J. Page, PhD; Matt Vassar, PhD

**IMPORTANCE** Reproducible research practices are essential to biomedical research because these practices promote trustworthy evidence. In systematic reviews and meta-analyses, reproducible research practices ensure that summary effects used to guide patient care are stable and trustworthy.

**OBJECTIVE** To evaluate the reproducibility in theory of meta-analyses in oncology systematic reviews cited by the 49 National Comprehensive Cancer Network (NCCN) guidelines for the treatment of cancer by site and evaluate whether Cochrane reviews or systematic reviews that report adherence to Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines use more reproducible research practices.

**DESIGN, SETTING, AND PARTICIPANTS** A cross-sectional investigation of all systematic reviews with at least 1 meta-analysis and at least 1 included randomized clinical trial (RCT) that are cited by NCCN guidelines for treatment of cancer by site. We scanned the reference list of all NCCN guidelines (n = 49) for potential systematic reviews and meta-analyses. All retrieved studies were screened, and data were extracted, independently and in duplicate. The analysis was carried out between May 6, 2018, and January 28, 2019.

**MAIN OUTCOMES AND MEASURES** The frequency of reproducible research practices, defined as (1) effect estimate and measure of precision (eg, hazard ratio with 95% confidence interval); (2) clear list of studies included for each analysis; and (3) for subgroup and sensitivity analyses, it must be clear which studies were included in each group or level.

**RESULTS** We identified 1124 potential systematic reviews, and 154 meta-analyses comprising 3696 meta-analytic effect size estimates were included. Only 2375 of the 3696 meta-analytic estimates (64.3%), including subgroup and sensitivity analyses, were reproducible in theory. Forest plots appear to improve the reproducibility of meta-analyses. All meta-analytic estimates were reproducible in theory in 100 systematic reviews (64.9%), and in 15 systematic reviews (9.7%), no meta-analytic estimates could potentially be reproduced. Data were said to be imputed in 29 meta-analyses, but none specified which data. Only 1 meta-analysis included a link to an online data set.

**CONCLUSIONS AND RELEVANCE** More reproducible research practices are needed in oncology meta-analyses, as suggested by those that are cited by the NCCN. Reporting meta-analyses in forest plots and requirements for full data sharing are recommended.
Concerns are growing about the reproducibility of biomedical research. Many of these concerns stem from research practices that lack transparency, including poor reporting of study methodology and failing to make study data publicly available. As a result, efforts to reproduce biomedically robust research findings have been thwarted. Most efforts to reproduce research findings have been dedicated to primary studies, such as randomized clinical trials, and little effort has been dedicated to reproduce higher levels of evidence, such as systematic reviews. The first studies to holistically evaluate the reproducibility of systematic reviews and meta-analyses in the biomedically robust literature found that authors frequently fail to use reproducible research practices. However, only a small proportion of the systematic reviews with meta-analyses evaluated in previous investigations were for oncology interventions, leaving unanswered questions for researchers in this field, oncologists, and policy makers.

For this investigation of the reproducibility of oncology systematic reviews, we identified systematic reviews cited in National Comprehensive Cancer Network (NCCN) clinical practice guidelines. The NCCN set of guidelines is one of many available to oncologists; however, a survey of oncologists showed that NCCN guidelines were more likely to influence clinical practice than other popular oncology guidelines. Further, NCCN guidelines cover all blood and solid cancers, thus making them ideal for a broad investigation such as this. The primary objective of this investigation was to evaluate the reproducibility of theory of meta-analyses in oncology systematic reviews cited by the 49 NCCN guidelines for the treatment of cancer by site. The secondary objective was to evaluate whether Cochrane reviews or systematic reviews that report adherence to Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines use more reproducible research practices.

Methods

The protocol for this investigation is publicly available via the Open Science Framework. We defined a systematic review according to the PRISMA for protocols definition: articles that explicitly stated methods to identify studies (ie, a search strategy), explicitly stated methods of study selection (eg, eligibility criteria and selection process), and explicitly described methods of synthesis (or other type of summary). Because NCCN guidelines update regularly throughout each year, all guidelines were manually downloaded as PDFs on May 6, 2018, to avoid citations being added to the guideline during the course of our investigation. To identify systematic reviews we manually screened the reference lists and Discussion narratives of all NCCN clinical practice guidelines for the treatment of cancer. We extracted all references with “systematic review,” “meta-analysis,” “meta-analysis,” and any references without the keywords in the title that were discussed as a systematic review or meta-analysis by guideline authors. We also extracted any cited references that were published in the Cochrane Database for Systematic Reviews. All extracted references were added to a PubMed collection and exported to Rayyan for title and abstract screening.

We screened articles using the liberal acceleration method whereby 1 author (C.W.) was required to mark a record for inclusion and 2 authors (C.W. and M.J.P.) were required to mark a record for exclusion. Next, 2 authors (C.W. and M.J.P.) screened the full text of potentially relevant articles for inclusion. Key inclusion criteria were systematic reviews published in 2011 or later with at least 1 meta-analysis that included at least 1 randomized clinical trial. We chose to include only systematic reviews published after 2011 to allow time for uptake of the 2009 PRISMA Statement. Thus, all included systematic reviews were accountable to currently accepted reporting quality standards. The systematic reviews of individual patient data or of primary studies other than randomized clinical trials, network meta-analyses, and pooled analyses of randomized clinical trials were excluded.

To extract data for this study we developed a pilot-tested Google Form based on the extraction form used in a similar, previous study. Extracted data items were related to the number of meta-analyses reported, reporting of summary statistics for each individual study, use of fixed-effect vs random-effect models, interpretation of tests of heterogeneity and small-study effects, and types of subgroup and sensitivity analyses performed. We extracted data for all meta-analyses, but certain items were dedicated to the index meta-analysis, which we defined as the primary meta-analysis for the primary end point. If there was no primary end point mentioned, we used the first reported meta-analysis as the index meta-analysis and inferred the primary end point from there. We counted meta-analyses by summing the number of summary effects in forest plots, written narrative, and supplemental appendices. Duplicate meta-analytic effects were only counted once. We counted subgroup effects that were derived from an analysis of at least 2 studies, as well as the overall summary effect that synthesized all subgroup effects. We only counted sensitivity analyses that were expressly described with a summary effect in the article or the supplemental material.

To be considered reproducible in theory an analysis must have 3 elements: (1) effect estimate and measure of precision (eg, hazard ratio with 95% confidence interval); (2) clear list of studies included for each analysis; and (3) for subgroup and...
Figure 1. PRISMA Diagram

1124 Records retrieved from NCCN guidelines
841 Records imported to Rayyan\textsuperscript{12} for screening
283 Duplicates removed

687 Excluded
283 SRs with no included RCTs
165 SRs with no meta-analyses
94 Published before PRISMA in 2011
64 IPD meta-analyses
14 Retrospective reviews
13 Clinical practice guidelines
12 Pooled analyses
11 Network meta-analyses
7 Observational studies
6 RCTs
5 No full text
4 Narrative reviews
4 SRs of diagnostic accuracy
3 Letters to the editor
1 Case series
1 Cross-sectional review

154 SRs with at least 1 meta-analysis of RCTs included

IPD indicates individual patient data; NCCN, National Comprehensive Cancer Network; PRISMA, Preferred Reporting Items for Systematic Reviews and Meta-Analyses; RCT, randomized clinical trial; SR, systematic review.

sensitivity analyses, it must be clear which studies were included in each group or level.

Data from all systematic reviews were extracted by C.W. A random sample of 15% of the included systematic reviews was extracted in duplicate by M.J.P. and M.V. adjudicated discrepancies in the double-extracted 15% sample. Any item that had at least 1 discrepancy was reviewed a second time in the 85% of other studies by C.W. A complete list of items with a discrepancy are available, along with our protocol and data, via the Open Science Framework.\textsuperscript{9}

Summary statistics and measures of central tendency (eg, median with interquartile range [IQR]) were calculated using Microsoft Excel (2016, Microsoft). We planned to use STATA statistical software (version 15.1, STATA Corp) to calculate unadjusted risk ratios (uRR) and 95% confidence intervals (CIs) for the comparisons between Cochrane and non-Cochrane systematic reviews, and between systematic reviews self-reporting the use of PRISMA vs not, but owing to disparate numbers of Cochrane and non-Cochrane systematic reviews, we only reported the comparisons of systematic reviews stratified by PRISMA adherence and year of publication. We conducted sensitivity analyses for meta-analyses presented in figures and for those published as supplementary material to investigate potential factors contributing to reproducibility.

Results

Characteristics

We identified 1124 potential systematic reviews from our survey of the 49 NCCN guidelines for the treatment of cancer by site. Five NCCN guidelines did not cite any systematic reviews. An additional 19 clinical practice guidelines did not have any systematic reviews that met the inclusion criteria. After removing duplicates and screening all articles, 154 systematic reviews with at least 1 meta-analysis were included (Figure 1).\textsuperscript{9} There was high agreement between reviewers (94.0%) for studies extracted in duplicate.

Of the 154 included systematic reviews, 77 (50.0%) were either a Cochrane review or mentioned adherence to PRISMA. Eighteen systematic reviews (11.7%) were Cochrane systematic reviews, and 60 adhered to PRISMA (39.0%). Of the 78 systematic reviews that received funding, public sources (eg, government) were most common (36 [46.2%]). The systematic reviews included a median of 14 (IQR, 7.25-29.75) meta-analytic effect estimates, including those from subgroup and sensitivity analysis. Additional characteristics of our sample are reported in eTable 1 in the Supplement.

Only 88 systematic reviews (57.1%) labeled their primary end point (eTable 2 in the Supplement). Thus, we inferred the primary end point in the remaining 66 systematic reviews from the index (first reported) meta-analysis. Seventy-three (47.4%) primary end points were all-cause mortality. A median of 8 (IQR, 5-12) primary studies with a median 1914 (IQR, 917-3941) patients were included in each index meta-analysis. Seventy-nine index meta-analyses (51.2%) included a subgroup analysis and 54 included a sensitivity analysis (35.1%).

Reproducible Research Practices: Overall

There were a total of 3696 meta-analytic effect estimates, including subgroup and sensitivity analyses in the 154 meta-analyses, but only 2375 (64.3%) were reproducible in theory. All meta-analyses were reproducible in theory in 100 meta-analyses (64.9%), and in 139 meta-analyses (90.3%) there was at least 1 meta-analysis that could potentially be reproduced. Summary statistics (eg, event rates) for studies included in the index meta-analysis were reported in 107 meta-analyses (69.5%), but only 39 (25.3%) mentioned whether or not missing data were imputed and included in the index meta-analysis. Missing data were reported to have been imputed in 29 of these 39 meta-analyses, but it was not clear which exact data points were imputed in all 29. Similarly, only 29 meta-analyses mentioned whether unpublished data were retrieved from primary study authors, with 17 affirming that authors were contacted. However, only 3 of 17 (17.6%) were clear about which data were retrieved.

Eighty-seven meta-analyses (56.5%) generated funnel plots to assess for publication bias, but only 49 of 87 (56.3%) presented the funnel plot in the meta-analysis or supplemental appendix. In 28 of 87 meta-analyses (32.2%), the number of studies included in the funnel plot was unclear. Only 62 meta-analyses cited the guide they used to interpret their $I^2$ statistic, with the most common guide being by Higgins et al.\textsuperscript{13} Sixty-one authors (39.6%) decided between a random- or fixed-effects model based on the statistical heterogeneity of the included studies, but 31 of 61 (50.8%) did not report the amount of heterogeneity necessary to use a random-effects model.
Based on meta-analyses, text analysis, which was likely confounded by forest plot interpretation, especially supplemental versus main-text meta-analyses, both sensitivity analyses were unadjusted and should be interpreted with caution, especially the supplemental versus main-text analysis, which was likely confounded by forest plot-based meta-analyses.

Both sensitivity analyses were unadjusted and should be implemented for primary analyses, but far less so for secondary, subgroup, and sensitivity analyses. Moreover, figure-based (eg, forest plot) meta-analyses were far more reproducible than other meta-analyses, and our sensitivity analysis showed that reproducible research practices are commonly implemented for primary analyses, but far less so for secondary, subgroup, and sensitivity analyses. Moreover, figure-based (eg, forest plot) meta-analyses were far more reproducible than other meta-analyses, and our sensitivity analysis shows that the

### Table. Reproducible Research Practices of Systematic Reviews With Meta-analyses That Underpin the National Comprehensive Cancer Network Clinical Practice Guidelines for the Treatment of Cancer by Site

<table>
<thead>
<tr>
<th>Reproducible Research Practice</th>
<th>All (n = 154)</th>
<th>PRISMA (n = 60)</th>
<th>Non-PRISMA (n = 94)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reported the data needed to recreate all meta-analytic effect estimates in the SR</td>
<td>100 (64.9)</td>
<td>36 (60.0)</td>
<td>64 (68.1)</td>
</tr>
<tr>
<td>Reported the data needed to recreate the index meta-analytic effect estimate</td>
<td>140 (90.9)</td>
<td>58 (96.7)</td>
<td>82 (87.2)</td>
</tr>
<tr>
<td>Reported summary statistics for each individual study in the index meta-analysis</td>
<td>107 (69.5)</td>
<td>42 (70.0)</td>
<td>65 (69.1)</td>
</tr>
<tr>
<td>Reported effect estimates and measures of precision for each individual study in the index meta-analysis</td>
<td>140 (90.9)</td>
<td>58 (96.7)</td>
<td>82 (87.2)</td>
</tr>
<tr>
<td>Reported that some data in the index meta-analysis had been imputed</td>
<td>39 (25.3)</td>
<td>14 (23.2)</td>
<td>25 (26.6)</td>
</tr>
<tr>
<td>Clear which data were imputed and how</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Reported that some data in the index meta-analysis had been obtained from the study author/sponsor</td>
<td>17 (11.0)</td>
<td>7 (11.7)</td>
<td>10 (10.6)</td>
</tr>
<tr>
<td>Clear which data were obtained</td>
<td>3 (1.9)</td>
<td>0</td>
<td>3 (3.2)</td>
</tr>
<tr>
<td>Reported (or inferred) the type of random-effects method used for the index meta-analysis, No./No. (%)</td>
<td>78/91 (85.7)</td>
<td>22/43 (51.2)</td>
<td>24/48 (50.0)</td>
</tr>
<tr>
<td>Reported the data needed to recreate each subgroup analysis for the index meta-analysis, No./No. (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>For all subgroup analyses</td>
<td>51/79 (64.6)</td>
<td>26/40 (65.0)</td>
<td>25/39 (64.1)</td>
</tr>
<tr>
<td>For some subgroup analyses</td>
<td>1/79 (1.3)</td>
<td>0</td>
<td>1/39 (2.6)</td>
</tr>
<tr>
<td>Not for any subgroup analysis</td>
<td>27/79 (34.2)</td>
<td>14/40 (35.0)</td>
<td>13/39 (33.3)</td>
</tr>
<tr>
<td>Reported the data needed to recreate each sensitivity analysis for the index meta-analysis, No./No. (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>For all sensitivity analyses</td>
<td>34/54 (63.0)</td>
<td>12/23 (52.2)</td>
<td>22/31 (71.0)</td>
</tr>
<tr>
<td>For some sensitivity analyses</td>
<td>2/54 (3.7)</td>
<td>2/23 (8.7)</td>
<td>0</td>
</tr>
<tr>
<td>Not for any sensitivity analysis</td>
<td>18/54 (33.3)</td>
<td>9/23 (39.1)</td>
<td>9/31 (29.0)</td>
</tr>
<tr>
<td>Mention of access to data sets and statistical analysis code used to perform analyses</td>
<td>1 (0.6)</td>
<td>0</td>
<td>1 (1.1)</td>
</tr>
</tbody>
</table>

**Abbreviations:** PRISMA, Preferred Reporting Items for Systematic Reviews and Meta-Analyses; SR, systematic review.

Random-effects models were used for 91 of 154 index meta-analyses (59.1%), but specific information about the between-trial variance estimator (eg, DerSimonian and Laird\(^4\)) were not reported in 45 of 91 (49.5%). Subgroup analyses were included in 79 of 154 index meta-analyses (51.3%), but only 51 of 79 (64.6%) were fully reproducible in theory. Of the 54 sensitivity analyses that accompanied index meta-analyses, only 34 (63.0%) were fully reproducible in theory. Only 1 meta-analysis—a Cochrane meta-analysis—included a link to an online data set.

When considering only the 2341 of 3696 meta-analytic estimates that were presented on forest plots, we determined that 2195 of 2341 (93.7%) were reproducible in theory because they included numerical point estimates (or event rates conducive to calculating point estimates) and a list of included studies. Compared with meta-analyses not published in Figures (180/1535), forest plot-based meta-analyses were more often reproducible in theory (uRR, 8.4; 95% CI, 7.2-9.7). When considering only meta-analytic estimates published as supplemental material, we determined that 368 of 642 (57.3%) were reproducible in theory. Compared with main-text meta-analyses (2007/3054), supplemental meta-analyses were less often reproducible in theory (uRR, 0.74; 95% CI, 0.65-0.86). Both sensitivity analyses were unadjusted and should be interpreted with caution, especially the supplemental versus main-text analysis, which was likely confounded by forest plot-based meta-analyses.

### Stratified Analyses of Reproducible Research Practices

We limited our analysis of Cochrane and non-Cochrane reviews to summary statistics owing to large differences in group sample sizes. One of 18 Cochrane meta-analyses (5.6%) and 59 of 136 non-Cochrane meta-analyses (43.4%) stated that they adhered to PRISMA guidelines. In 16 of 18 Cochrane meta-analyses (88.9%), all included meta-analyses were reproducible in theory compared with 85 of 136 non-Cochrane meta-analyses (62.5%). Regarding sensitivity and subgroup analyses, all were reproducible in theory in Cochrane meta-analyses. In non-Cochrane meta-analyses only 29 of 48 (60.4%) with sensitivity analyses and 49 of 77 (63.6%) with subgroup analyses provided enough information to make these analyses reproducible. All data for comparisons between meta-analyses that did and did not mention PRISMA are in the Table and Figure 2. Data for our analysis by year of publication are shown in Figure 3.

### Discussion

The results of our investigation of oncology meta-analyses demonstrate that reproducible research practices are commonly implemented for primary analyses, but far less so for secondary, subgroup, and sensitivity analyses. Moreover, figure-based (eg, forest plot) meta-analyses were far more reproducible than other meta-analyses, and our sensitivity analysis shows that the
Evaluation of Reproducible Research Practices in Oncology Systematic Reviews With Meta-analyses

Our findings are comparable to those from a previous, similar study that examined the reproducible research practices of a cross-section of systematic reviews and meta-analyses that were published in February of 2014. That study found that 73% of meta-analytic effects were reproducible in theory, compared with the 64.3% found in our study. For articles in this study, adhering to PRISMA and citing a guide to interpret statistical heterogeneity both seemed to improve the reporting of effect estimates and measures of precision for the index meta-analysis. These effects are either small or imprecise and should be interpreted accordingly.

Strengths and Limitations
This study has several key strengths and limitations. Our sample of 154 is 40% larger than the previous study of reproducible research practices and is focused on only 1 area of medicine. Unlike previous investigations of data reporting in SRs, we extracted whether data necessary to reproduce meta-analyses was based on it being published in a forest plot or not. Systematic reviews with meta-analyses cited by oncology practice guidelines may represent the most important cohort of oncology systematic reviews because these systematic reviews inform guideline recommendations, in some cases. Yet, despite recent improvements in the quality of systematic reviews after the publication of the PRISMA statement, we found that key items were missing from oncology meta-analyses, which may hinder their reproducibility. The ability to reproduce all meta-analytic effects—even for secondary end points because systematic reviews are not powered for 1 end point like clinical trials—is fundamentally important because scientific progress requires trustworthy results. Although the inability to reproduce study findings does not mean the study findings are false, it may affect the interpretation of results, especially because our study defined "reproducibility" for main effects as the reporting of a summary effect, measure of precision, and list of included studies.
analyses (eg, summary statistics or effect estimates) were available from published reports, and whether subgroups or sensitivity analyses differed from the index meta-analyses in this regard. Concerning limitations, our sample of systematic reviews may not be generalizable to all systematic reviews of oncology interventions because we relied on the citations in NCCN guidelines. It is possible that other specialized organizations (eg, American Society of Hematology for blood cancers) cite different systematic reviews. Further, it may be that other systematic reviews of oncology interventions are more or less reproducible in theory than those in this study. We used double data extraction for only 15% of the included studies, which may increase the chance of data extraction errors. Despite the high percentage of agreement between authors, to mitigate the possibility of these errors, we extracted data a second time for all items with a discrepancy and used a third-party adjudicator. These quality checks are consistent with previous studies.4,22 Further, the absence of data to reproduce a meta-analysis effect does not necessarily imply it was incorrectly estimated, only that the availability of the data to reproduce may improve confidence for some readers in its accuracy.

Conclusions

We recommend that authors of systematic reviews with meta-analyses incorporate more reproducible research practices and expect guideline authors to evaluate whether existing systematic reviews and meta-analyses are reproducible. We further recommend journals encourage authors to present all meta-analyses in figures because standard graphical output for meta-analyses in most statistical packages includes a list of included studies and numerical point estimates. In this study, these 2 items alone were necessary to reproduce a summary effect, in theory. A guideline development group may downgrade the quality of systematic review data if they feel that the findings are not trustworthy. We further recommend earnest adherence to PRISMA because many of the reproducible research practices that we investigated are addressed therein, indicating that authors may incompletely adhere to PRISMA recommendations. Authors should make use of data repositories, such as the Open Science Framework, to store data, supplemental material, or other necessary items that ensures the reproducibility of findings.