BMJ Open

How well can we assess the validity of non-randomised studies of medications? A systematic review of assessment tools

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

Strengths and limitations of this study

- ► This is the first systematic review to investigate whether existing tools adequately assess the validity of non-randomised studies evaluating the comparative safety and effectiveness of medications.
- ► Assessment tools were identified by searching through multiple sources: relevant databases, grey literature, websites of authoritative organisations, bibliographies of previous systematic reviews and experts' suggestions.
- ► The prepiloted framework adopted to evaluate the completeness of the tools included all the main methodological challenges suggested by an interdisciplinary (academia, industry and government agencies) and international team of experts in the field of pharmacoepidemiology and healthcare outcomes research.
- ► Tools not published in English or that could not be retrieved were omitted from this systematic review.
- ► The search for tools in the grey literature might not be comprehensive since it was performed through only one browser.

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.

Systematic review protocol and registration https://osf.io/es65q.

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.[4 5] 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.[6]

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.[7 8] 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,[9–13] investigated only a specific type of bias[14] or focused only on safety outcomes.[15] 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),[16] 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.[17] 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[18] 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

Term	Definition
Active comparator design	A study design that compares the effect of the drug of interest with another drug used in clinical practice instead of non-use.
Adjustment for causal intermediaries	Adjustment for an intermediate variable (or a descending proxy for an intermediate variable) on a causal path from exposure to outcome.
Case–control design	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.
Cohort design	A study design in which a group of patients (a cohort) is identified and followed to ascertain the occurrence of an outcome.
Confounding	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.
Depletion of susceptibles	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).
Detection or surveillance bias	Bias that occurs when the degree of outcome surveillance (or an associated symptom) is related to exposure and is differential among the exposure groups.
Immortal time bias	Time-related bias that derives from including a period of follow-up during which, by design, outcomes cannot occur.
Time-window bias	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.
Incorrect outcome model specification	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.
Loss to follow-up bias	Bias that occurs when there is difference in retention during the follow-up period after enrolment that are related to exposure status and outcome.
New-user design	A study design that starts following patients at the time they initiate a new drug (also known as incident-user design)
Non-contemporaneous comparator bias	Bias generated by differences in the timing of selection of comparator group(s) within a study influence exposures and outcomes resulting in biased estimates.
Reverse causation (or reverse causality)	Bias due to direction of cause and effect contrary to a common presumption, or a two-way causal relationship between exposure and outcome.
Recall bias	Bias that occurs when participants do not remember previous events or experiences accurately or omit details (not for claims-based studies).
Selection bias	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.

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 table 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 2014.[25 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

Table 2 Individual characteristics of the tools included in the systematic review

Tool identified*	Year	Type of tool	Scope of the tool	Study design evaluated	tems
RELEVANT	2019	Checklist	Critical appraisal and reporting	NRS	21
RAMboMAN - GATE-EPIQ	2019	Rating scale+summary judgement	Critical appraisal	Coh (+RCTs), CC	Coh (+RCTs) 21, CC 18
MMAT	2018	Checklist	Critical appraisal	NRS	5
CASP	2018	Checklist	Critical appraisal	Coh, CC	Coh 12, CC 11
SURE	2018	Checklist+summary judgement	Critical appraisal	Coh, CC	Coh 13, CC 11
JBI	2017	Checklist+summary judgement	Critical appraisal	Coh, CC	Coh 11, CC 10
ROBINS-I	2016	Checklist+summary judgement	Critical appraisal	NRS	34 (+8 optional question)
ISPOR-AMCP-NPC†	2014	Checklist+summary judgement	Critical appraisal	Coh CC	32
GRACE†	2014	Checklist+summary judgement	Critical appraisal	Coh CC	11
NIH–NHLBI	2014	Checklist+summary judgement	Critical appraisal	Coh (+CSS), CC	Coh (+CSS) 14, CC 12
HEBW	2014	Checklist+summary judgement	Critical appraisal	Coh	18
RoBANS	2013	Rating scale	Critical appraisal	NRS	6
RTI-Item Bank	2013	Checklist	Critical appraisal	NRS	13
Newcastle-Ottawa	2013	Rating scale +summary judgement	Critical appraisal	Coh, CC	Coh 8, CC 8
SIGN - V.3.0	2012	Checklist+summary judgement	Critical appraisal	Coh, CC	Coh 14, CC 11
Montreal	2011	Checklist	Critical appraisal	Coh CC (+RCTs)	10
EPHPP	2011	Rating scale	Critical appraisal	Coh CC (+RCTs)	17
STROBE – V.4	2007	Checklist	Reporting	Coh, CC	Coh 22, CC 22
TREND	2004	Checklist	Reporting	NRS	22
Margetts	2002	Checklist	Reporting	Coh CC	11
Zaza	2000	Checklist	Critical appraisal	Coh CC	15
Downs-Black	1998	Rating scale	Critical appraisal and reporting	Coh CC (+RCTs)	27
Elwood	1998	Checklist	Critical appraisal	Coh CC (+RCTs)	20
Hadorn	1996	Checklist	Critical appraisal	Coh (+RCTs)	7
London	1996	Checklist	Critical appraisal	Coh CC	33
Avis	1994	Rating scale+summary judgement	Critical appraisal and reporting	Coh CC (+RCTs)	24
Durant	1994	Checklist	Critical appraisal	CC	23
Levine	1994	Checklist	Critical appraisal	Coh CC (+RCTs)	10
Gyorkos	1994	Checklist	Critical appraisal	Coh, CC	Coh 6, CC 5
Cho†	1994	Rating scale+summary judgement	Critical appraisal	NRS (+RCTs)	24
COEH	1991	Checklist	Critical appraisal	NRS	54
Fowkes-Fulton	1991	Checklist+summary judgement	Critical appraisal and reporting	NRS (+RCTs)	6
Lichtenstein	1987	Checklist	Critical appraisal	CC	20
Gardner	1986	Checklist	Critical appraisal	NRS	12
Horwitz	1979	Checklist	Critical appraisal and reporting	CC	12

Nine tools from our bibliographic search provided two separate instruments to assess cohort or case–control studies. Thus, the overall number of included records is 35, while the number of included assessment tools is 44.

*Tool name or first author name, if the tool does not have an assigned name, and it was published in peer-review journals.

†Tool developed to assess NRS on the comparative safety and effectiveness of medications.

CASP, The Critical Appraisals Skills Programme; CC, case–control study; COEH, Centre for Occupational and Environmental Health of The University of Manchester; Coh, cohort study; CSS, cross-sectional study; EPHPP, Effective Public Health Practice Project Quality Assessment Tool; GRACE, The Good ReseArch for Comparative Effectiveness; HEBW, Health Evidence Bulletins Wales; ISPOR-AMCP-NPC, International Society for Pharmacoeconomics and Outcomes Research – Academy of Managed Care Pharmacy – National Pharmaceutical Council; JBI, The Joanna Briggs Institute; MMAT, Mixed Methods Appraisal Tool; NIH–NHLBI, The National Institute of Health - The National Heart, Lung, and Blood Institute; NRS, non-randomised studies; RAMboMAN, GATE-EPIQ, Recruitment Allocation Maintenance blind objective Measurements Analyses, Graphic Approach To Epidemiology – Effective Practice, Informatics and Quality Improvement; RCTs, randomised controlled trials; RELEVANT, The REal Life EVidence AssessmeNt Tool; RoBANS, Risk of Bias Assessment tool for Non-randomized Studies; ROBINS-I, Risk Of Bias In Non-randomized Studies of Interventions; RTI-Item Bank, Research Triangle Institute Item Bank; SIGN, The Scottish Intercollegiate Guidelines Network; STROBE, STrengthening the Reporting of OBservational studies in Epidemiology; SURE, Specialist Unit for Review Evidence; TREND, Transparent Reporting of Evaluations with Non-randomized designs.

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

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

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

*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.

reporting and were for a more general audience (both researchers and readers) (tables 2 and 3).

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,

Table 4 Methodological challenges addressed by the included assessment tools

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

*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.

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

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.[25 26 46]

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.[53] 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.[53]

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.[54] In the same example, potential for time-related bias was detected in more than two-third of the studies.[54] 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,[21] ISPOR-AMCP-NPC,[25] Recruitment Allocation Maintenance blind objective Measurements Analyses,[19] GRACE[26] and ROBINS-I[24] are among the most complete tools, addressing several of the critical elements underlined by the ISPE CER SIG, they all had limitations in the acknowledgement 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.[55][56] 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.[55][56] Trials might also not record treatments taken outside the study protocol.[47] Additionally, patients volunteered to participate in a trial are usually very motivated and so more adherent to therapy compared with the real-world population.[56] 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.[55][56] 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.[55] Thus, since RWE derived by NRS contribute significantly to generate evidence of comparative effectiveness research of medications, our synthesis can help numerous stakeholders to evaluate whether the NRS considered are valid enough to guide decision making.

Although checklists have been previously suggested for reviewing the risk of bias of general NRS,[57] we cannot strongly recommend a specific tool for NRS on comparative analyses of medications. As already mentioned, items or questions that address all those methodological flaws must be integrated in the existing tools. Based on our findings, most recent and comprehensive tools such as ROBINS-I[24] and GRACE[26] assessed a higher number of major methodological elements and could therefore be prioritised in this endeavour.

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 pharmacological interventions,[9][10] included randomised study designs[11–13] or investigated only a specific type of bias.[14] One systematic review of NRS tools for medications focused only on safety outcomes, and it is now outdated since published in 2012.[15] 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.

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 pharmacological 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.

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

ORCID iDs
Elvira D'Andrea http://orcid.org/0000-0002-5263-3964
Elisabetta Patorno http://orcid.org/0000-0002-8809-9898
Dimitri Bennett http://orcid.org/0000-0002-8387-9342
Thomas P. A. Debray http://orcid.org/0000-0002-1790-2719
Grammati Sarri http://orcid.org/0000-0001-5536-8038

REFERENCES

1 Eichler H-G, Bloechl-Daum B, Broich K, et al. Data rich, information poor: can we use electronic health records to create a learning healthcare system for pharmaceuticals? *Clin Pharmacol Ther* 2019;105:912–22.

2 Baumfeld Andre E, Reynolds R, Caubel P, et al. Trial designs using real-world data: the changing landscape of the regulatory approval process. *Pharmacoepidemiol Drug Saf* 2020;29:1201–12.

3 Yuan H, Ali MS, Brouwer ES, et al. Real-World evidence: what it is and what it can tell us according to the International Society for pharmacoepidemiology (IspE) comparative effectiveness research (CER) special interest group (SIG). *Clin Pharmacol Ther* 2018;104:239–41.

4 United States Food and Drug Administration. PDUFA VI: fiscal years 2018-2022. Available: https://www.fda.gov/ForIndustry/UserFees/PrescriptionDrugUserFee/ucm446608.htm [Accessed 9 Jul 2020].

5 Senate and House of Representatives of the United States of America. 21st century cures act. Available: https://www.congress.gov/114/plaws/publ255/PLAW-114publ255.pdf [Accessed 9 Jul 2020].

6 European Medicines Agency. Adaptive pathways. Available: https://www.ema.europa.eu/en/human-regulatory/research-development/adaptive-pathways

7 United States Food and Drug Administration. Submitting documents utilizing real-world data and real-world evidence to FDA for drugs and biologics. Available: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/submitting-documents-utilizing-real-world-data-and-real-world-evidence-fda-drugs-and-biologics

8 European Medicines Agency. HMA-EMA joint big data Taskforce—summary report. Available: https://www.ema.europa.eu/documents/minutes/hma/ema-joint-task-force-big-data-summary-report_en.pdf [Accessed 9 Jul 2020].

9 Quigley JM, Thompson JC, Halfpenny NJ, et al. Critical appraisal of nonrandomized studies-A review of recommended and commonly used tools. *J Eval Clin Pract* 2019;25:44–52.

10 Sanderson S, Tatt ID, Higgins JPT. Tools for assessing quality and susceptibility to bias in observational studies in epidemiology: a systematic review and annotated bibliography. *Int J Epidemiol* 2007;36:666–76.

11 Katrak P, Bialocerkowski AE, Massy-Westropp N, et al. A systematic review of the content of critical appraisal tools. *BMC Med Res Methodol* 2004;4:22.

12 Zeng X, Zhang Y, Kwong JSW, et al. The methodological quality assessment tools for preclinical and clinical studies, systematic review and meta-analysis, and clinical practice guideline: a systematic review. *J Evid Based Med* 2015;8:2–10.

13 Morton SC, Costlow MR, Graff JS, et al. Standards and guidelines for observational studies: quality is in the eye of the beholder. *J Clin Epidemiol* 2016;71:3–10.

14 Page MJ, McKenzie JE, Higgins JPT. Tools for assessing risk of reporting biases in studies and syntheses of studies: a systematic review. *BMJ Open* 2018;8:e019703.

15 Neyarapally GA, Hammad TA, Pinheiro SP, et al. Review of quality assessment tools for the evaluation of pharmacoepidemiological safety studies. *BMJ Open* 2012;2:e001362.

16 Sarri G, Patorno E, Yuan H, et al. Framework for the synthesis of non-randomised studies and randomised controlled trials: a guidance on conducting a systematic review and meta-analysis for healthcare decision making. *BMJ Evid Based Med* 2020. doi:10.1136/bmjebm-2020-111493. [Epub ahead of print: 09 Dec 2020].

17 Liberati A, Altman DG, Tetzlaff J, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. *BMJ* 2009;339:b2700.

18 Hasson F, Keeney S, McKenna H. Research guidelines for the Delphi survey technique. *J Adv Nurs* 2000;32:1008–15.

19 School of Population Health. EPIQ (effective practice, informatics and quality improvement). faculty of medical and health sciences, University of Auckland, 2019. Available: https://www.fmhs.auckland.ac.nz/en/soph/about/our-departments/epidemiology-and-biostatistics/research/epiq.html [Accessed 9 Jul 2020].

20 Hong QN, Pluye P, bregues S F, et al. Mixed Methods Appraisal Tool (MMAT), version 2018. Registration of Copyright (#1148552), Canadian Intellectual Property Office, Industry Canada.

21 Critical Appraisal Skills Programme. CASP Cohort Study. Checklist [online], 2018. Available: https://casp-uk.net/casp-tools-checklists/ [Accessed 9 Jul 2020].

22 Specialist Unit for Review Evidence (SURE). Questions to assist with the critical appraisal of cohort studies, 2018. Available: http://www.cardiff.ac.uk/insrv/libraries/sure/checklists.html [Accessed 9 Jul 2020].

23 The Joanna Briggs Institute. System for the unified management of the review and assessment of information (SUMARI), 2017. Available: http://joannabriggs-webdev.org/research/critical-appraisal-tools.html [Accessed 9 Jul 2020].

24 Sterne JA, Hernán MA, Reeves BC, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. *BMJ* 2016;355:i4919.

25 Berger ML, Martin BC, Husereau D, et al. A questionnaire to assess the relevance and credibility of observational studies to inform health care decision making: an ISPOR-AMCP-NPC good practice Task force report. *Value Health* 2014;17:143–56.

26 Dreyer NA, Velentgas P, Westrich K, et al. The grace checklist for rating the quality of observational studies of comparative effectiveness: a tale of hope and caution. *J Manag Care Spec Pharm* 2014;20:301–8.

27 The National Institutes of Health and The National Heart, Lung, and Blood Institute (NIH-NHLBI). Study quality assessment tools. Available: https://www.nhlbi.nih.gov/health-topics/study-quality-assessment-tools [Accessed 9 Jul 2020].

28 Health Evidence Bulletin. *Questions to assist with the critical appraisal of an observational study eg cohort, case-control, cross-sectional*. Wales: HEB, 2014.

29 Kim SY, Park JE, Lee YJ, *et al*. Testing a tool for assessing the risk of bias for nonrandomized studies showed moderate reliability and promising validity. *J Clin Epidemiol* 2013;66:408–14.

30 Viswanathan M, Berkman ND, Dryden DM. *AHRQ methods for effective health care. assessing risk of bias and confounding in observational studies of interventions or exposures: further development of the RTI item bank*. Rockville, MD: Agency for Healthcare Research and Quality (US), 2013.

31 et alWells GA, Shea B, O'Connell D. The Newcastle-Ottawa scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses. Available: http://www.ohri.ca/programs/clinical_epidemiology/oxford.asp [Accessed 9 Jul 2019].

32 Scottish Intercollegiate Guidelines Network. Available: https://www.sign.ac.uk/checklists-and-notes.html [Accessed 17 Oct 2019, 9 Jul 2020].

33 University of Montreal. Critical appraisal Worksheet. University of Montreal, 2011. Available: https://guides.bib.umontreal.ca/uploads/uploads/original/critical-appraisal-worksheet.pdf?1296211861 [Accessed 9 Jul 2020].

34 The STROBE Statement. Strengthening the reporting of observational studies (cohort, case-control, and cross-sectional), 2007. Available: https://www.strobe-statement.org/index.php?id=strobe-aims [Accessed 9 Jul 2020].

35 Public Health, Effective Public Health Practice Project (EPHPP). Effective public healthcare panacea project. Quality assessment tool for quantitative studies. Available: https://link.springer.com/content/pdf/bbm%3A978-3-319-17284-2/1 [Accessed 9 Jul 2020].

36 Des Jarlais DC, Lyles C, Crepaz N, *et al*. Improving the reporting quality of nonrandomized evaluations of behavioral and public health interventions: the trend statement. *Am J Public Health* 2004;94:361–6. doi:10.2105/AJPH.94.3.361

37 Margetts BM, Vorster HH, Venter CS. Evidence-Based nutrition—review of nutritional epidemiological studies. *South African J Clin Nutr* 2002;15:68–73.

38 Zaza S, Wright-De Agüero LK, Briss PA, *et al*. Data collection instrument and procedure for systematic reviews in the guide to community preventive services. Task force on community preventive services. *Am J Prev Med* 2000;18:44–74.

39 Downs SH, Black N. The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. *J Epidemiol Community Health* 1998;52:377–84.

40 Elwood JM. *Critical appraisal of epidemiological studies and clinical trials*. Oxford University Press: Oxford, 1998.

41 Hadorn DC, Baker D, Hodges JS, *et al*. Rating the quality of evidence for clinical practice guidelines. *J Clin Epidemiol* 1996;49:749–54.

42 Ashby J, Carlo G, Cohen SM. *Principles for evaluating epidemiologic data in regulatory risk assessment. federal focus, incorporated*. Washington, DC, 1996. http://www.fedfocus.org/science/london-principles.html

43 DuRant RH. Checklist for the evaluation of research articles. *J Adolesc Health* 1994;15:4–8.

44 Levine M, Walter S, Lee H, *et al*. Users' guides to the medical literature. IV. How to use an article about harm. evidence-based medicine Working group. *JAMA* 1994;271:1615–9.

45 Gyorkos TW, Tannenbaum TN, Abrahamowicz M, *et al*. An approach to the development of practice guidelines for community health interventions. *Can J Public Health* 1994;85(Suppl 1):S8–13.

46 Cho MK, Bero LA. Instruments for assessing the quality of drug studies published in the medical literature. *JAMA* 1994;272:101–4.

47 Centre for Occupational and Environmental Health. Critical appraisal. School of epidemiology and health sciences, University of Manchester, 2003. Available: http://research.bmh.manchester.ac.uk/epidemiology/COEH/undergraduate/specialstudymodules/criticalappraisal/ [Accessed 9 Jul 2020].

48 Fowkes FG, Fulton PM. Critical appraisal of published research: introductory guidelines. *BMJ* 1991;302:1136–40.

49 Lichtenstein MJ, Mulrow CD, Elwood PC. Guidelines for reading case-control studies. *J Chronic Dis* 1987;40:893–903.

50 Gardner MJ, Machin D, Campbell MJ. Use of check Lists in assessing the statistical content of medical studies. *Br Med J* 1986;292:810–2.

51 Horwitz RI, Feinstein AR. Methodologic standards and contradictory results in case-control research. *Am J Med* 1979;66:556–64.

52 Campbell JD, Perry R, Papadopoulos NG, *et al*. The real life evidence assessment tool (relevant): development of a novel quality assurance asset to rate observational comparative effectiveness research studies. *Clin Transl Allergy* 2019;9:21.

53 Suissa S. Immortal time bias in pharmaco-epidemiology. *Am J Epidemiol* 2008;167:492–9.

54 Bykov K, He M, Franklin JM, *et al*. Glucose-Lowering medications and the risk of cancer: a methodological review of studies based on real-world data. *Diabetes Obes Metab* 2019;21:2029–38.

55 Sørensen HT, Lash TL, Rothman KJ. Beyond randomized controlled trials: a critical comparison of trials with nonrandomized studies. *Hepatology* 2006;44:1075–82.

56 Schneeweiss S, Avorn J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. *J Clin Epidemiol* 2005;58:323–37.

57 Schünemann HJ, Cuello C, Akl EA, *et al*. Grade guidelines: 18. How ROBINS-I and other tools to assess risk of bias in nonrandomized studies should be used to rate the certainty of a body of evidence. *J Clin Epidemiol* 2019;111:105–14.