



Published in final edited form as:

Pharmacoepidemiol Drug Saf. 2025 February ; 34(2): e70100. doi:10.1002/pds.70100.

Validation of Immune-Related Adverse Event (irAE) Case Definitions in a Real-World Lung Cancer Population

James S. Heyward, MPH^{1,2}, Jodi B. Segal, MD, MPH^{1,3}, Hemalkumar B. Mehta, PhD^{1,*}, Joseph C. Murray, MD, PhD^{4,*}

¹Center for Drug Safety and Effectiveness, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD

²Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland

³Division of General Internal Medicine, Johns Hopkins School of Medicine, Baltimore, MD

⁴Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins School of Medicine, Baltimore, MD

Abstract

Background: Use of real-world data is increasing to examine immune-related adverse event (irAE) incidence and risk factors in immune checkpoint inhibitor (ICI) users. We aimed to validate five case definition algorithms for irAE in a Johns Hopkins lung cancer registry.

Methods: We conducted a retrospective cohort study using linked electronic health record (EHR) and cancer registry data from a large academic healthcare system. The Lung Immunotherapy irAE Monitoring Registry assesses irAEs in a group of patients treated for lung cancer at Johns Hopkins Medicine from 2013 through 2020. We used data from inpatient, outpatient, and emergency department encounters, including ICD-10 codes and medication administration records to classify the presence or absence of irAEs using five distinct algorithms. These algorithms included three that used both diagnosis (Dx) and medication (Rx) codes, one that used Rx codes only, and one that used Dx codes only, ranging from most numerous criteria (most stringent) to least numerous criteria (least stringent). We compared all five algorithms' performances against chart review-ascertained irAE status and reported sensitivity (Se), specificity (Sp), positive predictive value (PPV), negative predictive value (NPV), and C-statistic (C-stat), with 95% confidence intervals (CI). We also explored algorithm performance by specific organ system toxicities and by Common Terminology Criteria for Adverse Events (CTCAE) severity.

Results: The study cohort included 354 patients with ICI exposure for whom chart-review ascertained irAE status was available. A total of 89 (25.1%) experienced at least one irAE (38 pneumonitis, 12 arthritis, 12 colitis, 7 thyroiditis, and others). Across algorithm versions,

Heyward jheywar1@jhu.edu.

*Co-senior authors

Conflict of Interest Statement: All authors disclose no conflicts of interests.

Ethics statement: This study was reviewed by and received approval from the Johns Hopkins Medical Institutions IRB (IRB00303251 & IRB00355453).

sensitivity ranged from 59.3% to 93.2% in descending order of algorithm stringency; specificity ranged from 21.0% to 77.6% in ascending order of algorithm stringency, and positive predictive value (PPV) ranged from 19.1% to 34.7%. The c-statistic ranged from 0.57 (95% CI, 0.53 – 0.61) (Dx codes only) to 0.71 (0.64 – 0.77) (Rx codes only). For severe irAE (CTCAE Grade 3–5), all algorithms performed better than in the primary analysis, and four exceeded the threshold for usefulness as a measurement tool (maximum C-stat: 0.78 (0.71 – 0.85) (Rx codes only). For severe tissue-specific toxicities, algorithmic detection of irAE pneumonitis, colitis and hepatitis performed better than for the overall group of severe toxicities. Generally, the algorithm versions depicted a sensitivity-specificity tradeoff depending on algorithm stringency.

Conclusion: In this validation study of five irAE case definition algorithms, a combination of ICD-10 codes and medication administration codes generally perform well to identify more severe irAE (CTCAE Grade 3–5), and severe pneumonitis, hepatitis, and colitis (common irAEs) among all possible irAE severity levels and sites. Medication codes alone perform well at identifying severe irAE, while the most stringent algorithm (mirroring guideline-recommended irAE treatment) has the highest Sp and PPV. Algorithms have utility for comparing relative risk of irAE between regimens or patient subgroups.

Plain language summary:

Immune-related adverse events (irAE) are a commonly occurring side effect due to the use of immune checkpoint inhibitors (ICI). Because there is no code for irAE, they are difficult to identify in healthcare data. Although researchers use different algorithms to identify irAE, the accuracy of these algorithms is not reported. The authors used electronic health records and cancer registry data from one large health system and evaluated the accuracy of five diagnosis- and medication-code-based algorithms to identify irAE. They found that all algorithms performed well for severe irAE relative to mild and moderate irAE, with medication codes alone algorithms performing well to identify severe irAE cases. This study can help researchers select appropriate algorithms for studies that uses irAE as an outcome.

Keywords

Immune checkpoint inhibitors; immune-related adverse events; lung cancer; cancer immunotherapy; validation; electronic health records; case definition algorithm

INTRODUCTION

Non-small cell lung cancer (NSCLC) is the second most frequently-occurring cancer in the United States (240,000 new cases/year) and is the nation's leading cause of cancer deaths (127,000 deaths/year).^{1,2,3} Immune checkpoint inhibitors (ICIs) received Food and Drug Administration (FDA) approval as second-line therapy in 2014 and were approved for first-line use in 2017.⁴ ICIs facilitate an anti-tumor immune response by neutralizing T-cell disactivating "immune checkpoint" proteins, specifically cytotoxic T lymphocyte antigen 4 (CTLA-4)–CD and programmed cell death 1 (PD-1)/programmed cell death 1 ligand 1 (PD-L1), that are commonly activated by tumor cells, impeding endogenous immune detection.^{5,6}

Beyond intended anti-tumor effects, the immune dysregulation caused by ICIs diminishes self-tolerance and leads to immune-related adverse events (irAE). More than 70% of patients experience these adverse events, and 10% develop severe irAEs.⁷ IrAE can affect every tissue and organ system, although the site, timing, and severity of events differ by the specific ICI molecule and cancer type. Severe irAEs among treated individuals with NSCLC include pneumonitis, colitis, myocarditis, and hepatitis, and often lead to temporary or permanent withdrawal of the ICI therapy.⁸ While irAEs are typically recorded as adverse outcomes in clinical trials of ICI therapies, trials are often short, enroll small number of individuals, and conventionally exclude individuals with underlying autoimmune disorders and individuals with multiple comorbidities who may be at greater risk of severe irAE. Real-world data (RWD), such as billing data or electronic health record data, can complement clinical trial data by including broader populations, more observations, and longer follow-up, which enable more robust measurement of irAE, including among subgroups excluded from clinical trials.

However, because there are no International Classification of Disease (ICD)-10 diagnosis codes for irAE, the occurrence of irAE must be deduced based on other diagnosis and procedure codes, and perhaps pharmacy records and clinician notes, temporally associated with exposure to treatment.^{9,10,11} Despite this challenge, RWD sources have been used to assess the benefits and side effects of ICIs, particularly to evaluate patient populations not well represented in randomized controlled trials.^{9,10,12,13,14,15,16} Researchers have primarily used three algorithms to identify irAEs with RWD.^{9,12,15} However, no study has assessed the accuracy of these irAE case definitions. We aimed to validate case definitions for irAE in a large health system's irAE monitoring registry. We validated one existing and four modified irAE case definitions. A valid case definition will allow more reliable future studies of the safety of these therapies.

METHODS

We conducted a retrospective study using chart review as the reference standard to validate algorithms for identifying irAE. Institutional review board (IRB) approval was received for this study. Informed consent was waived due to minimal risk of harm to subjects, and data were de-identified.

Data Sources

We accessed data from a single healthcare system having linked electronic health records (EHR) and registry data. Data from the parent Johns Hopkins Thoracic Oncology Precision Medicine Center of Excellence (PMCOE) Registry were derived from the Precision Medicine Analytics Platform (PMAP), which is a local common data model designed to facilitate use of our health systems' EHR data, and other imported data, for research and care improvement.¹⁷ Data derived from EHR included laboratory and radiology orders and results; clinical characteristics of diagnosed and treated cancers; sociodemographic and clinical data of patients; inpatient and outpatient medication utilization (pharmacy data); and billing diagnosis and procedure codes. (IRB00303251). This information was linked via common patient identifiers to The Lung Immunotherapy irAE Monitoring Registry (JH-irAE

monitoring registry; IRB00355453), a prospective registry that assesses irAE in patients treated for lung cancer at Johns Hopkins Medicine. The JH-irAE monitoring registry was curated by review of all patients enrolled between 2013–2020 for adjudication of irAEs by physician specialists (oncologists, rheumatologists, endocrinologists, gastroenterologists, dermatologists, and more) who were also part of the multidisciplinary Immunotherapy Toxicity Team that also provided clinical guidance for patients who developed irAEs. The irAE monitoring registry is maintained using the REDCap online database management software.^{18,19}

Study Cohort

We included adults (age ≥ 18 years) with lung cancer who received care at one of six hospitals and one outpatient center in Maryland and the District of Columbia between 2013 and 2020. Individuals were included in this study if they were in the irAE monitoring registry and had at least one record indicating receipt of ICI at Johns Hopkins Medicine no later than October 1, 2020. We excluded individuals with COVID-19 infection diagnosed at in a health system hospital as acute COVID-19 can resemble irAE pneumonitis in the EHR data (eFigure 1). We made no exclusions based on treatment regimen.

Reference Standard Chart-Review Ascertained Definition of irAE

Registry investigators identified irAE based on a hierarchical protocol.²⁰ The registry prioritizes pathologic confirmation of irAE, followed by evidence of clinical improvement with medical management of irAE, or if neither of these were available, multidisciplinary consensus based on chart review by at least two medical oncologists and one specialist from a relevant subspecialty. Pathologic confirmation included laboratory (e.g., endocrinology lab results, or other specialty specific testing) and/or histological (e.g., colon or renal biopsy) confirmation. Radiographic confirmation was not commonly available, but commented on in review of some of the irAEs (e.g., pneumonitis). Clinical improvement was assessed based on use of irAE-directed therapies (e.g., corticosteroids). Consensus was required when cases were equivocal beyond laboratory or pathological diagnosis, and included subspecialists. IrAE were further categorized by the registry investigators according to the confidence of attribution of symptoms to irAE on a scale from 1 (definitely related) to 3 (unrelated), the specific diagnosis (e.g. pneumonitis or cardiomyopathy), and according to severity on the Common Terminology Criteria for Adverse Events (CTCAE) v4.0 scale (0 [no symptoms]-5 [fatal]).

The presence of registry-confirmed irAE of confidence level 1 or 2 was our reference standard (or gold standard) definition of irAE. For our analyses, we restricted true positive cases to the toxicities in 10 anatomical domains (cardiovascular, endocrine, pulmonary, nervous system, gastrointestinal, skin, hematologic, musculoskeletal, ocular, and renal) (eTable 1).⁹ IrAE ascertainment is complete through October 23, 2020.

Five Algorithm Derived Definitions of irAEs

We tested five irAE case definition algorithms in the validation dataset. We adapted one published irAE case definition algorithm and further developed four others, relying on data elements relevant to identification and management of irAE according to clinical guidelines

(Table 1/eTable 2). One of the four algorithms was developed to assess severe irAE only. The algorithms included inpatient, outpatient and emergency department diagnosis ICD-10 codes; inpatient medication administration data, or both. For algorithm version “Dx/Rx”, we adapted a published irAE case definition that includes relevant diagnosis codes based on National Comprehensive Cancer Network (NCCN)/American Society of Clinical Oncology (ASCO) Guidelines,²¹ along with pharmacy records indicating the use of NCCN/ASCO-recommended medications for treatment of the irAE within 40 days before or after the diagnosis date. The rationale for the criteria for the definitions was based on the published irAE case definition and data readily available from the EHR. General diagnosis codes and medication administration data were both available. For our adapted case definition, the specific ICD-10 diagnosis codes corresponding to toxicities in the published case definition were selected based on physician expertise, and irAE medications and dosages of interest were chosen based on the referenced clinical practice guidelines. Given that ICI administrations are scheduled on 14 or 28 days (two or four weeks) cycles for nivolumab and durvalumab, and 21 or 42 days (three or six weeks) cycles for pembrolizumab, the choice of within 40 days before or after irAE diagnosis suggests that the corticosteroid was dosed within one or two cycles of irAE diagnosis. We modified, added, or removed requirements, relative to version “Dx/Rx”, for the four other algorithm versions (Table 1/eFigure 2).

Statistical Analyses

We used descriptive statistics to characterize patients with and without irAE based on the reference standard. For the primary analyses, we compared the algorithms’ performances against reference irAE status and reported sensitivity (Se), specificity (Sp), positive predictive value (PPV), negative predictive value (NPV), and C-statistic, with 95% confidence intervals (CI), for each algorithm. The C-statistic, a measure of algorithm discrimination, is defined as the probability that the algorithm will correctly differentiate irAE status for two individuals with and without irAE.²² We calculated 95% CI based on the standard error of each statistic (sensitivity (Se), specificity (Sp), positive predictive value (PPV), negative predictive value (NPV)) and with bootstrapping (C-statistic (C-stat)). For our primary analysis, we used the C-statistic to test for differences in performance of the algorithms.

Characterizing False Negative and False Positive Occurrences

We characterized false negative and false positive irAE from algorithm version “Dx/Rx” to understand the factors associated with algorithm misclassifications.

Sub-analyses

We performed two secondary validation sub-studies; the first assessed the accuracy of the irAE algorithms to detect severe irAE (i.e., those classified by registry physicians as CTCAE Grade 3 or higher) and the second assessed the algorithms’ performances for identifying the most prevalent severe irAE (pneumonitis, colitis, and hepatitis). For these analyses, we restricted the reference standard (“true positives”) and algorithm ICD-10 codes accordingly. The Dx/Rx (discontinue) algorithm was used exclusively for the sub-analyses, due to its emulation of guideline-recommended management of severe irAE, but not mild/moderate

irAE. The choice of >60 days of no ICI administration generally aligns with missing at least two cycles of treatment. The authors of the published case definition also used >60 days of no ICI administration to define treatment discontinuation in their analysis of irAE-related outcomes.⁹

RESULTS

Validation Patient Population

There were 354 individuals in the irAE monitoring registry with confirmed exposure to ICI therapy who were not diagnosed with COVID-19 during the time period of irAE ascertainment (eFigure 1). The study cohort was 52% male and 79% White, with a mean age of 65.8 ± 9.8 years (Table 2). Index dates ranged from 2015–2020, with most between 2017–2019 (eTable 3).

Reference Standard: Chart Review-Ascertained irAE

Among the 354 patients, 89 experienced at least one chart review-ascertained irAE, including 63 patients with 1 irAE, 16 patients with 2 irAEs, 9 patients with 3 irAEs; and 1 patient with 4 irAEs. There were 38 irAEs coded as pneumonitis, 12 as arthritis, 12 as colitis, 7 as thyroiditis, 6 as maculopapular eruption, and 3 as hepatitis, with other toxicities less frequent (Figure 1). Additionally, 26 irAEs were coded as “other”, meaning the symptoms were not deemed to match a single diagnosis. Severity varied, with 3 irAEs coded as “no symptoms” (0), 27 as “mild” (1), 46 as “moderate” (2), 35 as “severe” (3), 8 as “life threatening” (4), 6 as “fatal” (5), and one was unclassified. “No symptoms” irAEs were thyroiditis cases, ostensibly ascertained from laboratory values; the unclassified irAE was pneumonitis. In total, 56 of the chart review-ascertained irAE cases (44.4%) required subspecialty consultation.

Algorithm Performance

All irAE—For the four algorithms used in the main analysis, sensitivity ranged from 59.3% (CI: 46.8–71.8) to 93.2% (CI: 86.8–99.6) in descending order of algorithm stringency, while specificity ranged from 21.0% (CI: 16.4–25.6) to 77.6% (CI: 72.8–82.4) in ascending order of algorithm stringency. PPV was low and ranged from 19.1% (CI: 14.6–23.6) to 28.1% (CI: 20.1–36.1), with a positive correlation with algorithm stringency (Table 3). Among the two algorithms requiring both ICD-10 and medication administration codes, the difference in performance between the less stringent (Dx/Rx relaxed) and more stringent (Dx/Rx) was modest: sensitivity (62.7% vs. 59.3%); specificity (74.6% vs. 77.6%); and PPV (33.0% vs. 34.7%). Comparing the two least stringent algorithms, Rx only had much lower sensitivity (69.5% vs. 93.2%) but much higher specificity (72.2% vs. 21.0%), and PPV (33.3% vs. 19.1%).

The C-statistic was highest for the Rx only algorithm (0.71) (CI: 0.64–0.77) and lowest for Dx only (0.57) (CI: 0.53–0.61).

Severe irAE—We included a fifth algorithm (Dx/Rx [discontinue]) in the sub-analyses assessing severe irAE only. Relative to the primary analysis, the algorithms performed

better at identifying severe irAE (CTCAE Grade 3 or higher) in the dataset, with sensitivity generally higher across algorithms (range: 72.4% [CI: 56.1–88.7] to 86.2% [CI: 73.6–98.8]) and specificity similar (range: 19.1% [CI: 14.8–23.4] to 76.9% [CI: 72.3–81.5]) and the rank order of algorithms was preserved from the primary analysis. Owing to increased rarity of the outcome (severe only), the PPV was lower (range: 8.7% [CI: 5.4–12.0] to 21.9% [CI: 13.6–30.2]). The C-statistic for four of the algorithms exceeded 0.70, with Rx only (0.78) (CI: 0.71–0.85) the highest and Dx only (0.53) (CI: 0.50–0.59) the lowest.

Severe organ-specific irAE—The sensitivity of the algorithms for measurement of severe organ-specific irAE (pneumonitis, colitis, and hepatitis) was similar to severe irAE (range: 70.4% [CI: 53.2–87.6] to 85.2% [CI: 71.8–98.6]), with higher specificity (range: 66.7% [CI: 61.6–71.8] to 90.8% [CI: 87.7–93.9]), and higher PPV (range: 16.2% [CI: 9.9–22.5] to 38.8% [CI: 25.2–52.4]). The C-statistic for all of the algorithms exceeded 0.70, with Dx/Rx (discontinue) the highest (0.81) (CI 0.72–0.89) and Dx only the lowest (0.72) (CI 0.63–0.80).

False negatives—Twenty-four patients experienced 29 irAE that were not identified with the Dx/Rx algorithm. By severity, 18/29 (62%) of false negative irAE were CTCAE Grade 2 (moderate) or lower, compared with 29/59 (49%) of the irAE true positive. Among the false negative irAE that were CTCAE Grade 3 (severe) or higher, many were missing a record of treatment for the irAE in the PMAP medication administration data. The frequency and severity of false negative irAE diagnoses was similar to the distribution in the whole validation population.

False positives—Sixty-six patients were identified as having 106 false positive irAE, including 35 thyroiditis; 21 pneumonitis; 14 diabetes; 11 rash; 8 adrenal insufficiency; 4 aplastic anemia; 4 cardiac arrhythmia; 3 colitis; 2 hepatitis; 1 each of encephalitis, renal dysfunction; eye pain; and Crohn's disease. Most false positive thyroiditis cases were due to use of levothyroxine in proximity to a hypothyroidism diagnosis. These false positives were generally related to pre-existing conditions, or cases in which symptoms were treated like irAE, but later attributed to other causes (e.g. radiation-related pneumonitis).

DISCUSSION

In this validation study, we identified EHR-based algorithms that are sufficiently valid for identifying individuals who had severe irAE (CTCAE Grade 3–5). The algorithms generally performed poorly in individuals with mild/moderate irAE. Differential algorithm validity by irAE severity likely stems from the fact that four of the five algorithms required medication codes, which align with guideline-recommended treatment of severe, but not necessarily moderate or mild irAEs. For this reason, too, the Dx code-only algorithm had a much poorer C-statistic for assessing severe irAE than other algorithm versions, though it appeared sufficient for identifying common, “high-confidence” severe irAE.

For all algorithm versions and irAE severity levels, algorithm PPV was low, indicating that a majority of irAE toxicities identified using the algorithms were not identified as true cases by clinician chart reviewers. This surprisingly low proportion arises from the

relatively low incidence of irAE in the validation population and the imperfect specificity of the algorithms (maximum: 90.8% for the most stringent algorithm in one sub-analysis). In the context of real-world clinical management of irAE, the low PPV also reflects the tendency to treat suspected irAE early using guideline-recommended treatments before a definitive diagnosis is made. Many of these may have been later deemed to be due to another cause. This practice was evident in the irAE registry's irAE ascertainment protocol, whereby evidence of improvement after pharmacologic irAE treatment was taken as indicative that the symptoms had been caused by irAE. It is plausible, then, that the low PPV is an artifact of this clinical uncertainty in the early stages of irAE.

Across the main and sub-analyses, the Rx only algorithm outperformed other algorithm versions as assessed using the C-statistic, including being the only algorithm to exceed 0.70 for the overall analysis. This case definition is relatively simple to implement, and appears sufficient to identify patients with severe irAE, relative to other, more complex case definitions. However, given the range of sensitivity and specificities, the algorithms may have different use cases. First, in a clinical research context, where there is interest in identifying "true" irAE to assess etiology and risk factors, the highest specificity algorithm (Dx/Rx Discontinue) might prove most useful. Second, in health care planning, the algorithm with the highest sensitivity will detect most completely all affected patients. Therefore, Dx only or Rx only depending on irAE severity may be more useful. Third, the Dx-only case definition, having a high sensitivity and high NPV, could be useful to identify patients potentially in need of chart review and outcome adjudication in order to inform potential follow-up care for irAE.²³ Fourth, the irAE case definition algorithms could also prove useful in real-world studies comparing the risks across ICI-containing regimens, across patient subgroups, or to measure changes in risk over time, assuming that outcome misclassification is nondifferential across those settings. An imperfectly calibrated measurement tool, while potentially unsuitable to estimate true incidence rates, may be used to contrast irAE, for example, for calculating risk ratios and other measures of association.

Validated irAE case definition algorithms could also be implemented in predictive models aimed at prognosticating the risk of irAE toxicity in ICI-eligible patients. Such predictive models would be valuable to aid clinical decision-making, such as by informing knowledge of individualized risk-benefit balance of ICI²⁴ or prospectively identifying risk factors for severe irAE.¹⁵ In this context, the specific goal of the prediction study would inform the choice of algorithm; but a more sensitive algorithm may result in patients erroneously being classified as high-risk (disadvantageous when follow-up is invasive and/or expensive), while a more specific algorithm might correctly identify irAE-prone patients at the cost of missing some cases (costly when preventing irAE is of predominating concern).

Our study has several strengths. This is the first study to validate EHR-derived case definitions for irAE, which is important given the growing number of studies using EHR and other RWD to assess this important clinical endpoint. The protocol for ascertaining true cases was rigorous, guideline-based, and implemented by an experienced clinical team.²³ We relied on data fields that are available in administrative claims data, enabling the implementation of these irAE algorithms across other data sources. The study also has limitations. Our validation population included research patients in clinical trials, which may

not be generalizable to the broader real-world use population. We also excluded patients with a diagnosis of COVID-19, limiting generalizability in settings where this diagnosis is likely to be prevalent. We used only one health system's data, which may not generalize to all patients and settings of care. We used diagnosis and medication codes only, as opposed to other EHR-available fields such as laboratory and radiology results for rule-out diagnoses, which may have improved algorithm performance. Unstructured clinical notes, in addition, contain important data that can be parsed by natural language processing models in settings such as this one, where diagnostic uncertainty is high.²⁵ Diagnosis codes alone can be unreliable to ascertain specific conditions, as signs and symptoms are possibly coded differentially depending on the presentation and billable medical care received (e.g., "hypoxic respiratory failure" instead of "pneumonitis"). Diagnosis codes also may not be consistently entered into EHR, leading to under-ascertainment of cases using this field alone. We used EHR medication administration records, which reflect medications administered by a nurse or other provider in an inpatient setting, as opposed to outpatient medications taken at home. This choice of data source may have caused us to miss mild cases that could be treated outside the hospital. Finally, although our EHR-based case definition is inter-operable with claims data, algorithm performance may differ when implemented in claims due to differential granularity of inpatient care records when relying on claims (e.g., due to diagnosis-related group (DRG)-based coding of inpatient admissions).

CONCLUSIONS

IrAE case definition algorithms using both diagnosis and medication codes perform best at identifying severe irAEs (CTCAE Grade 3–5), and commonly encountered irAEs such as pneumonitis, colitis, and hepatitis. IrAE thyroiditis, which is common and most often low-risk, appears to be unreliably measured using these algorithms. Ultimately, analyses can be adjusted according to known measurement error in irAE case definitions,^{26,27} and algorithms have utility for comparing relative risk of irAE between ICI regimens or between patient subgroups.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements:

We would like to acknowledge annotating contributors to the Johns Hopkins Lung Immunotherapy Database, including: Dr. Jenna Mammen, Dr. Jarushka Naidoo, Dr. Laura Cappelli. We also acknowledge the Thoracic Oncology Precision Medicine Center of Excellence personnel, in particular Durrant Burasa for his support with data access.

Funding: Dr. Mehta is supported by the National Institute on Aging (1K01AG070329). Mr. Heyward is supported by the National Heart, Lung and Blood Institute (T32HL139426). Dr. Murray is supported by the Center for Disease Control (NU01DD000025), National Cancer Institute (P30CA006973), Maryland Cigarette Restitution Fund, and Johns Hopkins inHealth Program.

References

1. Cancer of the Lung and Bronchus - Cancer Stat Facts. SEER. Accessed February 4, 2022. <https://seer.cancer.gov/statfacts/html/lungb.html>

2. Suresh K, Naidoo J, Lin CT, Danoff S. Immune Checkpoint Immunotherapy for Non-Small Cell Lung Cancer. *Chest*. 2018;154(6):1416–1423. doi:10.1016/j.chest.2018.08.1048 [PubMed: 30189190]
3. Duma N, Santana-Davila R, Molina JR. Non-Small Cell Lung Cancer: Epidemiology, Screening, Diagnosis, and Treatment. *Mayo Clin Proc*. 2019;94(8):1623–1640. doi:10.1016/j.mayocp.2019.01.013 [PubMed: 31378236]
4. Onoi K, Chihara Y, Uchino J, et al. Immune Checkpoint Inhibitors for Lung Cancer Treatment: A Review. *J Clin Med*. 2020;9(5):1362. doi:10.3390/jcm9051362 [PubMed: 32384677]
5. Coschi CH, Juergens RA. The Price of Success: Immune-Related Adverse Events from Immunotherapy in Lung Cancer. *Curr Oncol*. 2021;28(6):4392–4407. doi:10.3390/curroncol28060373 [PubMed: 34898551]
6. Martins F, Sofiya L, Sykietis GP, et al. Adverse effects of immune-checkpoint inhibitors: epidemiology, management and surveillance. *Nat Rev Clin Oncol*. 2019;16(9):563–580. doi:10.1038/s41571-019-0218-0 [PubMed: 31092901]
7. Fan Y, Xie W, Huang H, et al. Association of Immune Related Adverse Events With Efficacy of Immune Checkpoint Inhibitors and Overall Survival in Cancers: A Systemic Review and Meta-analysis. *Front Oncol*. 2021;11:633032. doi:10.3389/fonc.2021.633032 [PubMed: 33912454]
8. Choi J, Lee SY. Clinical Characteristics and Treatment of Immune-Related Adverse Events of Immune Checkpoint Inhibitors. *Immune Netw*. 2020;20(1):e9. doi:10.4110/in.2020.20.e9 [PubMed: 32158597]
9. Brown VT, Antol DD, Racska PN, Ward MA, Naidoo J. Real-World Incidence and Management of Immune-Related Adverse Events from Immune Checkpoint Inhibitors: Retrospective Claims-Based Analysis. *Cancer Invest*. 2021;39(10):789–796. doi:10.1080/07357907.2021.1913502 [PubMed: 33829943]
10. Petito LC, García-Albéniz X, Logan RW, et al. Estimates of Overall Survival in Patients With Cancer Receiving Different Treatment Regimens: Emulating Hypothetical Target Trials in the Surveillance, Epidemiology, and End Results (SEER)–Medicare Linked Database. *JAMA Netw Open*. 2020;3(3):e200452. doi:10.1001/jamanetworkopen.2020.0452 [PubMed: 32134464]
11. Sun VH, Heemelaar JC, Hadzic I, et al. Enhancing Precision in Detecting Severe Immune-Related Adverse Events: Comparative Analysis of Large Language Models and International Classification of Disease Codes in Patient Records. *J Clin Oncol*. Published online September 3, 2024;JCO.24.00326. doi:10.1200/JCO.24.00326
12. Tully KH, Cone EB, Cole AP, et al. Risk of Immune-related Adverse Events in Melanoma Patients With Preexisting Autoimmune Disease Treated With Immune Checkpoint Inhibitors: A Population-based Study Using SEER-Medicare Data. *Am J Clin Oncol*. 2021;44(8):413–418. doi:10.1097/COC.0000000000000840 [PubMed: 34081033]
13. Almutairi AR, Slack M, Erstad BL, McBride A, Abraham I. Association of immune-checkpoint inhibitors and the risk of immune-related colitis among elderly patients with advanced melanoma: real-world evidence from the SEER–Medicare database. *Ther Adv Drug Saf*. 2021;12:2042098621991279. doi:10.1177/2042098621991279 [PubMed: 33796257]
14. Anjohrin S, Sheahan A, Suruki R, Stark JL, Sloan VS. Occurrence of Possible Rheumatologic Immune-Related Adverse Events (rh-irAEs) Associated with Immune Checkpoint Inhibitor (ICI) Therapy. *Rheumatol Ther*. 2021;8(4):1651–1659. doi:10.1007/s40744-021-00359-z [PubMed: 34491550]
15. Kalinich M, Murphy W, Wongvibulsin S, et al. Prediction of severe immune-related adverse events requiring hospital admission in patients on immune checkpoint inhibitors: study of a population level insurance claims database from the USA. *J Immunother Cancer*. 2021;9(3):e001935. doi:10.1136/jitc-2020-001935 [PubMed: 33789879]
16. Zheng Y, Kim R, Yu T, et al. Real-World Clinical and Economic Outcomes in Selected Immune-Related Adverse Events Among Patients with Cancer Receiving Immune Checkpoint Inhibitors. *The Oncologist*. 2021;26(11):e2002–e2012. doi:10.1002/onco.13918 [PubMed: 34327774]
17. Precision Medicine Analytics Platform (PMAP) – Institute for Clinical and Translational Research. Accessed July 25, 2023. <https://ictr.johnshopkins.edu/service/informatics/pmap/>

18. 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. doi:10.1016/j.jbi.2019.103208 [PubMed: 31078660]
19. 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):377–381. doi:10.1016/j.jbi.2008.08.010 [PubMed: 18929686]
20. Hsu ML, Murray JC, Psoter KJ, et al. Clinical Features, Survival, and Burden of Toxicities in Survivors More Than One Year After Lung Cancer Immunotherapy. *The Oncologist.* 2022;27(11):971–981. doi:10.1093/oncolo/oyac140 [PubMed: 35972337]
21. Brahmer JR, Lacchetti C, Schneider BJ, et al. Management of Immune-Related Adverse Events in Patients Treated With Immune Checkpoint Inhibitor Therapy: American Society of Clinical Oncology Clinical Practice Guideline. *J Clin Oncol Off J Am Soc Clin Oncol.* 2018;36(17):1714–1768. doi:10.1200/JCO.2017.77.6385
22. Pencina MJ, D’Agostino RB Sr. Evaluating Discrimination of Risk Prediction Models: The C Statistic. *JAMA.* 2015;314(10):1063–1064. doi:10.1001/jama.2015.11082 [PubMed: 26348755]
23. Naidoo J, Zhang J, Lipson EJ, et al. A Multidisciplinary Toxicity Team for Cancer Immunotherapy-Related Adverse Events. *J Natl Compr Cancer Netw JNCCN.* 2019;17(6):712–720. doi:10.6004/jnccn.2018.7268 [PubMed: 31200355]
24. Lippenszky L, Mittendorf KF, Kiss Z, et al. Prediction of Effectiveness and Toxicities of Immune Checkpoint Inhibitors Using Real-World Patient Data. *JCO Clin Cancer Inform.* 2024;(8):e2300207. doi:10.1200/CCI.23.00207 [PubMed: 38427922]
25. Gupta S, Belouali A, Shah NJ, Atkins MB, Madhavan S. Automated Identification of Patients With Immune-Related Adverse Events From Clinical Notes Using Word Embedding and Machine Learning. *JCO Clin Cancer Inform.* 2021;(5):541–549. doi:10.1200/CCI.20.00109 [PubMed: 33989017]
26. Edwards JK, Cole SR, Troester MA, Richardson DB. Accounting for Misclassified Outcomes in Binary Regression Models Using Multiple Imputation With Internal Validation Data. *Am J Epidemiol.* 2013;177(9):904–912. doi:10.1093/aje/kws340 [PubMed: 24627573]
27. Lyles RH, Tang L, Superak HM, et al. Validation Data-Based Adjustments for Outcome Misclassification in Logistic Regression: An Illustration. *Epidemiol Camb Mass.* 2011;22(4):589–597. doi:10.1097/EDE.0b013e3182117c85

Key points:

1. No International Classification of Disease (ICD) code exists for irAE, but researchers are using different algorithms to identify irAE in real-world data. We validated irAE algorithms against chart review data.
2. We found that diagnosis- and medication-code based irAE algorithms perform well to identify severe and commonly occurring irAE in lung cancer patients who take immune checkpoint inhibitors (ICI).
3. irAE algorithms relying on medication codes alone are sufficient to identify severe cases of irAE
4. Our validation study can help researchers identify optimal algorithm for different use cases, including identification of potential cases for outcome adjudication; elucidation of risk factors for developing irAE; and comparison of incidence rates across ICI regimens and patient subgroups.

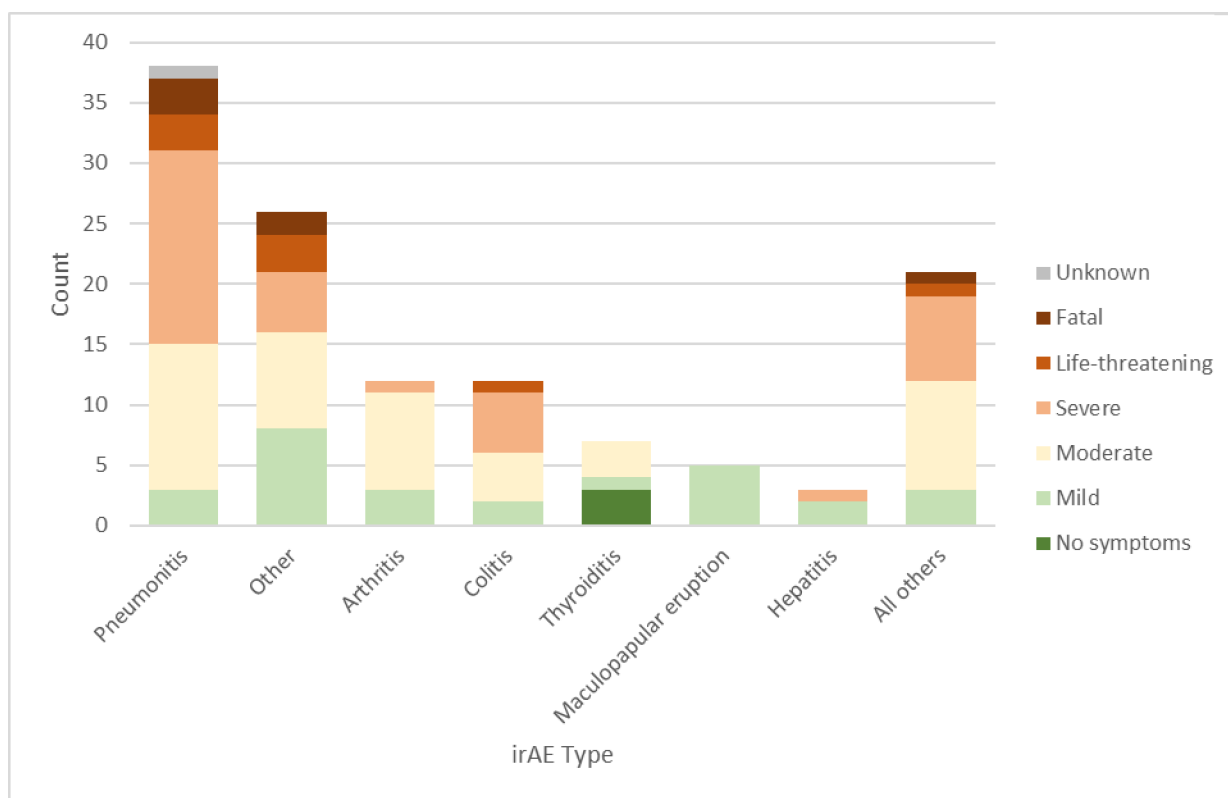


Figure 1. Frequency and Severity of irAE in Validation Population

Cases were ascertained via clinician chart review.

IrAE diagnoses not assessed by case definition algorithm: Arthritis, 12; Other, 26 (not including 3 classified as dermatitis that were included); Hypophysis, 3; Myasthenia gravis, 2; Pancreatitis, 2; Sicca syndrome, 2; Guillain-Barre syndrome, 1; ITP, 1; Optic neuritis, 1; Peripheral neuropathy, 1; Psoriasiform eruption, 1; Type 1 diabetes mellitus, 1; Vasculitis, 1. Specific irAE with $n < 3$ are not depicted here.

Table 1.**Components of irAE Case Definition Algorithms**

Criteria	Discontinue	Dx/Rx	Dx / Rx (relaxed)	Rx only	Dx only
"irAE" dx	✓	✓	✓	✗	✓
Corticosteroid/ DMARD use	✓	✓	✓	✓	✗
Corticosteroid timing	+/- 40 days of "irAE" dx	+/- 40 days of "irAE" dx	- [Time 0]/+ 40 days of "irAE" dx	Any time during ICI	✗
ICI treatment discontinuation	✓	✗	✗	✗	✗
					
Most stringent Least stringent					

Algorithm Dx/Rx was adapted directly from a publication, as previously described. For algorithm Dx/Rx relaxed, we expanded the date buffer for irAE treatment to include any date prior to the ICD-10 code, up to 40 days following. For algorithm Dx/Rx (discontinue), we added a requirement that ICI treatment be discontinued subsequent to the ICD-10 code, in concordance with guidelines for management of moderate-severe irAE. For algorithms Rx only and Dx only) we relaxed the requirement for the presence of both diagnosis and pharmacy data to define an irAE episode.

Table 2.

Characteristics of Validation Study Population

	Total (n = 354)	irAE (n = 89)	no irAE (n = 265)
Age (Mean, SD)	65.8 (9.8)	67.0 (10.2)	65.4 (9.7)
Age range (%)			
18–30 years	0.0	0.0	0.0
31–40 years	1.1	0.0	1.5
41–50 years	5.9	7.9	5.3
51–60 years	24.6	23.6	24.9
61–70 years	37.9	32.6	39.6
71–80 years	25.4	25.8	25.3
81+	5.1	10.1	3.4
Sex (% Male)	52.3	56.2	51.0
Body mass index (BMI) (Mean, SD)	26.6 (5.3)	26.9 (5.7)	26.5 (5.2)
Race/Ethnicity (%)			
White	78.5	82.0	77.4
Black	16.1	15.7	16.2
Asian	3.1	2.2	3.4
AAPI	0.6	0.0	0.8
Unknown/Other	1.7	0.0	2.3
Hispanic/Latino (any race)	1.1	0.0	1.5
Marital status (%)			
Married	67.0	71.9	65.3
Single	11.0	7.9	12.1
Divorced	9.3	4.5	10.9
Widowed	8.2	11.2	7.2
Other	4.5	4.5	4.5
ICI Treatment *			
Pembrolizumab	42.4	39.3	43.4
Nivolumab	41.8	53.9	37.7
Durvalumab	11.0	1.1	14.3
Atezolizumab	3.7	5.6	3.0
Ipilimumab	0.0	0.0	0.0
Other **	1.1	0.0	1.5

* Percentages may not sum to 100, as some patients received multiple ICI therapies

** Investigative therapies (relatlimab, sabatolimab, spartalizumab)

Table 3.

Algorithm performance

Test metric (95% CI)	Algorithm version				
	Dx/Rx (discontinue)	Dx / Rx	Dx / Rx (relaxed)	Rx	Dx
All toxicities (CTCAE Grade 0 [No symptoms] to Grade 5 [Fatal])					
Sensitivity	--	59.3% (46.8–71.8)	62.7% (50.4–75)	69.5% (57.8–81.2)	93.2% (86.8–99.6)
Specificity	--	77.6% (72.8–82.4)	74.6% (69.6–79.6)	72.2% (67.1–77.3)	21.0% (16.4–25.6)
PPV	--	34.7% (25.4–44)	33.0% (24.3–41.7)	33.3% (25–41.6)	19.1% (14.6–23.6)
NPV	--	90.5% (86.9–94.1)	90.9% (87.3–94.5)	92.2% (88.7–95.7)	93.9% (88.1–99.7)
C-statistic	--	0.68 (0.61–0.75)	0.69 (0.62–0.75)	0.71 (0.64–0.77)	0.57 (0.53–0.61)
Severe toxicities (CTCAE Grade 3 [Severe] to Grade 5 [Fatal])					
Sensitivity	72.4% (56.1–88.7)	72.4% (56.1–88.7)	75.9% (60.3–91.5)	86.2% (73.6–98.8)	86.2% (73.6–98.8)
Specificity	76.9% (72.3–81.5)	75.4% (70.7–80.1)	72.3% (67.4–77.2)	69.8% (64.8–74.8)	19.1% (14.8–23.4)
PPV	21.9% (13.6–30.2)	20.8% (12.9–28.7)	19.6% (12.2–27.0)	20.3% (13.2–27.4)	8.7% (5.4–12.0)
NPV	96.9% (94.8–99.0)	96.8% (94.6–99.0)	97.1% (95.0–99.2)	98.3% (96.6–100)	93.9% (88.1–99.7)
C-statistic	0.75 (0.66–0.83)	0.74 (0.65–0.83)	0.74 (0.65–0.81)	0.78 (0.71–0.85)	0.53 (0.50–0.59)
Severe common organ-specific toxicities (pneumonitis, colitis, hepatitis)					
Sensitivity	70.4% (53.2–87.6)	70.4% (53.2–87.6)	70.4% (53.2–87.6)	85.2% (71.8–98.6)	77.8% (62.1–93.5)
Specificity	90.8% (87.7–93.9)	89.9% (86.6–93.2)	85.6% (81.8–89.4)	69.4% (64.4–74.4)	66.7% (61.6–71.8)
PPV	38.8% (25.2–52.4)	36.5% (23.4–49.6)	28.8% (17.9–39.7)	18.7% (11.8–25.6)	16.2% (9.9–22.5)
NPV	97.4% (95.6–99.2)	97.4% (95.6–99.2)	97.2% (95.3–99.1)	98.3% (96.6–100)	97.3% (95.2–99.4)
C-statistic	0.81 (0.72–0.89)	0.80 (0.72–0.89)	0.78 (0.68–0.87)	0.77 (0.69–0.84)	0.72 (0.63–0.80)

Se sensitivity Sp specificity PPV positive predictive value NPV negative predictive value CI confidence interval -- Not included in main analysis