How Well Can We Assess the Validity of Non-Randomised Studies of Medications? A Systematic Review of Assessment Tools

Elvira D'Andrea
Vinals
Elisabetta Patorno
Jessica M. Franklin
Dimitri Bennett

See next page for additional authors

Follow this and additional works at: https://digitalcommons.uri.edu/php_facpubs

Citation/Publisher Attribution
How Well Can We Assess the Validity of Non-Randomised Studies of Medications? A Systematic Review of Assessment Tools

Authors
Elvira D'Andrea, Vinals, Elisabetta Patorno, Jessica M. Franklin, Dimitri Bennett, Joan A. Largen, Daniela C. Moga, Hongbo Yuan, Xuerong Wen, Andrew R. Zullo, Thomas P. A. Debray, and Grammati Sarri

This article is available at DigitalCommons@URI: https://digitalcommons.uri.edu/php_facpubs/203
How well can we assess the validity of non-randomised studies of medications? A systematic review of assessment tools

Elvira D’Andrea 1,2, Lydia Vinals 2, Elisabetta Patorno 1, Jessica M. Franklin 1, Dimitri Bennett 3,4, Joan A. Largent 5, Daniela C. Mogá 6, Hongbo Yuan 7, Xuereong Wen 8, Andrew R. Zullo 9,10, Thomas P. A. Debray 11,12, Grammati Sarri 13

ABSTRACT

Objective To determine whether assessment tools for non-randomised studies (NRS) address critical elements that influence the validity of NRS findings for comparative safety and effectiveness of medications.

Design Systematic review and Delphi survey.

Data sources We searched PubMed, Embase, Google, bibliographies of reviews and websites of influential organisations from inception to November 2019. In parallel, we conducted a Delphi survey among the International Society for Pharmacoepidemiology Comparative Effectiveness Research Special Interest Group to identify key methodological challenges for NRS of medications. We created a framework consisting of the reported methodological challenges to evaluate the selected NRS tools.

Study selection Checklists or scales assessing NRS.

Data extraction Two reviewers extracted general information and content data related to the prespecified framework.

Results Of 44 tools reviewed, 48% (n=21) assess multiple NRS designs, while other tools specifically addressed case–control (n=12, 27%) or cohort studies (n=11, 25%) only. Response rate to the Delphi survey was 73% (35 out of 48 content experts), and a consensus was reached in only two rounds. Most tools evaluated methods for selecting study participants (n=43, 98%), although only one addressed selection bias due to depletion of susceptibles (2%). Many tools addressed the measurement of exposure and outcome (n=40, 91%), and measurement and control for confounders (n=40, 91%). Most tools have at least one item/question on design-specific sources of bias (n=40, 91%), but only a few investigate reverse causation (n=8, 18%), detection bias (n=4, 9%), time-related bias (n=3, 7%), lack of new-user design (n=2, 5%) or active comparator design (n=0). Few tools address the appropriateness of statistical analyses (n=15, 34%), methods for assessing internal (n=15, 34%) or external validity (n=11, 25%) and statistical uncertainty in the findings (n=21, 48%). None of the reviewed tools investigated all the methodological domains and subdomains.

Conclusions The acknowledgement of major design-specific sources of bias (eg, lack of new-user design, lack of active comparator design, time-related bias, depletion of susceptibles, reverse causation) and statistical assessment of internal and external validity is currently not sufficiently addressed in most of the existing tools. These critical elements should be integrated to systematically investigate the validity of NRS on comparative safety and effectiveness of medications.

INTRODUCTION

There are high expectations that real-world data (RWD) and resultant real-world evidence (RWE) will become a key source of information for the development process of pharmacological or biological therapies. 1-3 The 21st Century Cures Act and the sixth Prescription Drug User Fee Act required the Food and Drug Administration (FDA) to explore the use of RWE and, consequently, well-designed and conducted non-randomised studies...
(NRS) for expediting drug approvals. Similarly, one of the goals of the European Medicines Agency (EMA) Adaptive Pathways Initiative is to supplement clinical trial data with RWD and to eventually produce RWE as part of the approval process of new medications or indications.

However, the growing demand for RWD has raised concerns about the reliability of NRS to generate RWE. Due to the inherent limitations of observational analyses, the validity of NRS depends largely on the implementation of complex design and analytic methodologies. In recent reports, both FDA and EMA emphasised the need to plan and execute NRS following standards that can ensure validity and reproducibility of RWE. Tools that assess the validity of NRS can be useful instruments for both researchers (eg, for authors and reviewers to prevent publication of poor quality pharmacoepidemiological research) and other stakeholders who are involved in clinical, managerial or economic decision making (eg, to correctly inform guidelines and clinicians or to guide resource allocation).

An analysis on the capability of existing tools to assess the validity of NRS of comparative safety and effectiveness of medications is currently lacking. Previously published systematic reviews on assessment tools for NRS were mostly descriptive and did not provide a critical evaluation of the tools content investigated only a specific type of bias or focused only on safety outcomes. Therefore, we conducted a systematic review to assess the content of eligible tools for NRS of medications. There is no agreement on an assessment framework for NRS of pharmacological interventions. Thus, we performed a Delphi survey among international experts in the field of pharmacoepidemiology and health outcome research in order to build consensus for the methodological challenges that may threaten the validity of NRS of medications and that should be evaluated by assessment NRS tools.

The main objective of this study was to determine whether the retrieved NRS tools sufficiently address the main methodological challenges recommended by the experts. This study is part of a research project to develop a framework for the synthesis of NRS and randomised controlled trials (RCTs), led by the Comparative Effectiveness Research Special Interest Group (CER SIG) of the International Society for Pharmacoepidemiology (ISPE).

METHODS

The systematic review was conducted according to the Preferred Reporting Items for Systematic Reviews and Meta-analyses statement. Systematic review protocol and registration are available at https://osf.io/es65q.

Systematic search and eligibility criteria

We searched PubMed and Embase from inception to November 2019 to identify existing tools that investigated the validity of NRS, specifically case-control and cohort design studies. We excluded guidelines or manuals, tools to review study protocols, tools targeting NRS of non-pharmacological interventions (eg, surgery) or assessing only one or a few specific types of bias, and tools not available in English language. In parallel, we searched the same electronic databases for systematic reviews of assessment tools of NRS. We then extracted the references of the tools included in the systematic reviews retrieved. We also performed a general search through Google for grey literature and reviewed any additional information from initiatives, programmes or organisations. Full details on the search strategy are reported in the supplement (online supplemental tables S1 and S2). Three reviewers (ED, GS and LV) independently removed duplicates and reviewed titles and abstracts of peer-reviewed publications or documents from the grey literature to select eligible tools. Discrepancies were resolved by consensus.

Delphi survey and prespecified framework

Concurrently, we performed a Delphi survey to reach a consensus among content experts about the main methodological challenges (domains) that may threaten the validity of NRS on comparative safety and effectiveness of medications. The survey is available in the online supplemental 2. The panel of experts involved members of the SIG for CER of the ISPE. Detailed information on the Delphi methods and results is reported in the online supplemental 1.

Domains and subdomains indicated by the Delphi respondents as major elements that can impact the validity of NRS of medications were used to develop and pilot a framework to evaluate the identified NRS tools. All domains were considered equally important. A glossary of terms used in the framework is reported in table 1.

Data extraction

Two reviewers (ED and LV) independently extracted general information of the identified tools (first author or name of the tool, year of publication or online availability of the most updated version, type of tool, scope of the tool, NRS designs evaluated and number of items) and content data related to the prespecified domains of the framework. Discrepancies were resolved by consensus. We categorised the tools as checklists, defined as itemised instruments (including questionnaires) developed to identify the presence or absence of critical elements, or rating scales, defined as itemised instruments aimed to identify the performance of a study at each critical element described in the tool, using a qualitative or quantitative scale.

Data synthesis

General characteristics of the identified tools were summarised with means and SD, for continuous variables, and relative frequencies, for categorical variables. The findings from the Delphi survey and the proportion of tools assessing the prespecified elements of the framework were reported in terms of relative frequencies.
Table 1  Glossary of terms

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Active comparator design</td>
<td>A study design that compares the effect of the drug of interest with another drug used in clinical practice instead of non-use.</td>
</tr>
<tr>
<td>Adjustment for causal intermediaries</td>
<td>Adjustment for an intermediate variable (or a descending proxy for an intermediate variable) on a causal path from exposure to outcome.</td>
</tr>
<tr>
<td>Case–control design</td>
<td>A study design in which cases (patients with outcomes) are identified and compared with controls (patients without outcomes) with respect to the exposure of interest.</td>
</tr>
<tr>
<td>Cohort design</td>
<td>A study design in which a group of patients (a cohort) is identified and followed to ascertain the occurrence of an outcome.</td>
</tr>
<tr>
<td>Confounding</td>
<td>A mixing of effects that arises when patients with different baseline risks are compared; the resulting effect measure is a mix of drug effects and risk factor effects.</td>
</tr>
<tr>
<td>Depletion of susceptibles</td>
<td>Selection bias that occurs when the initiation of exposure to a drug is associated with an early increased incidence rate of the study outcome, followed by a decreased incidence rate with longer duration of exposure (eg, users of new drugs are compared with users of older drugs).</td>
</tr>
<tr>
<td>Detection or surveillance bias</td>
<td>Bias that occurs when the degree of outcome surveillance (or an associated symptom) is related to exposure and is differential among the exposure groups.</td>
</tr>
<tr>
<td>Immortal time bias</td>
<td>Time-related bias that derives from including a period of follow-up during which, by design, outcomes cannot occur.</td>
</tr>
<tr>
<td>Time-window bias</td>
<td>Time-related bias, in the context of a case–control study nested in a cohort, that derives from the use of time-windows of different lengths between cases and controls to define time-dependent exposures.</td>
</tr>
<tr>
<td>Incorrect outcome model specification</td>
<td>Misspecification of a statistical model that leads to biased outcome results. Common causes are omission of a relevant variable, inclusion of an unnecessary variable, adopting the wrong functional form, incorrect specification of the error term, uncertainty about what the true model is and reciprocal causation.</td>
</tr>
<tr>
<td>Loss to follow-up bias</td>
<td>Bias that occurs when there is difference in retention during the follow-up period after enrolment that are related to exposure status and outcome.</td>
</tr>
<tr>
<td>New-user design</td>
<td>A study design that starts following patients at the time they initiate a new drug (also known as incident-user design)</td>
</tr>
<tr>
<td>Non-contemporaneous comparator bias</td>
<td>Bias generated by differences in the timing of selection of comparator group(s) within a study influence exposures and outcomes resulting in biased estimates.</td>
</tr>
<tr>
<td>Reverse causation (or reverse causality)</td>
<td>Bias due to direction of cause and effect contrary to a common presumption, or a two-way causal relationship between exposure and outcome.</td>
</tr>
<tr>
<td>Recall bias</td>
<td>Bias that occurs when participants do not remember previous events or experiences accurately or omit details (not for claims-based studies).</td>
</tr>
<tr>
<td>Selection bias</td>
<td>Bias that occurs when selection of participants or follow-up time is related to both intervention and outcome (eg, prevalent users of a drug are compared with non-users or incident users). Our framework has a separate subdomain that refers to selection bias due to lack of generalisability, applicability or transferability to patients who were excluded from the study.</td>
</tr>
</tbody>
</table>

RESULTS
Overview of tools
Of 44 tools that met our eligibility criteria,19–52 20 (45%) were identified through the database search of peer-reviewed literature and 24 (55%) through the general online search and other sources (online supplemental figure S1 and tables S2 and S3). Characteristics of the tools are shown in tables 2 and 3. The number of items across all tools ranged from 5 to 54, with a median of 13.5 (IQR 10.3–22). Only three tools were designed to specifically address studies on the comparative safety and effectiveness of pharmacological interventions: one published in 1994 by Cho and Bero,46 the The Good ReseArch for Comparative Effectiveness (GRACE) checklist and the International Society for Pharmacoeconomics and Outcomes Research – Academy of Managed Care Pharmacy – National Pharmaceutical Council (ISPOR-AMCP-NPC) tool, both published in 201425 26

Tool formats and scopes
Most of the tools were checklists (n=35, 80%), and 13 checklists included a final section to elaborate a summary judgement of the study appraisal (37%). The remaining tools were scales (n=9, 20%), and six of them provided a section for a summary judgement (67%).

Thirty-five tools (80%) were designed as critical appraisal tools for different scopes (eg, assessing the quality of NRS included in a systematic review, screening eligible NRS to
include in systematic reviews to support clinical guidelines, supporting peer-review processes or, more general, allowing readers to interpret NRS results critically. Four tools (9%) were developed to assess the quality of reporting and were mainly intended for researchers. Five other tools (11%) combined elements of both critical appraisal and quality.
Study designs addressed
Twenty-one tools (48%) were developed to assess multiple NRS designs (11 targeted cohort and case–control studies and 10 others addressed also other NRS designs or did not specify them). Other tools specifically addressed case–control (n=12, 27%) or cohort studies (n=11, 25%). Ten tools (23%) were designed to assess also RCTs.

Tool elements
The response rate of the Delphi survey was 73% (35 respondents out of 48 members). Detailed results are reported in the online supplemental figure S2. Domains and subdomains indicated by the respondents as major elements that can impact the validity of NRS of medications are reported in the first column of Table 4.

Methods for selecting participants
Nearly all tools assessed methods for selecting study participants to correct selection bias (n=43, 98%). Specifically, almost half of the tools included items related to sampling strategies (n=19, 43%), the definition of inclusion and exclusion criteria (n=27, 61%) and the generalisability of participants (ie, attempts to achieve a sample of participants that represents the target population) (n=21, 48%), while only one tool addressed the depletion of susceptibles (n=1, 2%) (Table 4 and online supplemental figure S3).

Measurement of exposure, outcomes, covariates and follow-up
Forty-two tools (95%) had at least one item assessing the definition and measurement of exposure, outcome, covariates and follow-up. Assessment of exposure and outcome was widely reported by the tools (n=40, 91%), while definition and measurement of covariates (n=12, 27%) or follow-up (n=17, 39%) were less often addressed (with the exception for tools addressing follow-up in cohort studies only, n=9, 82%) (Table 4 and online supplemental figure S4).

Design-specific sources of bias
Design-specific sources of bias (excluding selection bias which was investigated in ‘Methods for selecting participants’) were assessed by 91% of the tools (n=40) and generally included loss to follow-up bias (n=22, 50%), observer or interviewer bias (n=11, 25%), reverse causation bias (n=8, 18%), recall bias (n=6, 14%) and non-contemporaneous comparator bias (n=6, 14%). A few or no tools assessed detection or surveillance bias (n=4, 9%), time-related bias, such as immortal person-time bias or time-window bias (n=3, 7%), and biases due to lack of new-user design (n=2, 5%) or active comparator design (n=0). Other tools reported only a general item/question on the risk of bias (n=9, 20%), without any reference to a specific type of bias.

Tools specifically for cohort studies addressed more frequently loss to follow-up (n=9, 82%) and reverse causation biases (n=5, 45%) compared with the other tools, while tools for case–control studies addressed mostly recall (n=4, 27%).

Table 3 General characteristics of the assessment tools included in the systematic review

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>All, n=44</th>
<th>Cohort*, n=11</th>
<th>Case–control, n=12</th>
<th>NRS†, n=21</th>
</tr>
</thead>
<tbody>
<tr>
<td>Publication year, n (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1979–1989</td>
<td>3 (7)</td>
<td>0 (0)</td>
<td>2 (17)</td>
<td>1 (5)</td>
</tr>
<tr>
<td>1990–1999</td>
<td>12 (27)</td>
<td>2 (18)</td>
<td>2 (17)</td>
<td>8 (38)</td>
</tr>
<tr>
<td>2000–2009</td>
<td>5 (11)</td>
<td>1 (9)</td>
<td>1 (8)</td>
<td>3 (14)</td>
</tr>
<tr>
<td>2010–2019</td>
<td>24 (55)</td>
<td>8 (73)</td>
<td>7 (58)</td>
<td>9 (43)</td>
</tr>
<tr>
<td>Type of tool, n (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Checklist</td>
<td>22 (50)</td>
<td>4 (36)</td>
<td>6 (50)</td>
<td>12 (57)</td>
</tr>
<tr>
<td>Checklist+summary judgement</td>
<td>13 (30)</td>
<td>5 (45)</td>
<td>4 (33)</td>
<td>4 (19)</td>
</tr>
<tr>
<td>Rating scale</td>
<td>3 (7)</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>3 (14)</td>
</tr>
<tr>
<td>Rating scale+summary judgement</td>
<td>6 (14)</td>
<td>2 (18)</td>
<td>2 (16)</td>
<td>2 (9)</td>
</tr>
<tr>
<td>Scope of the tool, n (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Critical appraisal</td>
<td>35 (80)</td>
<td>9 (81)</td>
<td>10 (83)</td>
<td>16 (76)</td>
</tr>
<tr>
<td>Reporting</td>
<td>4 (9)</td>
<td>2 (18)</td>
<td>1 (8)</td>
<td>1 (5)</td>
</tr>
<tr>
<td>Critical appraisal and reporting</td>
<td>5 (11)</td>
<td>0 (0)</td>
<td>1 (8)</td>
<td>4 (19)</td>
</tr>
<tr>
<td>Tools designed for CER, n (%)</td>
<td>3 (7)</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>3 (14)</td>
</tr>
<tr>
<td>Number of items, median (IQR)</td>
<td>13 (10.3–21.8)</td>
<td>13 (9.5–16)</td>
<td>11.5 (10.8–18.5)</td>
<td>17 (11–24)</td>
</tr>
</tbody>
</table>

*Two tools evaluated both cohort and RCTs together; one tool evaluated both cohort and cross-sectional studies together.
†NRS tools refer to a single tool built to evaluate both cohort and case–control studies or a tool built to evaluate additional NRS (eg, cross-sectional studies and before–after studies) together with cohort and case–control studies. Eight NRS tools included also the evaluation of RCTs.
CER, Comparative Effectiveness research; NRS, non-randomised studies; RCTs, randomised controlled trials.
33%) and observer biases (n=3, 25%). Tools for multiple NRS covered commonly loss to follow-up (n=12, 57%) and interviewer or observer biases (n=7, 35%) (table 4 and online supplemental figure S5).

Confounding
Forty tools (91%) included at least one item or question related to confounding. Specifically, 26 tools (59%) searched whether study design was planned in a way to minimise

<table>
<thead>
<tr>
<th>Domains</th>
<th>Cohort tools*, n=11</th>
<th>Case–control tools, n=12</th>
<th>NRS tools†, n=21</th>
<th>Total, n=44</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Methods for selecting participants, n (%)</td>
<td>11 (100)</td>
<td>12 (100)</td>
<td>20 (95)</td>
<td>43 (98)</td>
</tr>
<tr>
<td>Sampling strategies to correct selection bias</td>
<td>4 (36)</td>
<td>6 (50)</td>
<td>9 (42)</td>
<td>19 (43)</td>
</tr>
<tr>
<td>Inclusion and exclusion criteria of target population</td>
<td>6 (55)</td>
<td>8 (67)</td>
<td>13 (61)</td>
<td>27 (61)</td>
</tr>
<tr>
<td>Depletion of susceptibles</td>
<td>1 (9)</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>1 (2)</td>
</tr>
<tr>
<td>External validity of target population</td>
<td>6 (55)</td>
<td>6 (50)</td>
<td>9 (43)</td>
<td>21 (48)</td>
</tr>
<tr>
<td>Others‡</td>
<td>11 (100)</td>
<td>12 (100)</td>
<td>18 (86)</td>
<td>41 (93)</td>
</tr>
<tr>
<td>2. Measurement of exposure, outcomes, covariates and follow-up, n (%)§</td>
<td>11 (100)</td>
<td>12 (100)</td>
<td>19 (90)</td>
<td>42 (95)</td>
</tr>
<tr>
<td>Measurement of exposure§</td>
<td>11 (100)</td>
<td>11 (92)</td>
<td>18 (81)</td>
<td>40 (91)</td>
</tr>
<tr>
<td>Measurement of outcomes§</td>
<td>11 (100)</td>
<td>11 (92)</td>
<td>18 (81)</td>
<td>40 (91)</td>
</tr>
<tr>
<td>Measurement of covariates</td>
<td>4 (36)</td>
<td>4 (33)</td>
<td>4 (19)</td>
<td>12 (27)</td>
</tr>
<tr>
<td>Measurement of follow-up</td>
<td>9 (82)</td>
<td>3 (25)</td>
<td>5 (24)</td>
<td>17 (39)</td>
</tr>
<tr>
<td>3. Design-specific sources of bias, n (%)</td>
<td>11 (100)</td>
<td>10 (83)</td>
<td>19 (90)</td>
<td>40 (91)</td>
</tr>
<tr>
<td>New-user design</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>2 (10)</td>
<td>2 (5)</td>
</tr>
<tr>
<td>Active comparator design</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Immortal time bias or time-window bias</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>3 (14)</td>
<td>7 (17)</td>
</tr>
<tr>
<td>Detection or surveillance bias</td>
<td>1 (9)</td>
<td>2 (17)</td>
<td>1 (5)</td>
<td>4 (9)</td>
</tr>
<tr>
<td>Loss to follow-up bias</td>
<td>9 (82)</td>
<td>1 (8)</td>
<td>12 (57)</td>
<td>22 (50)</td>
</tr>
<tr>
<td>Non-contemporaneous comparator bias</td>
<td>0 (0)</td>
<td>1 (8)</td>
<td>5 (24)</td>
<td>6 (14)</td>
</tr>
<tr>
<td>Reverse causation</td>
<td>5 (45)</td>
<td>1 (8)</td>
<td>2 (10)</td>
<td>8 (18)</td>
</tr>
<tr>
<td>Recall bias¶</td>
<td>1 (9)</td>
<td>4 (33)</td>
<td>1 (5)</td>
<td>6 (14)</td>
</tr>
<tr>
<td>Interviewer or observer bias¶</td>
<td>1 (9)</td>
<td>3 (25)</td>
<td>7 (35)</td>
<td>11 (25)</td>
</tr>
<tr>
<td>Ascertainment bias¶</td>
<td>0 (0)</td>
<td>1 (8)</td>
<td>1 (5)</td>
<td>2 (5)</td>
</tr>
<tr>
<td>General item/question on bias¶</td>
<td>3 (27)</td>
<td>3 (25)</td>
<td>3 (14)</td>
<td>9 (20)</td>
</tr>
<tr>
<td>Other biases**</td>
<td>0 (0)</td>
<td>2 (17)</td>
<td>5 (24)</td>
<td>7 (16)</td>
</tr>
<tr>
<td>4. Confounding, n (%)</td>
<td>11 (100)</td>
<td>11 (92)</td>
<td>18 (86)</td>
<td>40 (91)</td>
</tr>
<tr>
<td>Study design used to minimise confounding</td>
<td>6 (55)</td>
<td>7 (58)</td>
<td>13 (62)</td>
<td>26 (59)</td>
</tr>
<tr>
<td>Confounders measured and included in statistical analyses</td>
<td>10 (91)</td>
<td>10 (83)</td>
<td>18 (86)</td>
<td>38 (86)</td>
</tr>
<tr>
<td>Potential unmeasured confounding addressed in the analysis (eg, proxy analysis and IV analysis)</td>
<td>1 (9)</td>
<td>1 (8)</td>
<td>3 (14)</td>
<td>5 (11)</td>
</tr>
<tr>
<td>5. Lack of appropriateness of statistical analyses (with specific mention of overadjustment and/or incorrect outcome model specification), n (%)</td>
<td>2 (18)</td>
<td>3 (25)</td>
<td>10 (48)</td>
<td>15 (34)</td>
</tr>
<tr>
<td>6. Methods for assessing statistical uncertainty in the findings (eg, CIs reported for each analysis), n (%)</td>
<td>7 (64)</td>
<td>6 (50)</td>
<td>8 (38)</td>
<td>21 (48)</td>
</tr>
<tr>
<td>7. Methods for assessing internal validity (eg, sensitivity analysis addressing potential confounding, measurement errors or other biases), n (%)</td>
<td>3 (27)</td>
<td>3 (25)</td>
<td>9 (43)</td>
<td>15 (34)</td>
</tr>
<tr>
<td>8. Methods for assessing external validity (eg, post hoc subgroup analysis and comparison with other populations), n (%)</td>
<td>4 (36)</td>
<td>3 (25)</td>
<td>4 (19)</td>
<td>11 (25)</td>
</tr>
</tbody>
</table>

*Two tools evaluated both cohort and RCTs together; one tool evaluated both cohort and cross-sectional studies together.
†NRS tools refer to a single tool built to evaluate both cohort and case–control studies or a tool built to evaluate additional NRS (eg, cross-sectional studies and before–after studies) together with cohort and case–control studies. Eight NRS tools included also the evaluation of RCTs.
‡‘Others’ refers to items not included in our evaluation framework but included in the reviewed tools to investigate selection bias (eg, population characteristics sufficiently described to determine the applicability of the research question, sample size justification and power description, and ethical considerations).
§Items or questions on exposure misclassification and/or outcome misclassification are counted in this domain and relative subdomains.
¶Design-specific biases not included in the evaluation framework but addressed by the reviewed tools.
**Other design-specific biases not included in the evaluation framework but addressed by a few tools (eg, bias due to missing data, patients’ blinding, different length of follow-up between groups, Berkson’s bias and protopathic bias).
IV, instrumental variable; NRS, non-randomised studies; RCTs, randomised controlled trials.
confounding, 38 (86%) whether confounders were measured and included in the analyses and only five whether potential unmeasured confounding was assessed in the sensitivity analyses (11%) (table 4 and online supplemental figure S6).

Appropriateness of statistical analyses, external and internal validity

One-third of the tools (n=14, 32%) assessed the appropriateness of statistical analyses, although most of them did not explicitly mention overadjustment of causal intermediates and/or incorrect outcome model specification. Almost half (n=21, 48%) included methods for measuring uncertainty in the findings. Few tools addressed methods for evaluating internal (n=15, 34%) or external validity (n=11, 25%) (table 4 and online supplemental figure S7 in the online supplement).

These results were mostly consistent across the three different types of design addressed, cohort only, case-control only and multiple NRS, except for the assessment of follow-up (domain 2) and several design-specific sources of biases (domain 3) already mentioned above (table 4). None of the reviewed tools covered all the main domains and subdomains as identified by the CER SIG and listed in table 4.

Results for each selected tool on the proportions of items/questions that investigate the prespecified domains are shown in the online supplemental figures S8–S11.

DISCUSSION

In this systematic review, we identified assessment tools evaluating the validity of NRS on comparative safety and effectiveness of medications. Of 44 tools included, only three were specifically designed to assess NRS of pharmacological interventions.

Main findings

Overall, we found that existing tools assessed most of the methodological challenges identified by the domains of the CER SIG framework, but critical elements were often insufficiently addressed. For example, although many tools assessed the risk of selection bias, only half of them explicitly investigated sampling strategies and considered a prespecification of inclusion/exclusion criteria. Even more surprising was that only one tool explored the potential for selection bias due to depletion of patients that are susceptible to the outcome. This cohort-based phenomenon can occur when new users of a medication are depleted of all susceptible subjects to the outcome, documenting an increased incidence rate of the outcome in an early stage, followed by a decreased rate with a longer duration of exposure. Depletion of susceptibles is an important source of bias to account for when evaluating effects of new medications in incident users and can significantly undermine the validity of the results.

Similarly, many tools investigated misclassification or information bias of exposure and outcome. However, only about one-third assessed definition and measurement of covariates, and less than one-fourth of the case–control and multiple NRS designs tools assessed information on follow-up definition. Again, these are common causes of bias and should be integrated in tools that investigate the validity of NRS.

Design-specific sources of bias was a critical domain. Although overall 91% of the tools had at least one item/question investigating biases due to an inappropriate study design, only Risk Of Bias In Non-randomized Studies of Interventions (ROBINS-I) and the GRACE checklist addressed bias due to lack of new-user design and time-related bias (ie, immortal person-time bias or time-window bias), while no tools investigated bias due to lack of active comparator design. Since these biases can independently lead to major methodological flaws (defined as elements that by themselves can significantly compromise the validity of the results), their assessment must be included in appraisal tools for NRS of pharmacological interventions. For example, recent evidence on NRS of glucose-lowering medications reported that only one-fourth of the studies adopted a new-user design and less than half used an active comparator. In the same example, potential for time-related bias was detected in more than two-thirds of the studies. Integrating the evaluation of these major methodological flaws in existing tools and recommending the use of these tools before publication can increase awareness in the clinical research community with respect to main design-specific biases. This can ultimately decrease the amount of NRS with invalid findings on the safety or effectiveness of medications.

A high percentage of tools evaluated whether confounders were appropriately measured, controlled for in the analysis and considered in the study design. However, very few tools included at least one item/question on whether potential unmeasured confounding had been considered in the analysis or interpretation of findings.

One-third of the tools checked the appropriateness of statistical analyses, but most omitted specific reference to common flaws such as overadjustment or incorrect outcome model specification. Similarly, only one-third of the tools assessed internal validity (eg, through sensitivity analysis to address potential confounding, measurement errors or other biases), and only one-fourth assessed external validity (eg, post hoc subgroup analysis and comparison with other populations).

Implications for practice and research

While recently published tools such as The Critical Appraisals Skills Programme checklist, ISPOR-AMCP-NPC, Recruitment Allocation Maintenance blind objective Measurements Analyses, GRACE, and ROBINS-I are among the most complete tools, addressing several of the critical elements underlined by the ISPE CER SIG, they all had limitations in the acknowledgment of two or more major methodological challenges (eg, selection bias for depletion of susceptibles,
immortal-time bias or window-time bias, lack of new-user design, lack of active comparator design, reverse causation bias and adjustment for causal intermediaries). Assessment tools can be powerful instruments for researchers, authors, reviewers of scientific journals or readers, helping to identify the main limitations of a study and to correctly interpret the results, to acknowledge major methodological flaws and, ultimately, to prevent publication of studies with invalid findings.

Furthermore, other decision makers, such as clinicians, guideline developers and payers or investors, can benefit from instruments that help to ensure the validity of NRS findings. RCTs can be an insufficient source of evidence for decisions on pharmaceutical interventions. Despite well-designed and adequately powered RCTs being considered the ‘gold standard’ of the clinical research paradigm, they can often be too time intensive and money intensive. Trials are often relatively small, focus on short-term efficacy and safety in a controlled clinical environment, using surrogate outcomes or under-representing high-risk populations that can be most likely the target on the new medications in the real-world setting. Trials might also not record treatments taken outside the study protocol. Additionally, patients volunteered to participate in a trial are usually very motivated and so more adherent to therapy compared with the real-world population. NRS based on RWD can help to address these issues and could be supplement the evidence from RCTs to provide a more complete picture on the effectiveness of pharmaceutical interventions in less controlled environments. NRSs have the advantages to investigate large-scale populations, high-risk subpopulations, rare exposures, diseases or outcome, and long-term outcomes or other delayed health effects at low costs and rapidly. Moreover, since RWD are often collected for intents unrelated to research objectives (mainly administrative), biases such as recall bias, interviewer bias, non-response bias and bias for loss to follow-up are reduced or eliminated. Thus, since RWE derived by NRS contribute significantly to generate evidence of comparative effectiveness research of medications and reduce the publication of studies with unreliable findings.

CONCLUSION
In this systematic review, we found that available tools for NRS assessment failed to provide a comprehensive assessment of major methodological aspects that can affect the validity of NRS on the comparative safety and effectiveness of medications. Specifically, major aspects such as lack of new-user design, active comparator design, time-related bias (ie, immortal time bias and time-window bias) and statistical assessment of internal validity remain poorly covered. Including these critical elements into existing tools may provide a more accurate instrument to evaluate NRS of pharmaceutical interventions and increase awareness in the clinical research community about major addressable flaws in pharmacoepidemiology. This may improve the validity of NRS on the comparative safety and effectiveness of medications and reduce the publication of studies with unreliable findings.

Strengths and limitations
To our knowledge, this is the first systematic review that investigated whether existing tools adequately assess the validity of cohort and case–control studies evaluating the comparative safety and effectiveness of medications. Previously published systematic reviews on assessment tools for NRS were not specifically focused on pharmacoepidemiological interventions included randomised study designs or investigated only a specific type of bias.

One systematic review of NRS tools for medications focused only on safety outcomes, and it is now outdated since published in 2012. Our systematic review has multiple strengths: authors reviewed the results of the searches independently following a predefined protocol; the framework for data extraction was developed based on inputs of worldwide experts in the field of pharmacoepidemiology and healthcare outcomes research coming from different backgrounds (academia, industry and governmental agency) and different countries, and it included the most updated versions of the identified tools. This review also has limitations. Search for tools in the grey literature might not be comprehensive since it was performed through only one browser. The search was also restricted to tools published in English and excluded identified tools that could not be retrieved.

Author affiliations
1 Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women’s Hospital and Harvard Medical School, Boston, Massachusetts, USA
2 HEOR Department, Cytel Inc, Toronto, Quebec, Canada
3 Pharmacoepidemiology, Takeda Pharmaceutical, Cambridge, Massachusetts, USA
4 Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
5 Real-World Solutions, IQVIA, California, Los Angeles, USA
6 Department of Pharmacy Practice and Science, College of Pharmacy, University of Kentucky, Lexington, KY, USA
7 Canadian Agency for Drugs and Technologies in Health (CADTH), Ottawa, Ontario, Canada
8 Department of Pharmacy Practice, University of Rhode Island, Kingston, RI, USA
9 Department of Health Services, Policy, and Practice, Brown University, Providence, Rhode Island, USA
10 Center of Innovation in Long-term Services and Supports, Providence VA Medical Center, Providence, Rhode Island, USA
11 Department of Epidemiology, Julius Center for Health Sciences and Primary Care, Utrecht, The Netherlands
12 Smart Data Analysis and Statistics, Utrecht, The Netherlands
Funding
This project has received funding from the European Union’s Horizon 2020 research and innovation programme under ReCoDiG grant agreement No 825746.

Competing interests
DB is an employee of Takeda. ARZ has received salary support from Sanofi Pasteur through a grant to Brown University unrelated to the current work. TD provides consulting services via Smart Data Analysis and Performance Improvement.

Patient consent for publication
Not required.

Provenance and peer review
Not commissioned; externally peer reviewed.

Data availability statement
All data relevant to the study are included in the article or uploaded as supplemental information.

Supplemental material
This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

Open access
This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: http://creativecommons.org/licenses/by-nc/4.0/. 

ORCID iDs
Elvira D’Andrea http://orcid.org/0000-0002-5263-3964
Elisabetta Patorno http://orcid.org/0000-0002-8809-8988
Dimitri Bennett http://orcid.org/0000-0002-8387-9342
Thomas P.A. Debray http://orcid.org/0000-0002-1790-2719
Grammali Sarri http://orcid.org/0000-0001-5536-8038

REFERENCES


Seoensen HT, Lash TL, Rothman KJ. Beyond randomized controlled trials: a critical comparison of trials with nonrandomized studies.