









ORIGINAL RESEARCH



Identification of cancer chemotherapy regimens and patient cohorts in administrative claims: challenges, opportunities, and a proposed algorithm

Catherine M. Lockhart^a , Cara L. McDermott^a , Aaron B. Mendelsohn^b , James Marshall^b, Ali McBride^c , Gary Yee^d, Minghui Sam Li^e , Aziza Jamal-Allial^f , Djeneba Audrey Djibo^g , Gabriela Vazquez Benitez^h , Terese A. DeFor^h and Pamala A. Pawloski^h 

^aBiologics and Biosimilars Collective Intelligence Consortium, Alexandria, VA, USA; ^bHarvard Pilgrim Health Care Institute and Harvard Medical School, Boston, MA, USA; ^cUniversity of Arizona College of Pharmacy, Tucson, AZ, USA; ^dUniversity of Nebraska Medical Center, Omaha, NE, USA; ^eUniversity of Tennessee Health Science Center, Memphis, TN, USA; ^fHealthCore Inc, Wilmington, DE, USA; ^gCVS Health Clinical Trial Services, Blue Bell, PA, US; ^hHealthPartners Institute, Bloomington, MN, USA

ABSTRACT

Background: Real-world evidence is a valuable source of information in healthcare. This study describes the challenges and successes during algorithm development to identify cancer cohorts and multi-agent chemotherapy regimens from claims data to perform a comparative effectiveness analysis of granulocyte colony stimulating factor (G-CSF) use.

Methods: Using the Biologics and Biosimilars Collective Intelligence Consortium's Distributed Research Network, we iteratively developed and tested a de novo algorithm to accurately identify patients by cancer diagnosis, then extract chemotherapy and G-CSF administrations for a retrospective study of prophylactic G-CSF.

Results: After identifying patients with cancer and subsequent chemotherapy exposures, we observed only 12% of patients with cancer received chemotherapy, which is fewer than expected based on prior analyses. Therefore, we reversed the initial inclusion criteria to identify chemotherapy receipt, then prior cancer diagnosis, which increased the number of patients from 2,814 to 3,645, or 68% of patients receiving chemotherapy had diagnoses of interest. Additionally, we excluded patients with cancer diagnoses that differed from those of interest in the 183 days before the index date of G-CSF receipt, including early-stage cancers without G-CSF or chemotherapy exposure. By removing this criterion, we retained 77 patients who were previously excluded. Finally, we incorporated a 5-day window to identify all chemotherapy drugs administered (excluding oral prednisone and methotrexate, as these medications may be used for other non-malignant conditions) as patients may fill oral prescriptions days to weeks prior to infusion. This increased the number of patients with chemotherapy exposures of interest to 6,010. The final cohort of included patients, based on G-CSF exposure, increased from 420 from the initial algorithm to 886 using the final algorithm.

Conclusions: Medications used for multiple indications, sensitivity and specificity of administrative codes, and relative timing of medication exposure must all be evaluated to identify patient cohorts receiving chemotherapy from claims data.

ARTICLE HISTORY

Received 18 October 2022
Revised 28 February 2023
Accepted 1 March 2023

KEYWORDS

Claims data; algorithm; oncology; chemotherapy; real-world data

JEL CLASSIFICATION CODES



C80; C8; C; C90; C9; C


Introduction

Observational research and real-world evidence (RWE) are valuable sources of information for stakeholders and decision-makers in the healthcare system, including clinicians, patients, payers, and regulatory authorities^{1,2}. In 2016, the twenty first Century Cures Act amplified interest in RWE through a provision recommending that the United States Food and Drug Administration (FDA) to identify ways to incorporate it into regulatory decisions³. Typical sources of real-world data used to generate RWE include electronic

health records, insurance claims, disease registries, and patient-generated data². These sources are rich with data generated each time an individual interacts with the healthcare system, including clinic visits and prescription claims; however, these data are typically collected for reasons other than research (e.g. recording a clinical encounter or submitting claims for reimbursement) so there are challenges in adapting them into the context of real-world research⁴.

RWE shows great promise in informing decisions for oncology patient care given the complexity of the disease(s), the large number of guideline-based treatment options, and

CONTACT Catherine M. Lockhart  clockhart@bbcic.org  Biologics and Biosimilars Collective Intelligence Consortium, 675 North Washington Street, Suite 220, VA, 22314, USA

 Supplemental data for this article can be accessed online at <https://doi.org/10.1080/13696998.2023.2187196>

© 2023 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.
www.tandfonline.com/ijme

the continually growing arsenal of therapeutic choices. Administrative claims are commonly used in oncology research to assess real-world outcomes and evaluate drug utilization, which is particularly important in the US where healthcare is often fragmented and patients may receive care through different providers, necessitating the use of claims to capture the full breadth of healthcare use⁴⁻⁸. While there are some limitations to conducting research with claims data, such as a lack of laboratory findings or patient-reported outcomes, they can identify encounters and diagnoses, medications received, and clinical outcomes. However, conducting observational oncology studies from administrative claims presents unique challenges, particularly in identifying appropriate patient cohorts and the receipt of multi-agent chemotherapy regimens^{4,8-14}.

BBCIC experience

The Biologics and Biosimilars Collective Intelligence Consortium (BBCIC) is a non-profit, multi-stakeholder, scientific collaborative of partners from industry, professional organizations, non-profit agencies, and health care systems in the United States, with a mission to generate reliable RWE that examines the safety and effectiveness of biologics in order to improve public health¹⁵. BBCIC leverages the infrastructure built by the American Food and Drug Administration (FDA) Sentinel System (<https://www.sentinelinitiative.org/>), and the BBCIC distributed research network (DRN) includes curated claims data representing approximately 95 million patient years across five Research Partners (CVS Health Clinical Trial Services, Harvard Pilgrim Health Care Institute, HealthCore, Inc (Elevance Health), HealthPartners Institute, Kaiser Permanente Washington Research Institute). Curated claims are those which have been submitted to the Research Partners for payment of healthcare services and medications on behalf of hospitals, clinics, and pharmacies, with duplicate claims removed prior to analysis. While the primary mission of BBCIC is to conduct real-world comparative safety and effectiveness research of commercially available biologic drugs, BBCIC also drives the evolution of observational research methodology and identifies solutions to challenges and gaps in claims-based research. Herein we review some challenges of identifying appropriate patient cohorts and treatment regimens specific to cancer therapy and describe our experience while conducting a complex comparative safety and effectiveness study of granulocyte colony stimulating factor (G-CSF) use.

Patient protection

Institutional Review Board (IRB) review was conducted and found that this study did not meet the definition of human subjects research, consistent with the Office of Human Research Protections which deemed research conducted using secondary claims data to be exempt from Human Subjects Research.

Methods

We iteratively developed and tested a de novo algorithm to use administrative claims from the BBCIC DRN to accurately identify patients for a retrospective study of the safety and effectiveness of prophylactic G-CSF (filgrastim, pegfilgrastim), including G-CSF biosimilars in adults diagnosed with lung, breast, colon, ovarian, pancreatic, testicular, cervical, uterine, or non-Hodgkin lymphoma (NHL) cancer who received chemotherapy regimens per National Comprehensive Cancer Network (NCCN) guidelines.

We used an initial patient data set of binary values indicating the absence or presence of criteria for each patient to develop de novo computer code (SAS/STAT software, version 9.4; SAS Institute, Inc.) and identify patient cohorts according to the initial protocol inclusion and exclusion criteria applied in the order listed here:

Inclusion criteria

1. First occurrence of one inpatient diagnosis, or two outpatient diagnoses at least 30 days apart and within a 12-month period, for a cancer of interest during the study period (March 1, 2015-June 30, 2020).
2. Receipt of myelosuppressive chemotherapy regimens with high or intermediate FN risk (as defined by NCCN Guidelines) on or after the cancer diagnosis of interest and during the study period.
3. Medical or pharmacy claims for G-CSF occurring up to 7 days after completion of the chemotherapy regimen during the first chemotherapy cycle.
4. Minimum 365 days of health plan enrollment and continuous pharmacy coverage, with gaps of ≤ 45 days permitted, prior to the index date (defined as the first chemotherapy administration during the study period) and a minimum of 90 days after the index date.
5. Age 20 years or older at the index date.

Exclusion criteria

1. One inpatient or two outpatient cancer diagnoses at least 30 days apart in the 183 days prior to the index date for cancers that differ from the enrolling cancer diagnosis.
2. Receipt of any chemotherapy in the 183 days prior to the index date.
3. Any medical or pharmacy claim for a G-CSF product in the 183 days prior to the index date.
4. Two or more medical claims at least 30 days apart for a skilled nursing facility or hospice care in the 183 days prior to the index date.
5. Two or more diagnoses/procedure codes at least 1 day apart for cancer-related radiotherapy, indicators of metastatic disease, receipt of bone-targeted agents, bone marrow or stem cell transplant, diagnosis of HIV/AIDS, severe hepatic disease, chronic kidney disease, or any non-oncology related neutropenia from 183 days prior to the index date.

The program was tested with data from one Research Partner to resolve coding errors. A waterfall plot mapping attrition after each criterion was tabulated to assess the impact of each criterion on the final cohort size. One of the principal investigators (PP) conducted a chart review of a small random sample of patients in the cohort to verify that appropriate patients were included and excluded, plus medications of interest were ordered/administered. We took an iterative approach to program modifications to test the impact of protocol changes (i.e. revisions of inclusion and exclusion criteria) and to find an optimal algorithm. We describe the challenges and successes during algorithm development to identify patient cohorts.

Results

The initial test with data from one Research Partner resulted in unexpectedly low counts of patients across all diagnosis cohorts (Table 1). Overall, we initially identified 22,859 patients with diagnoses for any cancer of interest as defined in Inclusion Criterion 1, based on a diagnosis code list provided for the reader's review (Supplemental Materials). When we applied Inclusion Criterion 2, only 2,814 (12%) of cases received a chemotherapy regimen of interest per NCCN guidelines¹⁶. (Table 1) We have provided the chemotherapy regimens of interest as well as codes used to identify medication of interests from claims in Supplemental Materials. The final cohort included only 420 patients with cancer diagnoses who received chemotherapy and G-CSF, which was far lower than expected based on clinical experience and knowledge of the patient population (Table 1)¹⁷.

To better characterize the cohorts and to explain the lower-than-expected patient population, we categorized patients according to cancer diagnosis, and then applied the algorithm to identify chemotherapy regimens of interest. The low numbers of patients in each disease cohort receiving chemotherapy relative to the known population at the lead data site suggested that we were not capturing all eligible individuals for this study. We also compared our initial cohort to known treatment trends in national data. For example, a recent national assessment of cancer treatment and survivorship in the US found that among colorectal cancer patients, 9% of Stage I-II, 66% of Stage III, and 36% of Stage IV patients received chemotherapy¹⁸. Furthermore, of

patients diagnosed with colorectal cancer, 20% were Stage I at diagnosis, 22% at Stage II, 23% at Stage III, 20% at Stage IV, and 15% unknown stage. While our data do not include cancer stage, if we take the most conservative estimate that all patients have Stage I-II disease, which is unlikely, we expect at least 9% of patients with colorectal cancer identified in our cohort would receive chemotherapy. We initially found only 21 patients comprising 1% of our population, suggesting we failed to capture potentially eligible patients.

Several potential solutions were identified based on our detailed investigation of the data records, including patient chart review to confirm the algorithm approach, and subsequent testing in an iterative process as described in Table 2. Table 3 lists the results of each Revision that led to the final algorithm and cohort to include in the study. In Revision 1, we reversed the order of the first two inclusion criteria to first identify patients receiving chemotherapy drugs of interest during the study period followed by including only patients with a cancer diagnosis of interest any time before chemotherapy receipt. This resulted in correctly excluding patients who were screened for cancer, thus having a diagnosis code in their record even though they were found not to have cancer. This also included eligible patients who were previously excluded based on cancer diagnosis prior to the study period.

The study investigators manually reviewed a random sample of medical records for patients with breast cancer, colon cancer, or non-Hodgkin lymphoma who did not receive chemotherapy to better understand if patients who did not receive chemotherapy were excluded due to overly stringent regimen algorithms. Across these three cohorts, 59 records were reviewed. Most cases were correctly excluded for reasons including chemotherapy not clinically indicated ($n=33$), treatment receipt prior to the index date ($n=10$), patient declined treatment ($n=6$), and the chemotherapy regimen administered was not aligned with NCCN guidelines, thus was not of interest to the study protocol ($n=2$). Of the remaining eight cases, three had incomplete records (e.g. missing diagnosis codes), two should not have been included based on the chemotherapy administration timeline (i.e. chemotherapy treatment was completed prior to the study period) and incorrect diagnosis codes. Three of the colon cancer patients received appropriate chemotherapy according to medical records but were incorrectly excluded, which suggested the regimen algorithms were potentially too stringent.

We assessed each patient cohort and the chemotherapy algorithms to identify potential approaches to alleviate the identified data capture problem. Using the colorectal cancer cohort, we parsed out the number of patients with a code for at least one chemotherapy agent in the regimen including fluorouracil, leucovorin, and oxaliplatin (FOLFOX), with or without irinotecan (FOLFOXIRI). A total of 627 patients were initially identified. Since the index date marked the first receipt of a chemotherapy regimen of interest, we found that 132 patients were incorrectly excluded due to the cancer diagnosis occurring prior to the study period. This left 495 patients who received at least one chemotherapy agent

Table 1. Percent of patients with diagnoses of interest who also received chemotherapy regimens of interest based on the initial algorithm and expected based on literature.

Cancer Type	Chemotherapy Receipt (%)	
	Initial Algorithm	Expected
Breast	15%	>17% ²³
Cervical	<1%	>32% ²⁴
Colorectal	1%	>9% ¹⁸
Lung	20%	>18% ²³
NHL	16%	>82% ²³
Ovarian	33%	>63% ²⁵
Pancreatic	<1%	>43% ²⁶
Testicular	13%	>6% ²⁷
Uterine	<1%	>7% ²⁸

Abbreviations. NHL, non-Hodgkin lymphoma.

Table 2. Description and rationale of iterative revisions to identify patient cohorts.

	Description of Change	Rationale
Revision 1	Reverse order of first two inclusion criteria	<ul style="list-style-type: none"> Eliminated patients screened but who do not have cancer Retained patients with initial cancer diagnosis outside 12-month look-back period Timing of claims did not necessarily reflect clinical care (e.g. colorectal patients with claim for 5-FU on ≥ 1 day assumed to have received a 2-day regimen in FOLFOX) Multi-indication treatments that did not influence FN risk were not included Patients may fill oral prescription days to weeks prior to infusion Removed - not commonly used in clinical practice per clinician advisement Removed - potential interaction due to bleomycin and G-CSF per clinician advisement Patients may fill outpatient G-CSF prescriptions separately, so allowing 7 days before and after infusion allows for a more accurate capture of G-CSF use This exclusion was too narrow and eliminated patients even with early-stage or other cancers without G-CSF or chemotherapy exposure To improve the likelihood that patients are placed in the disease cohort that most likely reflects their chemotherapy and G-CSF use More accurately capture chemotherapy use, accounting for differences in the timing of claims vs. clinical care Exclusions were done after initial data extraction Moved to earlier in the exclusion process to streamline data extraction Were not always identified; the issue was resolved when the 5-day window for regimen identification was implemented Exclusions were done after initial data extraction
	Chemotherapy claim on ≥ 1 day assumed to receive all doses	
	Oral prednisone and methotrexate not used for case identification or exclusion	
	Remove cisplatin + vinorelbine in lung cancer	
	Remove BEP regimen in testicular cancer	
Revision 2	G-CSF treatment window	<ul style="list-style-type: none"> More accurately capture chemotherapy use, accounting for differences in the timing of claims vs. clinical care Exclusions were done after initial data extraction Moved to earlier in the exclusion process to streamline data extraction Were not always identified; the issue was resolved when the 5-day window for regimen identification was implemented Exclusions were done after initial data extraction
	Prior cancers exclusion	
	Categorize on most recent diagnosis	
Revision 3	5-day window to identify chemotherapy drugs	<ul style="list-style-type: none"> Exclusions were done after initial data extraction Moved to earlier in the exclusion process to streamline data extraction Were not always identified; the issue was resolved when the 5-day window for regimen identification was implemented Exclusions were done after initial data extraction
	Prior chemotherapy exclusion	
	Separate exclusion for prior radiation therapy	
	Include paclitaxel regimens for lung cancer	
Revision 4	Metastatic disease or bone-targeting agent exclusion ≤ 183 days	<ul style="list-style-type: none"> Exclusions were done after initial data extraction

Table 3. Waterfall describing the number of patients included after applying inclusion and exclusion criteria to data from one Research Partner.

A. Initial Algorithm		B. Revised Algorithm				
Inclusion Criteria	No. Included	Revision 1	Revision 2	Revision 3	Revision 4	
1 – cancer diagnosis	22,859	1 – chemotherapy	5,320	5,182	13,600 ^a	13,600 ^a
2 – chemotherapy	2,814	2 – cancer diagnosis	3,645	3,396	6,010	6,010
3 – G-CSF claim	1,318	3 – G-CSF claim	1,501	1,476	1,779	1,801
4 – 365 days enrollment	843	4 – 365 days enrollment	931	916	1,093	1,116
5 – age 20 years or older	842	5 – age 20 years or older	930	915	1,086	1,109
Exclusion Criteria						
–	–	1a – prior radiation	Not Excl.	Not Excl.	1,081	1,068
1 – prior cancer	781	1b – prior cancer	853	Not Excl.	Not Excl.	Not Excl.
2 – prior chemotherapy	631	2 – prior chemotherapy	702	727	Not Excl.	Not Excl.
3 – prior G-CSF	610	3 – prior G-CSF	668	689	962	958
4 – prior SNF or hospice	610	4 – prior SNF or hospice	668	689	962	958
5 – clinical exclusions	420	5 – clinical exclusions	449	462	620	886
Final Cohort	420					886

(A) First data run; (B) Results of each incremental revision to the cohort identification algorithm.

Abbreviations. G-CSF, granulocyte-colony stimulating factor; SNF, skilled nursing facility.

^aAll records within a 5-day window; individuals may contribute more than one record.

and had a colorectal cancer diagnosis during the study period. Most notably, 241 patients with records of FOLFOX receipt, and 136 patients who received FOLFOXIRI were not included in the original cohort but should have been. By loosening the criteria defining multi-agent chemotherapy regimens (e.g. if a claim for a multi-dose agent appeared on at least one day, we assumed the patient received the complete dosing regimen), which better accounted for misalignment in the timing of billing claims and clinical care receipt.

In Revision 2, we widened the G-CSF dosing window to account for patients filling outpatient G-CSF prescriptions within a week on either side of chemotherapy administration, and removed the exclusion for prior cancer diagnoses. Patients were also categorized according to the most recent cancer diagnosis code prior to the index date of chemotherapy receipt, which allows for patients who may have multiple diagnoses in their medical history, including prior cancers treated with chemotherapy not considered high- or

intermediate-risk for FN. While these changes resulted in lower numbers than Revision 1, it increased the available number of patients compared to the initial algorithm, and the percentage of patients included was higher as patients with chemotherapy in the distant past for a prior cancer would no longer be excluded.

In Revision 3 we similarly extended the time window for chemotherapy receipt to allow for differences between the timing of claims, and clinical administration of medications. We also applied some of the planned exclusion criteria manually after initial data extraction to manage complex cases, allowing for expert clinical judgment in categorizing patients. Such cases were reviewed by two pharmacists with extensive oncology training (PAP, CLM).

We further explored how the chemotherapy drugs were coded, and if the expected day of clinical administration (e.g. same day for multiple agents that are typically administered together) corresponded to the day when the billing code was recorded for the chemotherapy agents. Of 442 patients who received all FOLFOX or FOLFOXIRI agents at some point after a colorectal cancer diagnosis, only 355 (80%) patients had drug codes appearing all on the same day. Of those 355, we evaluated how the fluorouracil claims were recorded, and found that 181 (51%) had two separate fluorouracil claims on Day 1 and none on Day 2, 167 (47%) had one fluorouracil claim on Day 1 and none on Day 2, and 7 (2%) had one fluorouracil claim on each of Day 1 and Day 2. We know that specific medication dose is not always available in claims, and our analysis suggests the bolus and continuous infusion doses are not coded consistently in administrative claims. By requiring the timeline of billing claims to match expected clinical administration, we misclassified 377 (38%) individuals and excluded them from the original cohort, emphasizing that the timing of claims does not necessarily align with the schedule of clinical care.

In Revision 4 we moved the exclusion of metastatic disease to occur after the initial cohort definition to allow for a more accurate cohort definition as metastases are not well characterized in administrative claims¹³. This definition included the following International Classification of Disease (ICD) codes: C79.9 (ICD-10-CM, "metastatic disease, not otherwise specified (NOS)" and ICD-9-CM, "secondary malignant neoplasm of specified sites." Additionally, we used clinical judgment and the presence of codes for common anatomical sites of cancer metastases (e.g. brain, bone, liver) that were not specified as primary anatomical sites for study. .

Each Revision is detailed in [Table 3](#), resulting in a final cohort of 886, more than double the 420 patients identified initially. The final algorithm is illustrated in [Figure 1](#).

Discussion

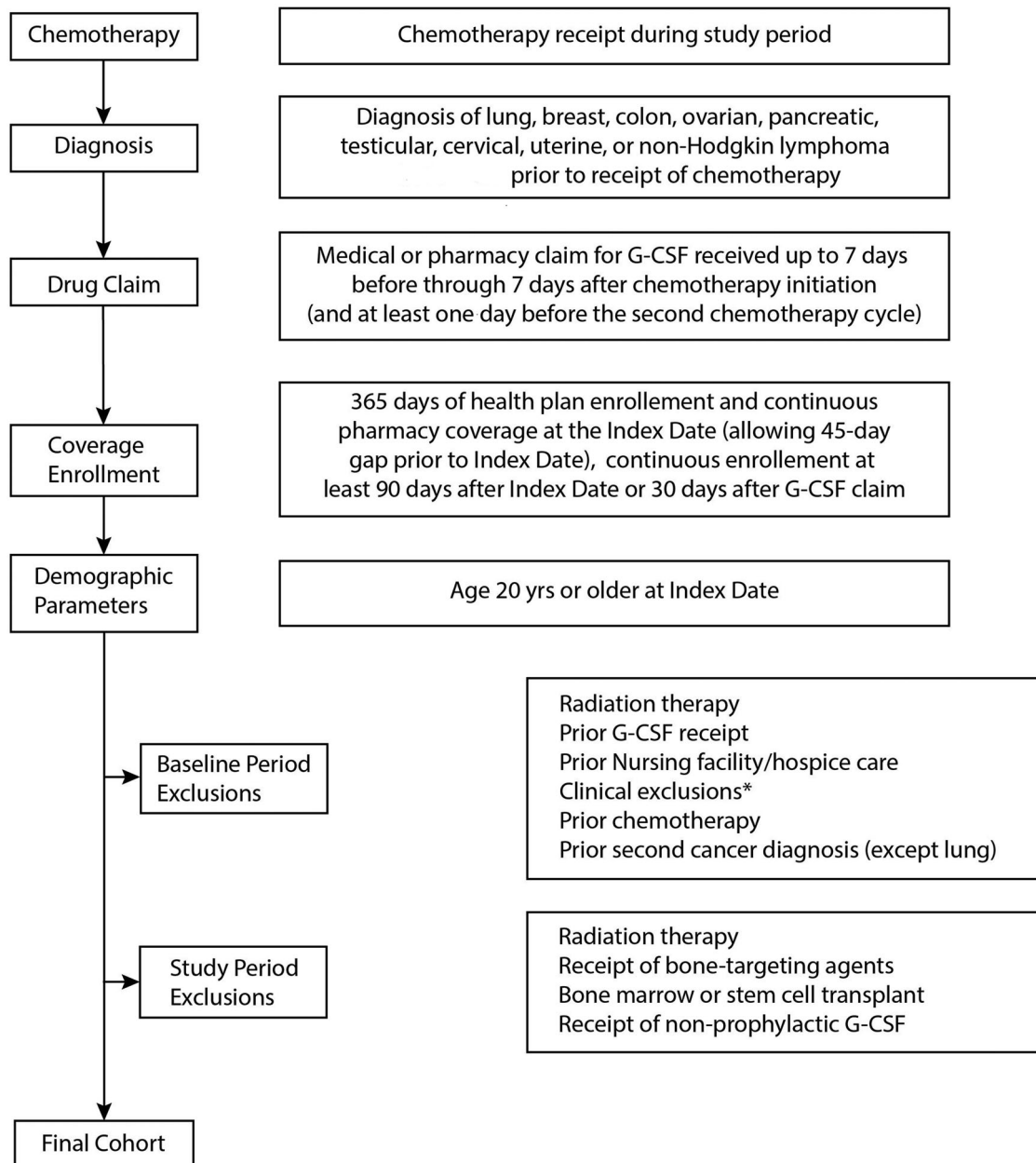
In this study we developed, tested, and applied multiple revisions to an algorithm to identify and categorize cohorts of patients receiving multi-agent chemotherapy regimens for select cancer diagnoses where G-CSF products are commonly used for FN prophylaxis in conjunction with chemotherapy receipt. Through an iterative process, in collaboration with

data scientists, programmers, and clinical experts, we investigated a series of algorithm revisions to improve the accuracy of correctly categorizing patients in cohorts. These revisions increased the cohort by over two-fold compared with the initial algorithm, and with greater confidence in the accuracy of the final study cohort.

Proper identification of cohorts that include multi-agent chemotherapy regimens from administrative claims is a known challenge^{19,20}. Following our discovery of the challenges, we identified opportunities and developed solutions to more reliably assemble an appropriate patient cohort. Several lessons emerged from this work: 1) Identify patients receiving a chemotherapy treatment regimen of interest prior to including patients with cancer diagnoses of interest; this results in the inclusion of patients who did in fact have an appropriate cancer diagnosis that may have initially occurred outside of the observation period, and also retains patients who may have administration codes for chemotherapy that occurred prior to diagnosis, highlighting the variability in the timing of claims versus actual treatment receipt. 2) Medications that may be administered orally with outpatient prescriptions filled separately from the infusion day (e.g. prednisone, methotrexate) can lead to the exclusion of patients who may fill the prescription days to weeks prior to infusion. Additionally, medications that may be used for other indications can result in incorrect exclusions. For example, methotrexate is often used alone or in combination with chemotherapy, but is also one of the most commonly used disease modifying drugs to treat rheumatoid arthritis and other autoimmune disorders²¹. Since prednisone and methotrexate themselves do not increase the risk of FN we removed those from the exclusion criteria. For example, removal of the prednisone and methotrexate exclusions resulted in a 10-fold increase in lymphoma patients identified.

When applying an algorithm, we cannot understate the importance of designing clinically meaningful inclusion and exclusion criteria and thoroughly evaluating the output relative to the population of interest. An ongoing, iterative review between programmers, investigators, and clinicians, plus a comparison against other data sources (e.g. EHR review and tumor registry case counts) was critical in determining whether the output was clinically sound. We recognize that such reviews can be burdensome and costly, which is why we performed this iterative review with one data partner before deployment to other sites. Our findings provided us with information that we were accurately capturing the population of interest. The process we detail here may be used in future studies to avoid exhaustive clinical reviews. Additionally, this approach can be used beyond chemotherapy identification or use of G-CSF, so long as researchers apply appropriate clinical context and administration guidance to the medication under analysis.

It is important to emphasize that administrative claims are created for reasons other than documenting clinical care, such as billing and payment, and are not specifically designed for research purposes. Therefore, the ability to use the algorithm in claims data linked with electronic or other



*Clinical exclusions = bone marrow or stem cell transplant, diagnosis of HIV/AIDS, severe hepatic disease, chronic kidney disease, any non-oncology related neutropenia

Figure 1. Final algorithm to identify cohorts of patients receiving multi-agent chemotherapy.

health records was essential to verifying the results of algorithmic changes and increasing investigator confidence in identified patient cohorts. We also found the criteria for defining regimens in claims should not be overly stringent, as claims do not always reflect the timing or mimic actual delivery of clinical care such as not billing each day of a multi-day chemotherapy regimen. For example, one recent study found that 4% of chemotherapy receipt was recorded in patient medical records, but did not appear in claims²². Furthermore, billing codes are occasionally incomplete and may not include every drug product in a regimen which could be due to coding errors, billing practices (e.g. bundled payments), or the timing of billing relative to expected

clinical treatment. Finally, while it may require more effort, it is worthwhile to implement some exclusions manually following a thorough clinical review of the data output after identification of the initial cohort to ensure appropriate patients are excluded.

Our approach has some limitations. Administrative claims data for observational research has inherent challenges in that they are collected for reasons other than research, which may lead to inaccurate or inconsistent results in cohort identification. However, given the limitation that the US healthcare system does not have a cohesive electronic records system, administrative claims are a way to capture all health care use per patient despite the known risk for coding

errors that may confound, or at least complicate, cohort identification. The BBCIC DRN represents a geographical distribution of patients across the US, it may not be nationally representative. While some Commercial Medicare (e.g. Medicare Advantage) data are available, government-administered Medicare or Medicaid data, or those from other government insurance plans are not included. Additionally, while claims are a good resource for RWE studies, some assumptions may be difficult to validate based on the nature of observational claims research. For example, as cancer stage is not clearly delineated in claims data, identifying metastatic disease poses additional challenges to researchers. This is of particular importance since treatment regimens vary depending on whether the patient is receiving treatment with curative or palliative intent. Finally, in testing and iterating at only one site there is a risk of overfitting the algorithm, resulting in a lack of generalizability to other data sources, in spite of our thorough understanding of all BBCIC DRN data. We took several additional steps to avoid that risk: 1) ensured capture of chemotherapy drugs associated with the treatment regimen and eliminated only drugs from the algorithm that would not be administered in the clinic during the infusion, 2) widened the treatment window to allow for lags or variations in billing relative to clinical care, and 3) changed the order of inclusion criteria and moved some exclusion criteria until after initial cohort extraction to allow for data review without adding or removing criteria to satisfy a single site.

This study provides critical information on the ability to accurately identify patients receiving multi-agent chemotherapy regimens for select cancers of interest. In noting our iterative research process, we offer potential solutions for other researchers to manage the challenges of performing oncology studies with administrative data. Future studies should include extrapolation of this approach to additional data sources, and application of these algorithms to generate cohorts for comparative safety and effectiveness studies from administrative claims.

Transparency

Declaration of funding

This work was supported by the Biologics and Biosimilars Collective Intelligence Consortium (BBCIC), a non-profit, multi-stakeholder collaborative. The work herein is that of the BBCIC consortium and does not reflect the views of any individual participant organization and the content is solely the responsibility of the authors and does not necessarily represent the official views of the BBCIC.

Declaration of financial/other relationships

No potential conflict of interest was reported by the author(s).

Author contributions

All authors were involved in protocol development and study design; PAP, CLM, were the co-primary investigators, GVB, TAD, and CML contributed extensively to the analysis and interpretation of the data with input from all other authors; all authors provided critical review and

thorough feedback on drafts and the final manuscript. All authors share accountability for this work.

Acknowledgements

Thank you to all members of the BBCIC G-CSF Research Team: Jaclyn Bosco (IQVIA), Maria Bottorff (HOPA), Heidi Bailly (HealthPartners Institute), Terese A Defor (HealthPartners Institute), Djeneba Audrey Djibo (CVS Health Clinical Trial Services), Elizabeth Englehardt (Anthem), Cynthia Holmes (Abbvie), Aziza Jamal-Allial (HealthCore Inc, Elevance Health), Annemarie Kline (CVS Health Clinical Trial Services), Edward Li (Sandoz), Sam Li (University of Tennessee), Nancy Lin (IQVIA), Catherine M. Lockhart (BBCIC), James Marshall (Harvard Pilgrim Health Care Institute), Ali McBride (University of Arizona), Cara McDermott (BBCIC), Cheryl McMahill-Walraven (CVS Health Clinical Trials Services), Aaron Mendelsohn (Harvard Pilgrim Health Care Institute), Pamala Pawloski (HealthPartners Institute), Gabriela Vazquez Benitez (HealthPartners Institute), Gary Yee (University of Nebraska).










Data availability statement

None.

Reviewer disclosures

Peer reviewers on this manuscript have no relevant financial or other relationships to disclose.

ORCID

Catherine M. Lockhart  <http://orcid.org/0000-0001-7401-2205>
 Cara L. McDermott  <http://orcid.org/0000-0002-2921-1421>
 Aaron B. Mendelsohn  <http://orcid.org/0000-0002-0560-1177>
 Ali McBride  <http://orcid.org/0000-0002-9430-4675>
 Minghui Sam Li  <http://orcid.org/0000-0003-1909-6260>
 Aziza Jamal-Allial  <http://orcid.org/0000-0001-7039-8833>
 Djeneba Audrey Djibo  <http://orcid.org/0000-0003-3378-0691>
 Gabriela Vazquez Benitez  <http://orcid.org/0000-0002-6226-3207>
 Pamala A. Pawloski  <http://orcid.org/0000-0002-1005-5764>

References

- [1] Beaulieu-Jones BK, Finlayson SG, Yuan W, et al. Examining the use of Real-World evidence in the regulatory process. *Clin Pharmacol Ther.* 2020;107(4):843–852.
- [2] Blonde L, Khunti K, Harris SB, et al. Interpretation and impact of Real-World clinical data for the practicing clinician. *Adv Ther.* 2018;35(11):1763–1774.
- [3] 21st Century Cures Act. HR 34. 114th Congress 2015-2016. Available from: <https://www.congress.gov/114/plaws/publ255/PLAW-114publ255.pdf>
- [4] Abraha I, Montedori A, Serraino D, et al. Accuracy of administrative databases in detecting primary breast cancer diagnoses: a systematic review. *BMJ Open.* 2018;8(7):e019264.
- [5] Du X, Goodwin JS. Patterns of use of chemotherapy for breast cancer in older women: findings from medicare claims data. *J Clin Oncol.* 2001;19(5):1455–1461.
- [6] Lamont EB, Herndon JE, 2nd, Weeks JC, et al. Measuring clinically significant chemotherapy-related toxicities using medicare claims from cancer and leukemia group B (CALGB) trial participants. *Med Care.* 2008;46(3):303–308.
- [7] Mendelsohn AB, Marshall J, McDermott CL, et al. Patient characteristics and utilization patterns of Short-Acting recombinant granulocyte Colony-Stimulating factor (G-CSF) biosimilars compared to their reference product. *Drugs Real World Outcomes.* 2021.

- [8] Warren JL, Harlan LC, Fahey A, et al. Utility of the SEER-Medicare data to identify chemotherapy use. *Med Care*. 2002;40: IV55–IV61.
- [9] Funch D, Ross D, Gardstein BM, et al. Performance of claims-based algorithms for identifying incident thyroid cancer in commercial health plan enrollees receiving antidiabetic drug therapies. *BMC Health Serv Res*. 2017;17(1):330.
- [10] Lamont EB, Lan L. Sensitivity of medicare claims data for measuring use of standard multiagent chemotherapy regimens. *Med Care*. 2014;52(3):e15–20–e20.
- [11] Lockhart CM, McDermott CL, Felix T, et al. Barriers and facilitators to conduct high-quality, large-scale safety and comparative effectiveness research: the biologics and biosimilars collective intelligence consortium experience. *Pharmacoepidemiol Drug Saf*. 2020;29(7):811–813.
- [12] Lund JL, Sturmer T, Harlan LC, et al. Identifying specific chemotherapeutic agents in medicare data: a validation study. *Med Care*. 2013;51(5):e27–34–e34.
- [13] Nordstrom BL, Whyte JL, Stolar M, et al. Identification of metastatic cancer in claims data. *Pharmacoepidemiol Drug Saf*. 2012;21(Suppl 2):21–28.
- [14] Warren JL, Mariotto A, Melbert D, et al. Sensitivity of medicare claims to identify cancer recurrence in elderly colorectal and breast cancer patients. *Med Care*. 2016;54(8):e47–54–e54.
- [15] Biologics and Biosimilars Collective Intelligence Consortium. www.bbcic.org
- [16] NCCN Guidelines: Treatment by cancer type. 2022. https://www.nccn.org/guidelines/category_1
- [17] Pawloski PA, McDermott CL, Marshall JH, et al. BBCIC research network analysis of First-Cycle prophylactic G-CSF use in patients treated with High-Neutropenia risk chemotherapy. *J Natl Compr Canc Netw*. 2021;jnccn20268.
- [18] Miller KD, Nogueira L, Mariotto AB, et al. Cancer treatment and survivorship statistics, 2019. *CA A Cancer J Clin*. 2019;69(5):363–385.
- [19] Krzyzanowska MK, Enright K, Moineddin R, et al. Can Chemotherapy-Related acute care visits be accurately identified in administrative data? *JOP*. 2018;14(1):e51–e8.
- [20] Ritzwoller DP, Carroll N, Delate T, et al. Validation of electronic data on chemotherapy and hormone therapy use in HMOs. *Med Care*. 2013;51(10):e67–73–e73.
- [21] Methotrexate Sodium - Prescribing Information. 2022. https://www.accessdata.fda.gov/drugsatfda_docs/label/2016/008085s066lbl.pdf
- [22] McDermott CL, Engelberg RA, Woo C, et al. Novel data linkages to characterize palliative and End-Of-Life care: challenges and considerations. *J Pain Symptom Manage*. 2019;58(5):851–856.
- [23] Cancer Treatment & Survivorship Facts & Figures 2019–2021. 2022. <https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/cancer-treatment-and-survivorship-facts-and-figures/cancer-treatment-and-survivorship-facts-and-figures-2019-2021.pdf>
- [24] Levinson KL, Bristow RE, Donohue PK, et al. Impact of payer status on treatment of cervical cancer at a tertiary referral center. *Gynecol Oncol*. 2011;122(2):324–327.
- [25] Cancer Trends Proress Report. 2021. <https://progressreport>
- [26] Doleh Y, Lal LS, Blauer-Petersen C, et al. Treatment patterns and outcomes in pancreatic cancer: retrospective claims analysis. *Cancer Med*. 2020;9(10):3463–3476.
- [27] Leveridge MJ, Siemens DR, Brennan K, et al. Temporal trends in management and outcomes of testicular cancer: a population-based study. *Cancer*. 2018;124(13):2724–2732.
- [28] Zakem SJ, Robin TP, Smith DE, et al. Evolving trends in the management of high-intermediate risk endometrial cancer in the United States. *Gynecol Oncol*. 2019;152(3):522–527.