METHODOLOGICAL ARTICLES

A Checklist for Ascertaining Study Cohorts in Oncology Health Services Research Using Secondary Data: Report of the ISPOR Oncology Good Outcomes Research Practices Working Group

Kathy L. Schulman, MA1,2,*, Karina Berenson, MPH2, Ya-Chen (Tina) Shih, PhD3, Kathleen A. Foley, PhD4, Arijit Ganguli, MBA, PhD5, Jonas de Souza, MD5, Nicholas A. Yaghmour, MPP5, Alex Shteynshlyuger, MD6

1Outcomes Research Solutions, Inc., Bolton, MA, USA; 2Covance, Gaithersburg, MD, USA; 3University of Chicago, Chicago, IL, USA; 4Truven Health Analytics, Philadelphia, PA, USA; 5Abbott Laboratories, Abbott Park, IL, USA; 6Phoenixville Urologic Associates, Phoenixville, PA, USA

ABSTRACT

Objectives: The ISPOR Oncology Special Interest Group formed a working group at the end of 2010 to develop standards for conducting oncology health services research using secondary data. The first mission of the group was to develop a checklist focused on issues specific to selection of a sample of oncology patients using a secondary data source. Methods: A systematic review of the published literature from 2006 to 2010 was conducted to characterize the use of secondary data sources in oncology and inform the leadership of the working group prior to the construction of the checklist. A draft checklist was subsequently presented to the ISPOR membership in 2011 with subsequent feedback from the larger Oncology Special Interest Group also incorporated into the final checklist. Results: The checklist includes six elements: identification of the cancer to be studied, selection of an appropriate data source, evaluation of the applicability of published algorithms, development of custom algorithms (if needed), validation of the custom algorithm, and reporting and discussions of the ascertainment criteria. The checklist was intended to be applicable to various types of secondary data sources, including cancer registries, claims databases, electronic medical records, and others. Conclusions: This checklist makes two important contributions to oncology health services research. First, it can assist decision makers and reviewers in evaluating the quality of studies using secondary data. Second, it highlights methodological issues to be considered when researchers are constructing a study cohort from a secondary data source.

Keywords: cohort, ISPOR checklist, oncology, sample selection, secondary data.

Published by Elsevier Inc.

Purpose

The use of secondary data sources is an increasingly accepted approach in oncology health services research, as exemplified by studies using such data to profile care patterns, measure patient outcomes, and estimate cancer-related costs [1–6]. Ascertainment of study cohorts from these data sources, however, can be difficult. Selection of a sample of cancer patients from secondary data sources is challenging when coding systems (e.g., International Classification of Diseases, Ninth Revision, Clinical Modification) used to identify these patients lack adequate clinical precision (e.g., cancer stage) or when clinical interventions are masked by payment system protocols. In these circumstances, the cohort selected may not be reflective of the larger population with the disease (poor sensitivity) or may contain large numbers of patients who do not actually have the disease at all (poor specificity). In either case, this may lead to inconsistent, spurious, or erroneous results and compromise the clinical relevancy of study findings. While this is a common problem in observational studies, the complexity of oncologic disease and its inexact representation in clinical data systems may exacerbate both selection and misclassification bias.

The purpose of this article is to both provide researchers who conduct oncology health services research using secondary data a list of methodological issues to consider when selecting study cohorts and assist researchers as well as journal reviewers/readers, payers, and policymakers in evaluating the quality of published studies involving secondary data analyses.

A working group to develop standards for oncology health services research using secondary data was established through the ISPOR Oncology Special Interest Group (SIG) to review and address issues related to this line of research. The first task of our working group was to develop a checklist for use in selecting samples of patients with a specific type of cancer from secondary data sets. We completed this task in two phases: a systematic literature review followed by development of the draft and final checklist. The formal review of the literature was conducted to
inform SIG leadership as to the current state of science in the literature prior to the development of the checklist. It was intended to complement the knowledge base and experience of the larger SIG as well as the existing validation literature. It was not conducted, as are many systematic reviews, to support meta-analysis.

SIG consensus was shaped first by discussions regarding the extent to which the literature review should be structured and subsequently the development of the elements of the structured review. SIG members were involved in the construction of search parameters, in the development of the abstraction form, in the actual abstraction of selected articles, and of course in the summary of findings. This process, which took place over the course of a year, included multiple meetings in which the literature search, its findings, and the potential importance to the checklist were discussed. It should be noted that disagreement between authors was resolved by third-party review. Consensus was also formed by discussions with the larger ISPOR membership during the dissemination of these results at the annual meeting and also during presentation of the draft deliverable to the larger SIG membership.

The literature review resulting from the above efforts reflects a concerted decision on the part of the SIG leadership to either confirm or refute a priori assumptions about the current state of oncology research using secondary data prior to drafting a checklist. For example, SIG members underestimated the frequency with which claims data were being used in isolation from other data sources and overestimated researcher reliance on previously published algorithms. Moreover, the checklist itself was highly influenced by the degree of difficulty in abstracting the necessary detail about sample selection in general because key elements either were missing altogether or were vague or nonspecific in their presentation. Finally, consensus regarding key elements of the checklist was inferred on the basis of results of the review in conjunction with the experience of SIG members, most notably the importance of identifying the key clinical elements of the cancer of interest.

The literature search was conducted in PubMed on English language articles published between January 1, 2006, and December 31, 2010. We chose the most recent 5-year time period at the time the literature review was performed because it captures the majority of work done with secondary data sources while also reflecting the most recent methodological thinking in this topic area. Search terms included “claims,” “hospital discharge,” “electronic medical records (EMR),” “registry,” and “administrative” in conjunction with the Mesh term “Neoplasms.” No other restrictions were placed on the query. Two reviewers (K.S. and K.B.) assessed abstracts for potential eligibility with subsequent full-text article review by the larger working group leadership team to determine appropriateness of inclusion. Findings from the literature review were used to characterize the use of secondary data sources in oncology, to identify potential methodological and reporting issues for oncology health services researchers, and to inform the second phase, in which a draft checklist was developed by consensus among the leadership working group. The draft checklist, presented at the ISPOR international meeting in 2011, and subsequently reviewed by the larger Oncology SIG, was then modified by incorporating comments and feedback to produce the final checklist.

**Review of Current Literature**

The literature search yielded 863 abstracts, of which 321 were eligible for full article review and 294 [6–298] were eligible for abstraction (Fig. 1).

The literature search demonstrated that secondary data sources are widely used for oncology research and that the ascertainment of study cohorts largely relies on registry or claims-based data. As evidenced in Table 1, 72.8% of the studies identified in our review relied to some degree on claims data. Of the 294 articles selected for review, 21.8% (n = 64) were claims-only studies [7–70], 5.1% (n = 15) were registry-only [71–85], 51% (n = 150) used claims and registry data together without any other supplemental data set [6,24,151–298], and the remaining 22.1% (n = 65) used other data sources, primarily chart-based systems or hospital discharge data sets, either alone or in combination with claims and/or registry data [86–150]. Chart-based systems included paper or electronic health records (EHRs) as well as electronic medical record (EMR) databases, programatically generated subsets of EHRs. The Surveillance Epidemiology and End Results (SEER)/Medicare database was the single most commonly used (N = 93) data set, representing 62% of combination claims and registry studies and 31.6% of all eligible studies.
Table 1 – Secondary data sources used in oncology outcomes research.*

<table>
<thead>
<tr>
<th>Cancer, unspecified, n (%)</th>
<th>Registry only (N = 15)</th>
<th>Claims only (N = 64)</th>
<th>Registry + claims (N = 150)</th>
<th>Other data sources (N = 65)</th>
<th>All data sources (N = 294)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cancer, unspecified, n (%)</td>
<td>1 (6.7)</td>
<td>8 (12.5)</td>
<td>5 (3.3)</td>
<td>13 (20.0)</td>
<td>27 (9.2)</td>
</tr>
<tr>
<td>Breast, n (%)</td>
<td>2 (13.3)</td>
<td>34 (53.1)</td>
<td>46 (30.7)</td>
<td>12 (18.5)</td>
<td>94 (32.0)</td>
</tr>
<tr>
<td>Colorectal, n (%)</td>
<td>7 (46.7)</td>
<td>13 (20.3)</td>
<td>45 (30.0)</td>
<td>10 (15.4)</td>
<td>75 (25.5)</td>
</tr>
<tr>
<td>Leukemia, n (%)</td>
<td>1 (6.7)</td>
<td>7 (10.9)</td>
<td>7 (4.7)</td>
<td>2 (3.1)</td>
<td>17 (5.8)</td>
</tr>
<tr>
<td>Lymphoma, n (%)</td>
<td>1 (6.7)</td>
<td>12 (18.8)</td>
<td>14 (9.3)</td>
<td>1 (1.5)</td>
<td>28 (9.5)</td>
</tr>
<tr>
<td>Lung, n (%)</td>
<td>2 (13.3)</td>
<td>19 (29.7)</td>
<td>27 (18.0)</td>
<td>14 (21.5)</td>
<td>62 (21.1)</td>
</tr>
<tr>
<td>Multiple myeloma, n (%)</td>
<td>1 (6.7)</td>
<td>5 (7.8)</td>
<td>5 (3.3)</td>
<td>1 (1.5)</td>
<td>12 (4.1)</td>
</tr>
<tr>
<td>Prostate, n (%)</td>
<td>2 (13.3)</td>
<td>10 (15.6)</td>
<td>40 (26.7)</td>
<td>5 (7.7)</td>
<td>57 (19.4)</td>
</tr>
<tr>
<td>Other, n (%)</td>
<td>7 (46.7)</td>
<td>20 (31.3)</td>
<td>45 (30.0)</td>
<td>26 (40.0)</td>
<td>98 (33.3)</td>
</tr>
</tbody>
</table>

* Column percentages may exceed 100% because an individual study may reference multiple cancer types.

Studies included in the review were primarily conducted in the United States or Canada but also included studies using secondary data from Taiwan, Korea, France, Denmark, England, Italy, Australia, Sweden, The Netherlands, Finland, Belgium, and Japan. Among the claims-only studies, there was very little reported use of either a previously published algorithm (n = 8, 12.9%) [20,27,37,48,49,58,60,61] or a validated algorithm (n = 4, 6.5%) [48,58,60,61]. Almost half (n = 31, 48.4%) of claims-only studies relied solely on diagnosis codes to select their study sample [7,9,12–15,19,20,25,27–29,31,32,34,35,36–38,39–41,46,47,49,50,56,59,60,63,69,70], with 35.5% (n = 11) of these studies using a single claim with a cancer diagnosis as a prerequisite for study entry [15,29,32,34,36,37,46,50,63,69,70]. Only 4.8% of the articles discussed the implications of their selection criteria on the study findings [27,48,49].

Framework of Checklist

A six-step checklist was developed to assist reviewers and researchers (Table 2). Elements of the checklist are presented in the order that researchers are anticipated to consider them as they design and implement a study. Items in the first category are to identify cancer-specific clinical characteristics that may influence the data source selection and the inclusion/exclusion criteria for the study sample. The second category, selection of an appropriate data source, is then followed by a review of the literature to identify potential selection criteria or algorithms that may be used or adapted for the study. Items in the fourth category summarize factors to consider in developing a custom algorithm, while those in the fifth category detail considerations in the validation of such algorithm. Finally, the checklist includes recommendations for reporting the selection criteria or algorithm. The entire checklist is presented in Table 2. Additional oncology resources can be found at http://www.ispor.org/OncologyORResources/SearchOcologyResources.aspx.

How Should the Checklist Be Used?

The checklist, focused on the four general types of secondary data identified from our literature review, is intended to assist decision makers in evaluating the quality of published studies and to provide researchers a stepwise list of methodological issues for consideration when developing case ascertainment criteria and reporting study findings in oncology research. Because the purpose of the checklist is to guide the creation of appropriate inclusion and exclusion criteria to properly identify patients with specific tumor characteristics, it should be used to complement other checklists and guidelines related to secondary data research, such as the ISPOR checklist on retrospective database studies [299] and the ISPOR Good Research Practices for Comparative Effectiveness Research [300–302].

Description of Checklist Elements

Step 1: Identify Clinical Aspects of Cancer

Step 1 of the checklist emphasizes the importance of understanding the biology, natural history, and etiology of the specific cancer of interest prior to establishing the inclusion/exclusion criteria of the study protocol. Knowledge of the natural history of the disease will facilitate an understanding of the patient experience, specifically key clinical characteristics and the expected duration of the window of observation at each stage of the disease. Knowledge of biology and/or etiology will identify clinical subtypes, in addition to cancer stage, in which patient prognosis may differ.

Biomarkers are anatomic, physiologic, biochemical, or molecular parameters associated with the presence and severity of specific disease states. They may be diagnostic, perhaps before the cancer is detectable by conventional methods, prognostic, forecasting how aggressive the disease process will be, or predictive, identifying which patient will respond to which drug [303–306]. Failure to differentiate these subtypes can be problematic when evaluating clinical outcomes (e.g., mortality, progression, and symptom sets) or when comparing the toxicity profiles of therapy. For example, a retrospective study of outcomes among patients with metastatic colorectal cancer will likely need to address the issue of how to differentiate patients with colorectal cancer with KRAS wild-type gene (a gene that encodes the protein GTPase KRAS) from patients with the KRAS mutated gene, as the response to therapeutic agents may differ between these two groups. Similarly, studies evaluating outcomes or treatment burden in women with breast cancer may need to be able to differentiate women who are human epidermal growth factor receptor 2 positive [307] from those who are not.

It is also recommended that treatment guidelines in use during the study period are reviewed. The National Comprehensive Cancer Network, for example, publishes guidelines for treatment of cancer by site. It is also important to understand past treatment practices to the extent they impact the study window as well as the extent to which real-world practices deviate from guidelines.

The researcher should also have a good understanding of how the specific cancer type is coded in oncology practice and whether those codes are clinically specific enough to identify the pertinent population. For example, International Classification of Diseases, Ninth Revision, Clinical Modification diagnosis codes, which are used in the United States, lack the specificity of International Classification of Diseases, Tenth Revision codes. They cannot be used alone, for example, to identify patients with non–small cell lung cancer because the code does not
differentiate types of lung cancer. While there is far greater specificity to body site for solid tumors in International Classification of Diseases, Tenth Revision, Clinical Modification and greater detail as to clinical type for neoplasms of lymphatic and hematopoietic tissue, specificity issues remain in International Classification of Diseases, Tenth Revision, Clinical Modification. The researcher should consult a nosologist familiar with both coding practice rules and real-world conventions because a descriptive code listing alone may not suffice. The researcher should also determine whether in situ codes are appropriate for inclusion because these are not clinically relevant for all cancers.

Step 1 is an iterative process. Study objectives and/or the target population will determine the initial inclusion/exclusion criteria, but inherent limitations of the data chosen for the study often require revision of either study objectives and/or the target population to ensure the validity of the study.
# Table 3 – Data type strengths and limitations, by data type.

<table>
<thead>
<tr>
<th>Description, and example</th>
<th>Strength</th>
<th>Limitations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chart-based data (EMR or EHR)</td>
<td>Most accurate way to identify a cancer-specific sample as data are directly compiled from patients' medical records</td>
<td>Data are often difficult to access and frequently specific to a given hospital or physician system.</td>
</tr>
<tr>
<td>Examples:</td>
<td>Data may contain better clinical detail, with information on not only the frequency of essential diagnostic testing (e.g., laboratory, histology, including biomarkers) but also the results of such testing, both at the time of diagnosis and during follow-up</td>
<td>Data extractions, especially those from paper charts, are resource intensive. May require a trained nurse abstractor.</td>
</tr>
<tr>
<td>- US Oncology iKnowMed</td>
<td>Allow text-based searching for keywords</td>
<td>Requires data to be parsed and extracted into database format.</td>
</tr>
<tr>
<td>- ACORN Data Warehouse</td>
<td>EMR and EHR data contain relatively complete demographic information</td>
<td>Data may not adequately track patients across sites of service. Community practice data are frequently missing inpatient and nononcology outpatient services. Hospital-based data are frequently missing care delivered in other institutions, both inpatient and outpatient. Pharmacy-dispensed medications may be unavailable altogether.</td>
</tr>
<tr>
<td></td>
<td>May be linked to billing systems, thereby providing access to charges</td>
<td>Data that are missing will likely not be missing at random, introducing biases to the analyses.</td>
</tr>
<tr>
<td></td>
<td>May be linked to patient-reported outcomes</td>
<td>Limited generalizability of cost data (inferred from charges) as they are institution-specific.</td>
</tr>
</tbody>
</table>

## Claims data

- Claims databases are by-products of patient enrollment and billing records submitted to insurance companies.
- The data typically contain inpatient admissions and outpatient visits billed over the course of an enrollment period. Some also include data of outpatient prescription drugs.
- Examples:
  - Medicare or Medicaid claims data
  - MarketScan
  - All Payer Claims Database (APCD)*
  - Clinical Practice Research Datalink (CPRD)
- Data access is affordable and resource sparing
- Large sample size provides ample statistical power; supports evaluation of rare events
- Data capture in national data sets may reflect the diversity of practice patterns
- Information on payment and resource utilization can be used to infer costs
- Test and procedure frequency, but not results, may be easily measured
- Billing codes (e.g., ICD-9, ICD-10, and CPT) can be used to identify specific treatment patterns and pathways
- When data are compiled from a single payer in countries with national health insurance or a financially integrated health care system (e.g., Kaiser), information from claims is in general comprehensive and has good longitudinal follow-up
- Often do not have sufficient information to identify the cancer of interest, particularly for cancers that cannot be captured by ICD-9 codes alone or those associated with predictive/prognostic biomarkers
- Information on race/ethnicity, mortality, and more detailed geographic locations may not be available
- Supplemental laboratory/histology data may be incomplete or unavailable. No biomarker data (e.g., genetic subtypes such as KRAS, BRAF, and HER2)
- Continuous enrollment requirements in commercial claims data may result in a large reduction in sample size, especially in data sets from a single insurer because changes in insurance will mean discontinuation from the data set. If data source is employment-based insurance, any interruption in employment (e.g., job change, short- or long-term disability) may affect longitudinal data availability
- Completeness and specificity of data are subject to reimbursement policies
- Carve outs of mental health data are common
- Data from Medicare HMO plans may be incomplete

## Registry data

- Cancer registry collects patients' time of diagnoses, tumor characteristics, summary of treatments within a fixed period after diagnosis, and vital status
- Use of registry data requires in-depth understanding of the data reporting standard, and the coding and staging
- Gathered detailed information on tumor characteristics at diagnosis, e.g., staging, nodal status, and tumor size
- Decent data on patients' demographics
- Reliable source to identify incident cohort of cancer patients
- Most registries still do not require collecting biomarker data, specifically genetic subtypes (e.g., KRAS and HER2)
- Information on treatment is often limited to the first course of treatment or treatments within the first 4-6 mo of diagnosis
- Cannot be used to construct samples of treatment-resistant
### Systems for the Tumor of Interest

- Surveillance, Epidemiology and End Results (SEER)
- National Cancer Database (NCDB)
- CDC National Program of Cancer Registries (NPCR)
- Swedish Cancer Registry
- National Cancer Data Repository (NCADR) (UK)
- Saskatchewan Services Databases

### Claims Data Linked to Registry or EMR

#### Claims Data Linked to Cancer Registry or Medical Records
- SEER-Medicare database
- Clinical Practice Research Datalink (CPRD)
- Saskatchewan Cancer Registry and Services Database
- HMO Cancer Research Network (CRN) *

#### Contains Detailed Information on Tumor Characteristics, Demographics, Vital Status, and Also Payment Data
- Reliable source to identify incident cohort of cancer patients

#### Information on Biomarkers/Genetic Subtypes (e.g., KRAS, CAF, and HER2), and Other Laboratory and/or Histology Data May Still Be Incomplete or Unavailable
- Issues related to switching of insurance plans and reimbursement policies remain

---


* APCD composed of medical, pharmacy, and dental claims, as well as information about member eligibility, benefit design, and providers for all payers covering Massachusetts residents.

* The HMO CRN, funded by the NCI, consists of data from a consortium of 14 HMOs to conduct research on cancer prevention, treatment, and surveillance.
**Step 2: Select an Appropriate Data Source**

The second step is to choose an appropriate data source given the cancer of interest, the study objectives, the potential methodological constraints, and the characteristics associated with candidate data sources. Specifically, the researcher must determine whether a given data source has enough clinical detail to accurately identify the cancer population of interest, is sufficiently comprehensive, and has adequate duration of follow-up to address the research question. Table 3 provides an overview of different types of secondary data and summarizes the strengths and limitations of each data type.

While medical records (paper, EHRs) are often considered the gold standard for accurately identifying a cancer-specific sample, they are used infrequently because of resource and budget constraints as well as data accessibility. As a compromise, registries are frequently used to establish the gold standard. Information captured in a registry, however, is usually limited to the time at or immediately following entry in the registry. For example, the SEER registry currently reports only the first course of treatment. Hence, this data characteristic limits its ability to construct a sample of treatment-resistant patients (e.g., men with hormone-refractory prostate cancer) or to capture all patients with advanced disease, especially those who initially presented with early stage disease. Unless linked to another source, registries alone may also not capture detailed information on long-term follow-up, compromising the ability to measure longer term outcomes other than mortality. In addition, while registries collect information on disease staging, they may be missing the data elements needed to identify some cancer subtypes; of special concern are subtypes determined by genetic dispositions. Researchers should have a strong understanding of how the registry data are captured, collected, and coded, including a history of changes in registry methodology over the time period of the study. These should include familiarity with the coding and staging system used in the registry (e.g., International Classification of Diseases for Oncology, 3rd Edition), the population source and geographic coverage for the registry, and the validity of linkage algorithms if the registry was linked to another data source, such as the SEER-Medicare data.

Although the checklist developed by our working group is focused on the ascertainment of study cohorts, the nature of cancer requires that this be done in conjunction with assessing the researcher’s ability to fully observe the patient both before and after the pertinent index event. Cancer may be chronic, recurrent, and/or associated with high mortality. As such, it may be important to use a data source with both sufficient long-term follow-up and accurate mortality information. Because cancer patients receive care in inpatient, outpatient, and multiple office settings, a data source that does not uniformly capture information from each of these settings may be insufficient. This is especially problematic if the outcomes of interests are costs or resource utilization. Cancer patients may also experience job loss or transition to short- or long-term disability, either of which can result in their periodic termination from select databases. Researchers may therefore not be able to fully observe previous episodes of care (left censoring). This can result in misassignment of the primary tumor type, failure to distinguish between incident cases and prevalent or recurrent cases, and/or their ability to profile care or measure study outcomes during follow-up (right censoring).

For researchers interested in claims data, one of the primary concerns is whether the data capture sufficient clinical detail to identify the cancer of interest, especially if this includes cancer types with poor code specificity, or those strongly associated with predictive/prognostic biomarkers. Supplemental laboratory or histology data in claims data sets may be incomplete, especially if the study requires repeated measures such as prostate specific antigen measurements in prostate cancer or complete blood cell counts and polymerase chain reaction results in leukemia. For claims data sources tied to employment-based insurance, data for patients who have a change in employment status or go on long-term disability may not be fully captured. Moreover, mortality data are usually limited in commercial claims data sets for patients dying at home. Researchers should specifically inquire about the availability of date of death before purchasing a claims data set, as providers of commercial data are actively working to improve their capture of date of death. While death proxy algorithms exist for claims data, they have not been validated and have only been used in studies of patients with end-stage disease [308,309]. Information on race and ethnicity is not always available in claims data; these should also be verified prior to choosing a data set. Finally, reimbursement factors may impact either data completeness or how specific data are manifest in claims data. For example, researchers using SEER-Medicare data are recommended to exclude patients who enrolled in a health maintenance organization because claims data from these plans tend to be incomplete.

EMR data offer good clinical detail, including laboratory and histology, both at the time of diagnosis and during follow-up. These data, however, must be programmatically parsed and extracted from EHRs into database format. Moreover, there are frequently multiple sources of EMR data, each unique to the site of care (hospital, oncology clinic, primary care), even in countries with national health insurance. Accordingly, these databases may suffer from significant amounts of missing data or coding inconsistencies, although it should be noted that, in our experience, having a research nurse review the original EHR is a viable strategy to ameliorate this problem. Hospital discharge data, especially those culled from information technology systems designed to document clinician care orders, may also be restricted in their ability to track patients across sites of service. These data generally are able to track a patient’s care within the reporting hospital and can capture outpatient clinic data only in those financially integrated health care networks that include data linkages across sites of service.

**Step 3: Review Published Criteria or Algorithms of Case Ascertainment**

This step recommends a review of the literature to identify previously published, cancer-specific, selection criteria. Even if it is determined that the published algorithms are not appropriate to identify the cancer of interest, elements in these criteria can still highlight the full complement of pertinent diagnostic or procedure codes, the impact of varying clean periods, and/or the necessity of considering other variables in the inclusion/exclusion logic.

Researchers considering the use of an existing algorithm should examine the comparability of the data source. Algorithms may not be generalizable across data sources if coding systems differ across populations or if the sample used for validation is not a representative cross-section. This concern is especially pertinent for studies using algorithms published in one country that are then adapted for use in another country. For example, while the Current Procedural Terminology codes are the standard source to identify specific procedures or services in the United States, they are not a universally accepted coding system worldwide. Therefore, studies using a US-based algorithm to identify an incident cohort of cancer patients from the National Health Insurance claims in Taiwan will need to find the claims code in Taiwan that correspond to the Current Procedural Terminology codes used in the published algorithm. Researchers should also be mindful about the differential time periods between the time
the algorithm was developed and that of their own studies. Changes in practice over time or newly added codes for emerging technologies may limit the validity of an algorithm used in a study period that differs from the period of the original algorithm.

**Step 4: Develop the Custom Criteria/Algorithm**

If researchers conclude from step 3 that none of the existing algorithms suits the purpose of their study, a decision needs to be made whether to develop a new algorithm or to modify the study objective and target population to better fit the currently available algorithm (i.e., return to step 1). An important consideration in developing custom algorithms is whether the data source contains a component that distinguishes between physicians’ definitive versus differential diagnosis. Differential diagnoses, for example, may be recorded in claims data as “working” or “rule-out” diagnosis codes. Although the Center for Medicare & Medicaid Services policies prohibit such coding schemes in outpatient settings [310], they remain common in practice. Hence, if only one instance of the diagnosis code is used to determine a patient’s cancer status, the researcher risks selecting patients who do not have the cancer of interest either because they were ultimately determined to not have cancer at all or were subsequently determined to have a cancer other than the one under study.

Setoguchi et al. [311] tested four different algorithms for identifying cancer patients in claims data linked to a state cancer registry. When using just one diagnosis on a medical claim to identify patients with one of six cancers, they found that only 18.8% to 50.2% of such patients actually had cancer. When they also required at least one cancer-related procedure to occur with the cancer diagnosis, however, the percentage accurately identified increased from 41.1% to 81.7%. Accuracy of claims-based algorithms was greater for solid tumors than it was for hematologic malignancies, such as lymphoma or leukemia [311]. Similarly, Friedlin et al. [312] concluded that in the case of pancreatic cancer, using a single International Classification of Diseases, Ninth Revision diagnosis code was associated with high sensitivity, but poor specificity. Hence, while the criteria identified most of the patients with pancreatic cancer in the database, 62% of the patients in the final sample did not actually have the malignancy.

Information documented in the above studies, combined with working experiences of the workgroup members, suggests that it is essential to require more than one diagnosis code to select the tumor of interest and/or exclude claims for diagnostic procedures (e.g., laboratory, pathology, venipuncture, and noninterventional radiology). Including codes for cancer-specific procedures or therapies may further improve the accuracy. Study criteria that require a breast cancer diagnosis code and a mastectomy or lumpectomy procedure code is an example of a strategy that pairs a specific diagnosis code with a second component, cancer-specific treatment. This approach is likely to maximize specificity in selecting breast cancer patients with modest impact on sensitivity. A review of the approaches used by Nattinger et al. [313] provides further guidance.

Using a second component in addition to the diagnosis code, however, is not appropriate for all studies. If the study objective is to characterize treatment patterns, document disparities in treatment, or evaluate economic burden, the use of a treatment procedure as one of the components to identify the sample thus compromises the study objective. The researcher should also be aware that some therapies apply to multiple cancers, all of which may be on the physicians’ differential diagnosis list, thus impacting comparative analyses. Hence, requiring treatment of disease as part of the selection criteria may not be sufficient. For example, patients presenting with a symptom set suggestive of leukemia may have multiple working diagnoses suggestive of acute or chronic lymphocytic or myelogenous leukemia, some of which are treated with a common agent. The researcher may thus be faced with conflicting diagnosis codes and a nonspecific therapy.

Given the challenges associated with adding a procedure code as part of the ascertainment criteria in addressing certain study objectives, our recommendation is that researchers should consider requiring, for example, two or more diagnoses, on non-diagnostic claims, and on different dates between 30 and 365 days apart. This algorithm may require modification depending on the cancer type and study objective, especially when selecting cohorts of patients with terminal cancer.

Requiring a period free of cancer-specific diagnoses codes (known as the “clean” or “wash-out” period) also helps differentiate incident (new) cases from prevalent or recurrent cases. Distinguishing between incident and recurrent cancer however may be challenging in the presence of left censoring because previous cancer episodes may not be observed. Left censoring occurs if baseline windows of observation are inadequate, or if data sets do not comprehensively capture cancer care for any of the reasons detailed in step 2. Furthermore, it may be difficult to verify whether the selected cancer is the primary tumor and not a secondary metastatic tumor site. Linkage to a cancer registry may mitigate this problem, although not completely. In claims data not linked to registries, breast cancer that spreads to the lung, for example, may be incorrectly coded as an incident primary lung cancer rather than metastatic breast cancer. While identifying the primary tumor is not always possible in some data sets, researchers should understand the extent of misclassification in developing custom algorithms or applying published algorithms to make informed decisions that prevent the introduction of bias and minimize measurement error.

**Step 5: Validation and Sensitivity Analyses**

When developing custom algorithms, if the researcher has access to a data source that can serve as a “gold standard” for determining the cancer of interest, then the algorithm should be validated. Ideally, the new algorithm should be validated in a representative, cancer-specific, cross-section of patients. If that is not feasible, it is still valuable to validate the algorithm in a subsample. If validation is attempted, researchers should subsequently report their findings, clearly stating the population used for validation as well as the performance (sensitivity, specificity, positive predictive value [PPV], and negative predictive value) of the custom algorithm [314]. If validation is not feasible, sensitivity analyses should be conducted to assess the impact of alternate selection algorithms on study findings. For example, when selecting an incident cohort, the researcher may wish to test “clean” or “wash-out” periods of varying lengths or explore the impact of requiring three diagnoses instead of two to identify the study cohort. To benchmark the performance of custom algorithms, researchers may wish to include, as part of the sensitivity analysis, the published algorithm that is closest to the custom algorithm in development.

In general, there are few published algorithms for a given type of cancer. Breast cancer is the lone exception. Most of the algorithms used to identify breast cancer rely on matching to gold standard. The reported sensitivity of these algorithms varies from 46% to 99.5%, with the accompanying PPV suggesting that between 94% and 93% of the population the algorithm identified as having breast cancer truly had the disease [315–317]. Regression-based algorithms were somewhat less variable, with sensitivity ranging from 76.2% to 90%, specificity from 90% to 99.3%, and the PPV from 36.3% to 93% [318–319]. It should be
noted that sensitivity and specificity are characteristics of the “test,” that is, the algorithm being validated. In contrast, the PPV and the negative predictive value are influenced by the prevalence of disease in the population that is being tested. Hence, researchers implementing algorithms in different populations may not achieve similar predictive values.

**Step 6: Report the Selection Criteria/Algorithm**

Accurate and detailed reporting of the sample selection criteria is essential. A schematic detailing each step of inclusion and exclusion criteria and the change in sample size following each data step is highly recommended. It provides the opportunity for a critical evaluation of methods for case ascertainment, ensures that study findings can be replicated, and enhances comprehensibility of study findings. Data sources should be described fully, including the number of databases and data files used, number of lives covered, geographic regions included, types of insurance captured (e.g., fee for service, capitated), and source of information (e.g., health plan, payer, employer, hospital, community oncology clinics, and registry or other consortium). General descriptions referring to the data as “administrative” or “proprietary database” should be avoided. Without providing specific details on the database, readers may question the quality of the data and validity of the study findings.

Researchers should also provide detailed descriptions of the specific billing codes used, the number and types (i.e., diagnosis vs. procedure) of claims that were required, and any relevant time periods between claims or between diagnosis and treatment that were implemented as part of inclusion/exclusion criteria. Any change in coding over time should also be clearly documented. Finally, researchers should present findings from the validation and/or sensitivity analysis including an assessment of the algorithm’s capacity to accurately capture the target population and the biases possibly associated with certain inclusion/exclusion criteria. This will ensure that readers understand the context in which the study has been defined and the implication and/or generalizability of study findings.

**Conclusions**

Using secondary data in the analysis of cancer treatment and outcomes is of increasing importance. There is a growing need for clinical outcomes, health economic data, and patient-reported outcomes data from nonresearch settings that can be used to complement the data from randomized controlled trials. In addition, these data sources can be used to explore the off-label use of oncology products to both facilitate future development and enhance comprehension of the value equation for patients. As such, it is essential that the research conducted with secondary data be of the highest quality.

This article addresses the first issue—how to ascertain a specific cancer patient population—in a series of technical issues that researchers must address when using secondary data for oncology health services research. In addition to identifying an appropriate patient population, algorithms are needed to identify disease stage, line of therapy, and other clinical status indicators such as performance status. While some research has already been conducted in these areas [320,321], checklists such as the one presented here are needed for these types of algorithms as well. Beyond algorithm development, areas for future work in enhancing oncology data assets include the incorporation of more mortality data, biomarker data, and patient-reported outcomes. All these components are needed to fully utilize the vast resource that secondary data provide to oncology researchers.

This article provides a step-by-step checklist of issues to consider, along with recommendations, for drawing a cancer-specific sample in a variety of secondary data sources. It is important to note that we are not asking researchers to include in the article every element on the checklist. We do, however, believe that all articles should contain a detailed description of the sample selection criteria, the rationale for the criteria, the validity of the algorithm used to implement the criteria, and the implications of such algorithm on study results.

We hope that this checklist will facilitate the standardization of reporting and reviewing of future publications in oncology health services research using secondary data.

**Source of financial support:** No funding was received in support of this study.

**REFERENCES**


References


under the Medicaid expansion for cancer care. Womens Health Issues 2010;20:400–5.


