Drug Safety (2023) 46:87–97
https://doi.org/10.1007/s40264-022-01254-4

ORIGINAL RESEARCH ARTICLE

Development and Evaluation of the Algorithm CErtaInty Tool (ACE-IT) to Assess Electronic Medical Record and Claims-based Algorithms' Fit for Purpose for Safety Outcomes

Sonal Singh[1,6] · Julie Beyrer[2] · Xiaofeng Zhou[3] · Joel Swerdel[4] · Raymond A. Harvey[4] · Kenneth Hornbuckle[2] · Leo Russo[3] · Kanwal Ghauri[5] · Ivan H. Abi-Elias[6] · John S. Cox[6] · Carla Rodriguez-Watson[5]

Accepted: 25 October 2022 / Published online: 17 November 2022

Abstract

Introduction Electronic health record (EHR) or medical claims-based algorithms (i.e., operational definitions) can be used to define safety outcomes using real-world data. However, existing tools do not allow researchers and decision-makers to adequately appraise whether a particular algorithm is fit for purpose (FFP) to support regulatory decisions on drug safety surveillance. Our objective was to develop a tool to enable regulatory decision-makers and other stakeholders to appraise whether a given algorithm is FFP for a specific decision context.

Methods We drafted a set of 77 generic items informed by regulatory guidance documents, existing instruments, and publications. The outcome of ischemic stroke served as an exemplar to inform the development of draft items. The items were designed to be outcome independent. We conducted a three-round online Delphi panel to develop and refine the tool and achieve consensus on items (> 70% agreement) among panel participants composed of regulators, researchers from pharmaceutical organizations, academic clinicians, methodologists, pharmacoepidemiologists, and cardiologists. We conducted a qualitative analysis of panel responses. Five pairs of reviewers independently evaluated two ischemic stroke algorithm validation studies to test its application. We developed a user guide, with explanation and elaboration for each item, guidance on essential and additional elements for user responses, and an illustrative example of a complete assessment. Furthermore, we conducted a 2-h online stakeholder panel of 16 participants from regulatory agencies, academic institutions, and industry. We solicited input on key factors for an FFP assessment, their general reaction to the Algorithm CErtaInty Tool (ACE-IT), limitations of the tool, and its potential use.

Results The expert panel reviewed and made changes to the initial list of 77 items. The panel achieved consensus on 38 items, and the final version of the ACE-IT includes 34 items after removal of duplicate items. Applying the tool to two ischemic stroke algorithms demonstrated challenges in its application and identified shared concepts addressed by more than one item. The ACE-IT was viewed positively by the majority of stakeholders. They identified that the tool could serve as an educational resource as well as an information-sharing platform. The time required to complete the assessment was identified as an important limitation. We consolidated items with shared concepts and added a preliminary screen section and a summary assessment box based on their input. The final version of the ACE-IT is a 34-item tool for assessing whether algorithm validation studies on safety outcomes are FFP. It comprises the domains of internal validity (24 items), external validity (seven items), and ethical conduct and reporting of the validation study (three items). The internal validity domain includes sections on objectives, data sources, population, outcomes, design and setting, statistical methods, reference standard, accuracy, and strengths and limitations. The external validity domain includes items that assess the generalizability to a proposed target study. The domain on ethics and transparency includes items on ethical conduct and reporting of the validation study.

Conclusion The ACE-IT supports a structured, transparent, and flexible approach for decision-makers to appraise whether electronic health record or medical claims-based algorithms for safety outcomes are FFP for a specific decision context. Reliability and validity testing using a larger sample of participants in other therapeutic areas and further modifications to reduce the time needed to complete the assessment are needed to fully evaluate its utility for regulatory decision-making.

Extended author information available on the last page of the article

Key Points

The Algorithm CErtaInty Tool (ACE-IT) provides a structured, transparent, flexible, and qualitative approach for decision-makers to appraise whether algorithms that validate drug safety outcomes are fit for purpose for a decision context.

Further testing across different therapeutic areas is needed to evaluate its utility for regulatory decision-making.

1 Introduction

Validation of algorithms based on electronic health records (EHRs) or medical claims-based definitions of outcomes (e.g., combination of medical and pharmacy codes) are important to evaluate the effectiveness and safety of marketed products using real-world data. The US Food and Drug Administration (FDA) underscores the importance of validating algorithms designed to represent clinical endpoints, including algorithms derived from EHRs or administrative claims [1].

There is a lack of a transparent and consistent approach to evaluate whether algorithms for safety outcomes are fit for purpose (FFP) to support regulatory decisions. Regulators use real-world studies using EHR or medical claims-based algorithms to make decisions about the safety and effectiveness of drugs. It is necessary for such algorithms to be valid and reliable. There is no single resource with criteria or questions to appraise whether an algorithm is FFP for different types of regulatory contexts of use. Existing tools to evaluate EHR or medical claims-based algorithms recommend evaluation of important aspects of internal validity but may lack specificity for evaluating other items relevant to the algorithm's use in studies designed to inform regulatory decisions [2]. Other tools such as the Quality Assessment of Diagnostic Accuracy Studies instrument [3] and the Standards for Reporting of Diagnostic Accuracy Studies tool [4], are specific for evaluating diagnostic studies but do not support the evaluation of algorithms for regulatory decisions. The Good ReseArch for Comparative Effectiveness (GRACE) checklist [5] and the Structured Template for Assessment and Reporting of Real-World Evidence (STaRT-RWE) template [6] identify important elements of observational studies. However, these two tools rarely assess whether or how the operational definition of the outcome reflects the underlying conceptual or case definitions. The integrity of the underlying data source and the replicability of algorithms between data sources are of special importance to regulators. Existing tools evaluate internal validity but do not include assessment of overall FFP including external validity for the algorithm's intended use or target application, which regulatory decision-making requires [7]. The Algorithm CErtaInty Tool (ACE-IT) was developed to meet this evidence gap.

2 Objectives and Scope

Our objective was to develop a tool to enable regulatory decision-makers and other stakeholders to appraise whether a given EHR or medical claims-based algorithm is FFP for a specific decision context. The primary objective of this tool is to enable regulatory decision-makers and other stakeholders, including manufacturers and healthcare professionals, to appraise the FFP nature of an algorithm for a target study.

3 Methods

3.1 Development of Draft Items for the Delphi Panel

We drafted a set of items to present to the panel. Seven experts on real-world evidence (RWE) algorithms from different organizations drafted the initial set of items, informed by existing tools to evaluate diagnostic studies [3, 4], tools to evaluate observational studies on RWE [5, 6], and relevant published literature on the topic [8–10], including regulatory documents from the FDA [7, 11], European Medical Agency [12–14], and government organizations such as the Agency for Health Care Research and Quality [15] and the Patient Centered Outcomes Research Institute [16]. These included items from a recent stakeholder review by some of our team members [10] and were also informed by algorithms identified from a scoping review on ischemic stroke.

The initial list of items was intended to stimulate discussion by the Delphi panel. It was intended to be comprehensive and inclusive of all items potentially relevant to the objectives, data sources, study design, accuracy, reference standards, and limitations of algorithm validation studies. We also included items related to the assessment of external validity to a target study. These items were intended to be generic and outcome independent. All the authors, representing different organizations and different areas of expertise, were involved in drafting the items.

3.2 Selection of Safety Outcome of Ischemic Stroke

We selected the safety outcome of ischemic stroke to motivate the development of draft items. We selected ischemic stroke because of its importance to regulators and other stakeholders, the heterogeneity of operational definitions for "ischemic stroke," and the availability of several validated algorithms. We had initially intended to evaluate algorithms on the composite endpoint of major adverse cardiovascular outcome, which comprises the endpoints of myocardial infarction, stroke, and cardiovascular death. However, most available algorithms reported on the individual endpoints rather than the composite.

3.3 Scoping Review for Ischemic Stroke

We conducted a scoping review to identify EHR or medical claims-based algorithms for ischemic stroke. The results of this review informed the development of draft items for the Delphi panel. We excluded studies that did not validate algorithms against a reference standard. We included studies that evaluated the diagnostic accuracy, i.e., sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) of International Classification of Disease (ICD)-9- or ICD-10-based administrative algorithms for ischemic stroke. We included studies that reported on any of these parameters. We excluded prognostic utility studies, randomized controlled trials, case series, review articles, and systematic reviews or studies with no original data.

3.4 Delphi Process and Consensus

The Delphi process synthesizes opinions of experts into group consensus [17, 18] through the provision of structured feedback and statistical group response, with the responses of each individual reflected anonymously in final ratings [19]. A reproducible method for selection of participants, prespecification of the definition of consensus, and prespecification of the number of rounds are key qualities to ensure a robust Delphi study [17]. The required level of consensus for keeping an item needs to be prespecified prior to each round. It has varied from as low as 50% agreement [20] to more than 80% agreement [21]. Percent agreement (e.g., > 80%) has been the most commonly used definition of consensus, followed by proportion of participants agreeing on a particular rating [17]. Given the small size of our expert panel and the need for a tool that prioritizes sensitivity of items over specificity, we selected > 70% agreement as the definition of consensus for the selection of the final round of items. We conducted three sequential rounds of review and evaluated the level of agreement and open-ended responses to decide which items to include in subsequent rounds. We combined items with similar concepts when possible. We planned to stop the Delphi process after three rounds regardless of the level of agreement or stability of ratings.

3.5 Identification and Recruitment of the Delphi Panel

We considered the following areas of expertise critical to our panel: (1) regulatory expertise from the FDA Office of Surveillance and Epidemiology as well as the Office of Medical Policy; (2) cardiovascular methodology; (3) development and validation of algorithms in large databases; and (4) causal inference and pharmacoepidemiology. We also identified experts from pharmaceutical sponsors who conduct validation studies. We identified experts from relevant networks, membership of societies (e.g., International Society of Pharmacoepidemiology), and relevant publications on the topic. All were invited to participate by email. All but one expert agreed to participate. The expert who refused to participate identified another relevant expert from his/her perspective. The final seven-member panel included two regulators, four members from academia, including a cardiovascular epidemiologist, cardiologists, and pharmacoepidemiologists, and regulatory representatives. One member from industry provided expertise in the conduct of validation studies.

3.6 First Round of the Delphi Panel

We generated items for the first round based on domains and items informed by the tools and guidance documents noted above. To assist the panelists, we also presented a hypothetical vignette for a proposed safety study of a hypothetical product for ischemic stroke for regulatory surveillance (see Table 1). We presented the results of the draft tool to the panel after a 90-min online orientation session. The panel rated the importance of the draft items and proposed additional items independently and anonymously to reduce the risk of bias. The tool included items relevant to internal validity, external validity, ethics, and transparency of reporting. The panelists provided categorical responses to whether the selected items should be included to accurately identify the construct of interest (internal validity) or support the use of the algorithm in a target, i.e., safety surveillance study (external validity). They also provided open ended responses about additional items to evaluate algorithm validation studies. There was no limit to the number of items that could be generated. We conducted a qualitative analysis and grouped similar items together, removed redundant

Table 1 Vignette

Preclinical studies suggest an excess of ischemic stroke with hypothetical product X. A post-marketing real-world evidence study using the Food and Drug Administration's (FDA's) Best Practices for Conducting and Reporting Pharmacoepidemiologic Safety Studies Using Electronic Healthcare Data Sets (2013) is requested to rule out a relative risk ratio > 1.3 of ischemic stroke with product X. The sponsor would like to identify reliable, valid, and robust algorithms that are fit for purpose to identify ischemic stroke for this post-marketing safety study. If the algorithms are not fit for purpose, the sponsor would like to design a reliable and valid algorithm validation study for ischemic stroke. Criterion validity is considered most relevant for this scenario

items, and included new items generated by the panel in the participants' words as much as possible. We clarified concepts and terms and reorganized the structure of the tool (e.g., moving items from one section to another). The items were structured in the form of sufficiently complex questions for the next round to minimize ceiling effects.

3.7 Second Round of the Delphi Panel

In the second round, the panel rated the revised set of items on a 5-level scale (strongly disagree, disagree, neutral, agree, or strongly agree) to determine agreement for inclusion of items. A rating of strongly agree signified that item was important and needed to be included in the tool, whereas a rating of strongly disagree indicated that the item was not important and should not be included. We quantified the level of agreement and disagreement among all rated items. As prespecified, all items that were rated as being agree or strongly agree (≥ 4) by >70% of the panel were retained for the next round of the survey [21].

3.8 Third Round of the Delphi Panel

The objective of the third round of the Delphi panel was to achieve consensus on the final list of items and stability of ratings, if possible. The panel reviewed and rated the items using similar methods as round 2 above. All items rated as being agree or strongly agree (≥ 4) by > 70% of the panel were retained for the final tool [21].

3.9 Operationalizing the ACE-IT

Five pairs of experts in clinical epidemiology, pharmacoepidemiology, and validation independently used the final version of the tool to evaluate two selected ischemic stroke studies from the final list of included studies [22, 23]. There was no overlap between members of the Delphi panel and these experts. We selected two studies on ischemic stroke for appraisal for pragmatic considerations [22, 23]. These studies were selected from the final list of included studies by the study team as they were deemed appropriate for further testing of the ACE-IT. These two studies provided sufficient heterogeneity in data sources, population, and reporting of accuracy metrics. One study measured the PPV of acute ischemic stroke among intravenous immunoglobin users in the Sentinel Distributed Network [22]. The other study validated ischemic stroke among women in Medicare using the Women's Health Initiative Cohort [23]. One study reported precision around accuracy parameters [22], whereas the other study did not report any precision around these estimates [23]. One study was conducted in an administrative claims database and validated stroke outcomes against chart reviews [22]. The other study was conducted among Medicare participants and validated stroke outcomes compared to stroke diagnoses in the Women's Health Initiative Study [23].

The objective of this evaluation was to clarify ambiguous terms and concepts. When items included more than one component, (e.g., an item on the data source included considerations on data cleaning and data quality), raters made their best judgment and provided a rationale for their ratings. Raters used the above vignette to guide their work and assess the FFP of the two validation studies to a target data source (IQVIA PharMetrics® Plus database) (January 2010–September 2019) for the target study [24]. Using the hypothetical vignette shown in Table 1, the raters were asked to evaluate the internal validity, external validity, and reporting of the two ischemic stroke algorithms to the target study population [24]. This rigorous evaluation by a team of experts allowed us to test the performance of the tool.

3.10 Evaluation by the Stakeholder Panel

We convened 16 independent experts in the field for a 2-h virtual meeting via Zoom. There was one member of the Delphi panel, who provided the regulatory perspective, who also participated in the stakeholder panel. The members in attendance included researchers from academic institutions, representatives from pharmaceutical companies, healthcare technology companies, and the FDA. Participants' expertise spanned epidemiology, pharmacoepidemiology, regulatory, and RWE analytics. The meeting was recorded and transcribed for note taking purposes and followed a semi-formal guide moderated by a member that was not part of the core research team. Stakeholders were provided the tool and

user guide prior to the meeting for review and were asked to provide their input on four key questions: (1) What factors affect a decision about whether a given RWE algorithm is FFP? (2) What is your general reaction to the ACE-IT? (3) What are the limitations of the tool? and (4) How do you see you or your organization using this tool? We transcribed their comments and summarized their responses to the above questions with representative quotes.

3.11 Development of the ACE-IT and User Guide

We developed the final version of the tool, which included an explanation and elaboration document. Each item included in the tool was followed by an explanation and elaboration section, which included the rationale for inclusion of the item and the essential and additional elements needed to inform the assessment. We distinguished between elements that were essential to assess the validity of the study findings versus additional elements that may enhance study credibility and may only be applicable in certain scenarios. We also included a verbatim excerpt from an example validation study followed by an example of the assessment by a user.

3.12 Data Collection and Analysis

The tool was completed via REDCap, a secure data management system [25, 26]. We assessed percent agreement in rounds 2 and 3.

4 Results

Overview: The ACE-IT is a 34-item tool for assessing whether algorithm validation studies on safety outcomes are FFP. This is shown in Table 2. It comprises the domains of *internal validity* (24 items), *external validity* (seven items) and *ethical conduct and reporting of the validation study* (three items). The domain of internal validity includes sections on *objectives*, *population*, *design and setting*, *data sources*, *reference standard*, *strengths and limitations*, *outcomes*, *statistical methods*, and *accuracy*. The items refer to individual items within each of the respective sections and/or domains.

4.1 Delphi Panel

Our search identified eight studies on ischemic stroke [22, 23, 27–32]. This is shown in Supplementary File 1 (see the electronic supplementary material). A draft set of 77 items informed by regulatory guidance documents, existing instruments, scoping review, and publications was presented to the Delphi panel. This is shown in Supplementary File 2. After revision and input from the Delphi panel in the first round, the revised tool for the second Delphi panel included 47 items. The panel achieved consensus on 38 items after the Delphi round 2. We presented 38 items for the third round. The rating remained stable from the second and third round of the Delphi panel. The panel achieved consensus on all 38 items. Testing the application of the tool by the pairs of reviewers resulted in the identification of four overlapping items. Among these 38 items, four items were not incorporated into the 34-item ACE-IT because of overlapping themes.

4.2 Stakeholder Assessment

The stakeholders identified that an FFP algorithm should always start with a clearly defined research question. The stakeholders noted that such algorithms should ideally be relevant, reliable, valid, clearly outline the population of interest, and be derived from a reliable data source [7, 10]. They also noted that it should have high PPV and sensitivity but acknowledged that tradeoffs between PPV and sensitivity may be necessary. They also identified the need for transparency and replicability. Such an algorithm should clearly outline the population of interest and be derived from a reliable data source. It should also have the ability to account for bias and misclassification. The majority of participants approved of the structured, flexible, and systematic approach to an assessment of the algorithm outlined by the ACE-IT. The majority of participants agreed that a qualitative assessment rather than a quantitative score was appropriate. They identified that the tool could serve as an educational resource as well as an information-sharing platform within their organization.

The stakeholder panel also identified some limitations. A few stakeholders identified the time required to complete the assessment as an important limitation. Stakeholders suggested that assessors identify the critical components for their assessment before using the tool. They suggested that a full assessment of the algorithm's FFP was unnecessary if critical components were absent from the validation study based on a preliminary screen. A few stakeholders advised that the tool should also have a segment to summarize the overall assessment of FFP. In response to these comments, we consolidated items with shared concepts and added a preliminary screen section and a summary assessment box based on their input.

Another limitation noted was the inability to assess validity due to the limitations of study reporting versus the actual study design and conduct of the algorithm validation study. Detailed results of stakeholder assessment with representative quotes are shown in Supplementary File 3 (see the electronic supplementary material).

Table 2 Domains and Items in the Algorithm CErtaInty Tool (ACE-IT)

Domain	Section	Item #	Item
Internal validity	Objectives and scope	1	Were the objectives and scope of the index algorithm validation study clearly stated and adequately detailed?
	Data source	2	Are the index algorithm validation data sources adequately described and relevant and reliable to validate the algorithm (e.g., type of data source, validation data source quality assurance and control, data cleaning, transformation procedures including validated data linkages, and the impact of missing and miscoded data or data lag on completeness of data)?
	Study population	3	Is the number of excluded participants, the reasons for exclusion for study participants in the index validation data source, and its influence sufficiently documented in the index validation study?
		4	Does the sampling approach in the index validation study support obtaining unbiased estimates of accuracy measures for the outcome?
		5	Does the index algorithm validation study design fully describe the relevant study design and cohort(s) for measuring accuracy based on the statistical parameters to be estimated for the outcome?
	Study outcomes	6	Are the conceptual/case definitions of the outcomes clearly defined and described?
		7	Are the operational definitions (or the index algorithm) of the outcomes of interest clearly defined and described?
		8	Did the index algorithm validation study report whether the outcome was incident or prevalent?
		9	Was the index algorithm defined independently from the exposure?
	Study design and setting	10	Does the index validation study report on patient characteristics relevant to the study where the algorithm will be used (e.g., age, sex, race/ethnicity, geographic region, disease severity, and comorbidities)?
		11	Is the time period of the index validation study appropriate for measuring the outcome in the target study?
	Statistical methods	12	Are the statistical methods appropriate and adequately described for validating the index algorithm?
		13	Do the index validation study methods adequately describe and justify the estimation of sample size to the desired level of precision? Will the sample size enable detecting a clinically meaningful difference if comparing algorithms?
		14	Do the index validation study methods adequately report how confidence intervals were estimated?
		15	Does the study report on relevant accuracy parameters and data to support quantitative bias analysis to assess for misclassification of outcomes when applicable?
	Limitations of the validation study	16	Are the algorithm limitations, including misclassification of subgroups, the impact of missing data, and the possible magnitude, direction, and uncertainty of bias (e.g., channeling bias and immortal time bias) reported and addressed?
	Reference standard	17	Is the reference standard relevant and appropriate to identify the outcome?
		18	Was the reference standard independent of the index algorithm?
		19	Were the diagnostic thresholds in the reference standard adequately justified and documented to comprehensively identify the outcome?
		20	Did the study describe whether the assessment of the reference standard was conducted by a single assessor or more than one assessor?
		21	Was the reference standard measured in a consistent manner using standardized protocol and criteria for all participants? Was measurement error, changes in the measurement method(s) over time, and quality control procedures described and adequately justified?

Table 2 (continued)

Domain	Section	Item #	Item
	Accuracy	22	Are the estimated accuracy parameters (e.g., sensitivity, specificity, PPV, and NPV) appropriately justified for the index algorithm validation study?
		23	Is the prevalence of outcome reported in the validation data source when PPV is reported in the validation study?
		24	Does the index algorithm validation study report accuracy parameters stratified by important variables relevant to its intended use (e.g., important subgroups such as age, sex, race/ethnicity, geographic region, site [if multi-site study across various databases], disease severity, and comorbidities)?
External validity		25	Is the algorithm sufficiently generalizable to the target data source and population based on characteristics of the target study data source (healthcare delivery or insurance data source, health system characteristics, population characteristics, time period, prevalence of outcome in the target data source)?
		26	Is the index validation study location and setting (inpatient, outpatient, emergency, and other settings) clearly described and appropriate for measuring the outcome in the target population?
		27	Is the algorithm sufficiently generalizable to the target data source and population based on assessment of measurement error?
		28	Is the measurement of outcome consistent within and between validation and target data sources?
		29	Does the validation study report details on algorithm specifications, timeframes including index and study periods, code lists, and development history including modifications/updates to enable replication in the target data source?
		30	Are adequate data available for the selected timeframe, including accounting for data lag in measurement of outcomes, in the target data source?
		31	Is the index algorithm performance on measures of validity similar to results from other data sources?
Ethics and transparency		32	Does the validation study disclose the funding source?
		33	Was the validation conducted according to a prespecified protocol, which is also registered, and changes to the study documented in a protocol amendment?
		34	Did the validation study authors adhere to privacy, ethical, and regulatory requirements for study conduct?

NPV negative predictive value, *PPV* positive predictive value

5 ACE-IT

The final items included in the ACE-IT are shown in Table 2. The ACE-IT and user guide are shown in Supplementary File 4 (see the electronic supplementary material). It comprises the domains of *internal validity* (24 items), *external validity* (seven items), and *ethical conduct and reporting of the validation study* (three items). The domain of internal validity includes sections on *objectives*, *population*, *design and setting*, *data sources*, *reference standard*, *strengths and limitations*, *outcomes*, *statistical methods*, and *accuracy*. The items refer to individual items within each of the respective sections and/or domains. The external validity domain focuses on assessment of the feasibility and generalizability of re-using the algorithm in the target study (i.e., the ischemic stroke target study described above in the vignette). The domain on ethics and transparency contains items related to the validation study including prespecification of the study protocol, disclosure of funding, and meeting privacy, ethical, and regulatory requirements. The excerpt from the validation study followed by an example of the assessment by a user are shown in Supplementary File 5. The glossary of terms used throughout the document are shown in Supplementary File 6.

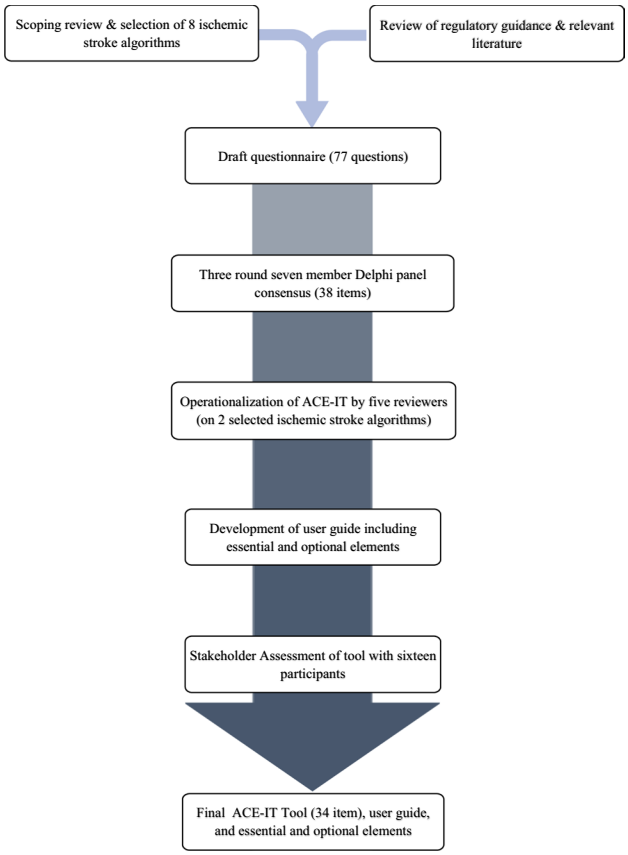

Fig. 1 Overview of development of the ACE-IT tool

5.1 Using the ACE-IT

To determine whether a full ACE-IT assessment may be warranted, the assessors should screen the validation study for a clear conceptual definition and the presence of relevant accuracy metrics, and a relevant study population. Although the specific combination of items will vary, the lack of a clear conceptual definition (e.g., there may be no standard clinical definition for long coronavirus disease [COVID]), absence of reporting of relevant accuracy parameters in the validation study necessary for the target study, and validation in a study population or data source whose findings are not generalizable to the target study population or data source may preclude the need for a full assessment.

If the preliminary screen indicates that the study meets these minimum requirements for the FFP assessment, the assessors should identify the key items relevant for their FFP assessment a priori before answering the other items in the ACE-IT. The assessors should have access to the validation study results, any supplementary appendices, protocol, underlying data sources, and adjudication methods. We recommend that two assessors with complementary expertise evaluate the study independently and consider adjudicating any differences. Users of the tool also need sufficient details on the target data source and study design to enable adequate appraisal of the section on external validity.

6 Discussion

6.1 Statement of Principal Findings

The ACE-IT supports a structured, transparent, and flexible approach for decision-makers to appraise whether EHR or medical claims-based algorithms of safety outcomes are FFP to evaluate an outcome in the context of a safety surveillance study. We provide examples and explanations along with essential and additional elements to facilitate use of the ACE-IT. Understanding the rationale for ratings on the items in the tool will improve the ability of users to interpret the strengths and weaknesses of the algorithm. A high degree of adherence to items included in the ACE-IT likely reflects the transparency and credibility of reported algorithms.

The ACE-IT is intended to support and inform an FFP assessment of the use of an algorithm by highlighting its strengths and limitations for a specific decision context. However, assessment of FFP using the ACE-IT cannot substitute for careful contextual considerations. Although not listed as a specific item in the ACE-IT, the ACE-IT encourages the incorporation of contextual considerations into the assessment of FFP. In advance of the assessment, the assessors should identify important contextual considerations. At the end, the summary assessment should also account for any contextual considerations. An example of such contextual considerations could include the broader impact on the target patient population, including the reasonably anticipated benefits and risks [33]. An example of such considerations may include an assessment of the importance of the safety outcome and the expected clinical benefits and risks of a drug. The assessment of FFP using the ACE-IT requires an examination of both the methods and results of validation. The end user can use the ratings, the presence or absence of elements, and the rationale provided by an assessment of the study, along with other contextual considerations, to complete their assessment.

We propose that end users consider the ACE-IT as an aid to inform and support a qualitative, rather than a quantitative, approach to FFP assessment of an algorithm. A quantitative assessment may be intuitively appealing, but applying quantitative weights to items of varying and unequal importance is methodologically flawed. The user of this tool will need to decide how important each item is to justify the algorithm's FFP. As an example, for some target studies, the user may choose to assign higher importance to the similarity of the data sources between the index validation study and the target study; for other studies, the user may

assign the highest importance to the presence of all accuracy parameters for bias analysis.

6.2 Strengths and Limitations

We believe that the structured and transparent approach to assessing an algorithm using the ACE-IT may offer some advantages to an FFP assessment as compared to the currently prevailing unstructured approach to decision-making. Certain important elements are ignored using an unstructured approach.

Our tool overcomes some limitations of existing reporting checklists for evaluating diagnostic algorithms [2] and other checklists for observational studies [5] by including unique and important items that allow for the assessment of FFP of algorithms for the regulatory decision context. These items include item #6, item #7, item #9, item #13, item #15, item #16, and item #19 in the internal validity domain in the ACE-IT. These items emphasize the need for clearly specifying both the conceptual and operational definition and the need for algorithm validation studies to provide sufficient data to support quantitative bias analysis and address various biases. The items on adequate sample size and power are also important. Finally, by including an additional domain on external validity, our tool allows the assessment of FFP for a decision context. The domain on ethics and reporting in the ACE-IT encourages transparency. However, as opposed to the objective of the ACE-IT to evaluate whether an algorithm validation study is FFP, the objective of the GRACE checklist is broader, with the goal to evaluate the quality of observational studies, and validation represents a single component of that evaluation.

The development of the ACE-IT should be viewed within the larger context of FDA efforts to expand the use of RWE, with impetus provided by the *21st Century Cures Act* and *Prescription Drug User Fee Act* (PDUFA). Although the items in this tool were derived from existing regulatory guidance and other best practice considerations, users should consider that using this tool to evaluate algorithm validation studies does not obviate the need for being aware and adhering to these best practice considerations [5, 7, 9, 11, 34].

Although our tool was primarily designed for the regulatory decision context, other users, such as reviewers of articles, may also find this tool useful to evaluate the credibility and transparency of algorithms and to justify or refute an algorithm's FFP.

Our tool has certain limitations. The selection of a single outcome of ischemic stroke and evaluation of a limited set of studies may have limited the scope of items that could inform the tool. We selected the outcome of ischemic stroke before the development of the tool to inform the items for the Delphi panel. Although the prior knowledge of ischemic stroke as the outcome of interest may have the potential to introduce bias, this is likely minimal. We deliberately designed the items to be outcome independent, and the items in the ACE-IT are intended to be generic. The assessment of only two studies by experts may also have the potential to introduce bias because these may not be fully representative of the spectrum of included studies. Future reliability and validity testing on other effectiveness and safety outcomes, covariates, or cohorts may result in refinement to the ACE-IT.

Another limitation was the small size of our Delphi panel. However, the panel had sufficient expertise with diverse perspectives. The definition of consensus is somewhat arbitrary with no clear guidance and is context dependent. We did not report on measures of inter-rater reliability because we anticipated such differences. Quantitative measures of inter-rater reliability may not be as helpful as understanding the differences in ratings and the rationale for their ratings. These differences resulted in further clarification of the items. Some assessors noted that a substantial amount of time was needed to complete the assessment.

Another limitation was the inability to assess validity due to the limitations of study reporting versus the design and conduct of the algorithm validation study. However, this represents a limitation of the algorithm validation study rather than the ACE-IT and can be construed as a strength. The tool could stimulate future validation studies to improve their reporting of items in the ACE-IT. It is possible that some studies may not have reported all relevant details (e.g., data sources) due to limitations on word counts in manuscripts. Assessors should attempt to gather all relevant references when applicable. The availability of online supplements should make it easier for future algorithm validation studies to address this limitation (Fig. 1).

6.3 Future Research

Future studies should explore whether one can arrive at an overall FFP assessment using the ACE-IT after accounting for careful contextual considerations. It would be important to explore whether such an assessment can be used to categorize studies as being FFP based on levels of certainty, for example, as either *optimal*, *sufficient*, or *probable* [9]. We developed the tool using an example of claims-based algorithms for a safety endpoint for ischemic stroke; however, the goal is for the tool to be used more broadly across other therapeutic areas. We should also explore whether the ACE-IT needs to be modified to evaluate algorithms for outcomes across other therapeutic areas and for different contexts of use or types of studies, such as effectiveness outcome measures for external control arms or predictive validation for surrogate outcomes. Studies should also explore

the performance of assessment of algorithms by the ACE-IT versus those of expert assessments, other external standards, or correlation or prediction with events.

7 Conclusions

The ACE-IT supports a structured, transparent, and flexible approach for decision-makers to appraise whether EHR or medical claims-based algorithms for safety outcomes are FFP. Reliability and validity testing using a larger sample of participants in other therapeutic areas and further modifications to reduce the time needed to complete the assessment are needed to fully evaluate its utility for regulatory decision-making.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s40264-022-01254-4.

Acknowledgements The authors are very grateful to all the experts who participated in the Delphi process, Stakeholder meeting, and review process, including the following: Jeff Allen, PhD; Noelle Cocoros, DSc, MPH; John Concato, MD, MPH; Efe Eworuke, PhD, MSc; Adrian Hernandez, MD; Stephan Lanes, PhD, MPH; Charles Leonard, PharmD, MSCE (Delphi participants); Susan Andrade, ScD; Sigrid Behr, PhD; Ulka Campbell, PhD, MPH; Jessica Chubak, PhD; Efe Eworuke, PhD, MSc; Nicole Gatto, PhD, MPH; Jamie Geier, PhD, MPH; Rosa Gini, PhD; Nirosha Lederer, PhD, MSPH; Nancy Lin, PhD; Kenneth Quinto MD, MPH; Donna R. Rivera, PharmD, MSc; John D. Seeger, PharmD, MPH, DrPH, FISPE; Sengwee Toh, ScD; Shirley V Wang, PhD, ScM (Stakeholder meeting attendees); Sangmi Kim, PhD; David R. Nelson, MS (Reviewers); and other stakeholders (anonymity retained).

Declarations

Funding This activity is one part of a multi-part Foundation project supported by the Food and Drug Administration (FDA) of the U.S. Department of Health and Human Services (HHS) as part of an award of $50,000 of federal funds (25% of the project) and by $150,000 from non-governmental sources (75% of the project). The contents are those of the author(s) and do not necessarily represent the official views of, nor an endorsement, by FDA, HHS, or the U.S. Government. For more information, please visit FDA.gov.

Conflict of interest Sonal Singh was supported by a grant from the Reagan Udall Foundation for the FDA to UMass Chan Medical School for the conduct of this study. Julie Beyrer and Kenneth Hornbuckle are employees and shareholders of Eli Lilly and Company. Xiaofeng Zhou and Leo Russo are employees and shareholders of Pfizer Inc. Raymond A. Harvey is employed by Janssen, a Johnson and Johnson company that participated in the funding of this research and is a shareholder of Johnson and Johnson. Joel Swerdel is an employee and shareholder of Johnson and Johnson. Kanwal Ghauri, Ivan H. Abi-Elias, John S. Cox, and Carla Rodriguez-Watson have no conflicts of interest that are directly relevant to the contents of the study.

Ethics approval The UMass Chan Medical School Institutional Review Board designated the study as not Human Subject Research.

Consent to participate Not applicable.

Consent for publication Not applicable.

Availability of data and material All data generated during the study are provided as electronic supplementary material. No additional datasets were generated or analyzed during the current study.

Code availability Not applicable.

Author contributions All authors contributed to the study conception and design. Material preparation, data collection, and analysis were performed by SS, JB, XZ, JS, RAH, KH, LR, KG, IHA-E, JSC, and CR-W. The first draft of the manuscript was written by SS, and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript for publication.

References

1. U.S. Department of Health and Human Services, Food and Drug Administration Center for Drug Evaluation and Research (CDER) Center for Biologics Evaluation and Research (CBER) Oncology Center of Excellence (OCE). Guidance for Industry and FDA Staff Best Practices for Conducting and Reporting Pharmacoepidemiologic Safety Studies Using Electronic Healthcare Data. Silver Springs, MD: FDA; 2013.
2. Benchimol EI, Manuel DG, To T, Griffiths AM, Rabeneck L, Guttmann A. Development and use of reporting guidelines for assessing the quality of validation studies of health administrative data. J Clin Epidemiol. 2011;64(8):821–9.
3. Whiting PF, Rutjes AW, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med. 2011;155(8):529–36.
4. Cohen JF, Korevaar DA, Altman DG, et al. STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. BMJ Open. 2016;6(11): e012799.
5. Dreyer NA, Bryant A, Velentgas P. The GRACE checklist: a validated assessment tool for high quality observational studies of comparative effectiveness. J Manag Care Spec Pharm. 2016;22(10):1107–13.
6. Wang SV, Pinheiro S, Hua W, et al. STaRT-RWE: structured template for planning and reporting on the implementation of real-world evidence studies. BMJ. 2021;372: m4856.
7. U.S. Department of Health and Human Services, Food and Drug Administration Center for Drug Evaluation and Research (CDER) Center for Biologics Evaluation and Research (CBER) Oncology Center of Excellence (OCE). Real-World Data: Assessing Electronic Health Records and Medical Claims Data To Support Regulatory Decision-Making for Drug and Biological Products: Guidance for Industry. DHHS. https://www.fda.gov/media/152503/download. Published 2021. Updated September 2021. Accessed January 31st, 2022.
8. Lanes S, Brown JS, Haynes K, Pollack MF, Walker AM. Identifying health outcomes in healthcare databases. Pharmacoepidemiol Drug Saf. 2015;24(10):1009–16.
9. Cocoros NM, Arlett P, Dreyer NA, et al. The certainty framework for assessing real-world data in studies of medical product safety and effectiveness. Clin Pharmacol Ther. 2021;109(5):1189–96.
10. Beyrer J, Abedtash H, Hornbuckle K, Murray JF. A review of stakeholder recommendations for defining fit for purpose real-world evidence algorithms. J Comp Eff Res. 2022;2:2.
11. U.S. Department of Health and Human Services, Food and Drug Administration Center for Drug Evaluation and Research (CDER) Center for Biologics Evaluation and Research (CBER) Oncology Center of Excellence (OCE). Real-World Data: Assessing

Registries to Support Regulatory Decision-Making for Drug and Biological Products Guidance for Industry. https://www.fda.gov/media/154449/download. Published 2021. Accessed January 13, 2022.
12. European Medicines Agency. Head of Medicines Agencies. Guideline on good pharmacovigilance practices (GVP) Module VIII – Post-authorization safety studies (Rev 3). Updated October 13, 2017. Accessed April 23, 2018. http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2012/06/WC500129137.pdf
13. European Medicines Agency. Scientific guidance on post-authorisation efficacy studies. Updated November 6, 2015. Accessed: November 22, 2021. http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2015/11/WC500196379.pdf
14. European Network of Centres for Pharmacoepidemiology and Pharmacovigilance. The ENCePP guide on methodological standards in pharmacoepidemiology (Revision 5). Updated July 5, 2016. Accessed: October 22, 2021. http://www.encepp.eu/standards_and_guidances/documents/ENCePPGuideofMethStandardsinPE_Rev5.pdf
15. Velentgas P, Dreyer NA, Nourjah P, Smith SR, (Eds.). TM. Developing a protocol for observational comparative effectiveness research: a user's guide. Agency for Healthcare Research and Quality (AHRQ). Effective Health Care Program on developing a protocol for observational comparative effectiveness research: a user's guide. . In. Rockville (MD): (US)2013 Jan.: Accessed November 30, 2018. https://www.ncbi.nlm.nih.gov/books/NBK126190/pdf/Bookshelf_NBK126190.pdf.
16. Patient-Centered Outcomes Research Institute (PCORI). PCORI methodology standards. Updated February 26 AN, 2018. https://www.pcori.org/research-results/about-our-research/research-methodology/pcori-methodology-standards.
17. Diamond IR, Grant RC, Feldman BM, et al. Defining consensus: a systematic review recommends methodologic criteria for reporting of Delphi studies. J Clin Epidemiol. 2014;67(4):401–9.
18. Keeney S, Hasson F, McKenna H. Consulting the oracle: ten lessons from using the Delphi technique in nursing research. J Adv Nurs. 2006;53(2):205–12.
19. Pill J. The Delphi method: substance, context, a critique and an annotated bibliography. Socioecon Plann Sci. 1971;5(1):57–71.
20. Loughlin KG, Moore LF. Using Delphi to achieve congruent objectives and activities in a pediatrics department. J Med Educ. 1979;54(2):101–6.
21. Green B, Jones M, Hughes D, Williams A. Applying the Delphi technique in a study of GPs' information requirements. Health Soc Care Community. 1999;7(3):198–205.
22. Ammann EM, Leira EC, Winiecki SK, et al. Chart validation of inpatient ICD-9-CM administrative diagnosis codes for ischemic stroke among IGIV users in the Sentinel Distributed Database. Medicine. 2017;96(52): e9440.
23. Lakshminarayan K, Larson JC, Virnig B, et al. Comparison of medicare claims versus physician adjudication for identifying stroke outcomes in the women's health initiative. Stroke. 2014;45(3):815–21.
24. Berger JS, Laliberté F, Kharat A, et al. Real-world effectiveness and safety of rivaroxaban versus warfarin among non-valvular atrial fibrillation patients with obesity in a US population. Curr Med Res Opin. 2021;37(6):881–90.
25. Harris PA, Taylor R, Minor BL, et al. The REDCap consortium: building an international community of software platform partners. J Biomed Inform. 2019;95: 103208.
26. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)–a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform. 2009;42(2):377–81.
27. Kumamaru H, Judd SE, Curtis JR, et al. Validity of claims-based stroke algorithms in contemporary medicare data. Circulation. 2014;7(4):611–9.
28. Niesner K, Murff HJ, Griffin MR, et al. Validation of VA administrative data algorithms for identifying cardiovascular disease hospitalization. Epidemiology. 2013;24(2):334–5.
29. Roumie CL, Mitchel E, Gideon PS, Varas-Lorenzo C, Castellsague J, Griffin MR. Validation of ICD-9 codes with a high positive predictive value for incident strokes resulting in hospitalization using Medicaid health data. Pharmacoepidemiol Drug Saf. 2008;17(1):20–6.
30. Thigpen JL, Dillon C, Forster KB, et al. Validity of international classification of disease codes to identify ischemic stroke and intracranial hemorrhage among individuals with associated diagnosis of atrial fibrillation. Circ Cardiovasc Qual Outcomes. 2015;8(1):8–14.
31. Tirschwell DL, Longstreth WT Jr. Validating administrative data in stroke research. Stroke. 2002;33(10):2465–70.
32. Wahl PM, Rodgers K, Schneeweiss S, et al. Validation of claims-based diagnostic and procedure codes for cardiovascular and gastrointestinal serious adverse events in a commercially-insured population. Pharmacoepidemiol Drug Saf. 2010;19(6):596–603.
33. Leptak C, Menetski JP, Wagner JA, Aubrecht J, Brady L, Brumfield M, Chin WW, Hoffmann S, Kelloff G, Lavezzari G, Ranganathan R, Sauer JM, Sistare FD, Zabka T, Wholley D. What evidence do we need for biomarker qualification? Sci Transl Med. 2017;9(417):l4599. https://doi.org/10.1126/scitranslmed.aal4599.
34. Food and Drug Administration. 2013. "Best Practices for Conducting and Reporting Pharmacoepidemiologic Safety Studies Using Electronic Healthcare Data."

Authors and Affiliations

Sonal Singh[1,6] · Julie Beyrer[2] · Xiaofeng Zhou[3] · Joel Swerdel[4] · Raymond A. Harvey[4] · Kenneth Hornbuckle[2] · Leo Russo[3] · Kanwal Ghauri[5] · Ivan H. Abi-Elias[6] · John S. Cox[6] · Carla Rodriguez-Watson[5]

✉ Sonal Singh
sonal.singh@umassmemorial.org

1 Department of Family Medicine and Community Health, UMass Chan Medical School, Worcester, MA, USA

2 Eli Lilly and Company, Indianapolis, IN, USA

3 Pfizer, Global Medical Epidemiology, Paoli, PA, USA

4 Janssen R&D, LLC, New Brunswick, NJ, USA

5 Reagan-Udall Foundation for the FDA, Washington, DC, USA

6 Division of Health Systems Science, Department of Medicine, UMass Chan Medical School, Worcester, USA