Accepted Manuscript

Submission Date: 2024-10-06
Accepted Date: 2025-01-15
Accepted Manuscript online: 2025-01-24

Applied Clinical Informatics

Application of an Externally Developed Algorithm to Identify Research Cases and Controls from Electronic Health Record Data: Failures and Successes

Nelly Estefanie Garduno Rapp, Simone D Herzberg, Henry H Ong, Cindy Kao, Christoph U Lehmann, Srushti Gangireddy, Nitin B Jain, Ayush Giri.

Affiliations below.

DOI: 10.1055/a-2524-5216

Please cite this article as: Garduno Rapp N, Herzberg S D, Ong H H et al. Application of an Externally Developed Algorithm to Identify Research Cases and Controls from Electronic Health Record Data: Failures and Successes. ACI 2025. doi: 10.1055/a-2524-5216

Conflict of Interest: The authors declare that they have no conflict of interest.

This study was supported by National Center for Advancing Translational Sciences (http://dx.doi.org/10.13039/100006108), UL1TR003163, National Institute of Arthritis and Musculoskeletal and Skin Diseases (http://dx.doi.org/10.13039/100000069), R01AR074989.

Abstract:
Background: The use of Electronic Health Records (EHRs) in research demands robust, interoperable systems. By linking biorepositories to EHR algorithms, researchers can efficiently identify cases and controls for large observational studies (e.g., Genome-Wide Association Studies (GWAS)). This is critical for ensuring efficient and cost-effective research. However, the lack of standardized metadata and algorithms across different EHRs complicates their sharing and application. Our study presents an example of a successful implementation and validation process.
Objective: To implement and validate a rule-based algorithm from a tertiary medical center in Tennessee to classify cases and controls from a research study on rotator cuff tear nested within a tertiary medical center in North Texas and to assess the algorithm‘s performance.
Methods: We applied a phenotypic algorithm (designed and validated in a tertiary medical center in Tennessee) using EHR data from 492 patients enrolled in case-control study recruited from a tertiary medical center in North Texas. The algorithm leveraged ICD (International Classification of Diseases) and CPT (Current Procedural Terminology) codes to identify case and control status for degenerative rotator cuff tears. A manual review was conducted to compare the algorithm‘s classification with a previously recorded gold standard documented by clinical researchers.
Results: Initially the algorithm identified 398 (80.9%) patients correctly as cases or controls. After fine-tunning and corrections of errors in our gold standard dataset, we calculated a sensitivity of 0.94 and specificity of 0.76.
Discussion: The implementation of the algorithm presented challenges due to the variability in coding practices between medical centers. To enhance performance, we refined the algorithm‘s data dictionary by incorporating additional codes. The process highlighted the need for meticulous code verification and standardization in multi-center studies.
Conclusion: Sharing case-control algorithms boosts EHR research. Our rule-based algorithm improved multi-site patient identification and revealed 12 data entry errors, helping validate our results.

Corresponding Author:
Dr. Nelly Estefanie Garduno Rapp, UT Southwestern, Clinical Informatics Center, 5323 Harry Hines Blvd., 75390 Dallas, United States, Nelly.GardunoRapp@utsouthwestern.edu

Affiliations:

Nelly Estefanie Garduno Rapp, UT Southwestern, Clinical Informatics Center, Dallas, United States
Simone D Herzberg, Vanderbilt University Medical Center, Nashville, United States
Henry H Ong, Vanderbilt University Medical Center, Center for Precision Medicine, Department of Biomedical Informatics., Nashville, United States
[...]
Ayush Giri, Vanderbilt University Medical Center, Divison of Quantitative and Clinical Sciences, Nashville, United States

Application of an Externally Developed Algorithm to identify research cases and controls from EHR Data: Failures and Successes

Nelly Estefanie Garduno-Rapp, MD, MSHI[1*], Simone Herzberg, PhD[2,3*], Henry H. Ong, PhD[4] Cindy Kao[1], Christoph U Lehmann, MD[1], Srushti Gangireddy[4], Nitin B Jain, MBBS, MSPH[5], Ayush Giri, PhD[2,6],

[1] Clinical Informatics Center, University of Texas Southwestern Medical Center

[2]Division of Epidemiology, Department of Medicine, Vanderbilt University Medical Center

[3]Medical Scientist Training Program, Vanderbilt University School of Medicine

[4]Center for Precision Medicine, Department of Biomedical Informatics, Vanderbilt University Medical Center

[5] Department of Physical Medicine and Rehabilitation, University of Michigan

[6]Division of Quantitative and Clinical Sciences, Department of Obstetrics and Gynecology, Vanderbilt University Medical Center

* All authors contributed equally.

Corresponding authors:

Nelly Estefanie Garduno-Rapp, MD, MSHI

Clinical Informatics Center, University of Texas Southwestern Medical Center

Ayush Giri, PhD

Division of Epidemiology, Department of Medicine, Vanderbilt University Medical Center Division of Quantitative and Clinical Sciences, Department of Obstetrics and Gynecology, Vanderbilt University Medical Center

Background

The use of Electronic Health Records (EHRs) in research demands robust, interoperable systems**.** By linking biorepositories to EHR algorithms, researchers can efficiently identify cases and controls for large observational studies (e.g., Genome-Wide Association Studies (GWAS)).

This is critical for ensuring efficient and cost-effective research. However, the lack of standardized metadata and algorithms across different EHRs complicates their sharing and application. Our study presents an example of a successful implementation and validation process.

Objective

To implement and validate a rule-based algorithm from a tertiary medical center in Tennessee to classify cases and controls from a research study on rotator cuff tear nested within a tertiary medical center in North Texas and to assess the algorithm's performance.

Methods

We applied a phenotypic algorithm (designed and validated in a tertiary medical center in Tennessee) using EHR data from 492 patients enrolled in case-control study recruited from a tertiary medical center in North Texas. The algorithm leveraged ICD (International Classification of Diseases) and CPT (Current Procedural Terminology) codes to identify case and control status for degenerative rotator cuff tears. A manual review was conducted to compare the algorithm's classification with a previously recorded gold standard documented by clinical researchers.

Results

Initially the algorithm identified 398 (80.9%) patients correctly as cases or controls. After fine-tunning and corrections of errors in our gold standard dataset, we calculated a sensitivity of 0.94 and specificity of 0.76.

Discussion

The implementation of the algorithm presented challenges due to the variability in coding practices between medical centers. To enhance performance, we refined the algorithm's data dictionary by incorporating additional codes. The process highlighted the need for meticulous code verification and standardization in multi-center studies.

Conclusion

Sharing case-control algorithms boosts EHR research. Our rule-based algorithm improved multi-site patient identification and revealed 12 data entry errors, helping validate our results.

Keywords: Phenotypic algorithms, Data Validation, Clinical Research Informatics

Background

As the use of Electronic Health Records (EHRs) for large-scale research is increasing[1], there is a pressing need to develop robust infrastructures and innovative research tools to provide syntactic and semantic interoperability among health systems and organizations[2,3]. To achieve this concept, researchers must overcome the lack of harmonization of national and institution-specific terminologies, formats, and structures into standardized formats such as the OMOP (Observational Medical Outcomes Partnership) CDM (common data model)[2,4-6]. Such advancements could transform EHRs into powerful research tools and ultimately contribute to improved patient outcomes. A critical aspect of this transformation involves the development of harmonized models, techniques, tools, and algorithms that can be applied to large datasets across multiple health systems[5,7,8]. One prominent type of research that leverages large-scale datasets and often involves data collected from multiple sites are Genome-Wide Association Studies (GWAS)[9], which are increasingly prevalent and identify genetic variants that predispose individuals to complex disorders (association between genotype and phenotype)[10]. These studies hold great promise for advancing our understanding and treatment of various diseases such as degenerative rotator cuff tear (DCT), with the caveat that data from EHRs, originally collected for patient care rather than research, are curated in a principled manner.[11,12]

A fundamental component of the success of population studies, including GWAS, is the correct classification of cases and controls[13,14]. While various cohort discovery tools, such as i2b2 (Informatics for Integrating Biology at the Bedside), TriNetX, and OHDSI/ATLAS (Observational Health Data Sciences and Informatics), quickly facilitate the identification of potential research participants, these tools are most effective for direct, single-step queries[15,16]. These platforms have fixed structures for how the data are stored and organized, which could limit the flexibility in how data are queried or analyzed. Thus, they fall short when handling complex clinical scenarios and meeting specific criteria that require multi-step temporal logic to answer research questions[17].

Our study addresses this gap by implementing and validating an external rule-based algorithm, leveraging CPT and ICD coding. Algorithms based on CPT and ICD codes offer a more effective approach, due to their flexibility to tailor data and rules to classify cases and controls in a more precise way. This allows for more accurate categorization in complex scenarios, overcoming the limitations of traditional cohort discovery tools.[18-20]

Nonetheless, research has shown that structured algorithms must be clear and well defined to avoid poor interpretation. For instance, asking for “patients that are 40 years of age or older” does not indicate at what point in the disease course the patient should be at least 40 [21,22].

The algorithm used in this study was developed using a unique combination of CPT and ICD codes and it involved consideration of frequency and temporality associated with other codes. It was designed and internally validated at Vanderbilt University Medical Center (VUMC) from a de-identified clinical records database. The database supports queries of structured clinical information such as diagnostic codes, CPT codes, medications, laboratory data, allergies, and demographics and unstructured clinical information including medical reports, radiology notes, and surgical notes. More details of the VUMC algorithm are described elsewhere[23].

Briefly, UTSW and VUMC are both tertiary medical centers with diverse populations in the southern United States. This makes our study particularly valuable by demonstrating the algorithm's performance across different EHR instances.

In this paper, we provide a comprehensive account of the algorithm's implementation and validation processes. We demonstrate how applying this external algorithm contributed to greater consistency and reliability in our case and control classifications within the gold-standard dataset.

Hypothesis

We hypothesized that the algorithm developed at VUMC would initially underperform and miss cases and controls from our gold standard dataset at UTSW, and that targeted improvements could enhance its performance and usability across other tertiary medical centers.

Objective

To implement and validate a rule-based algorithm designed at VUMC to classify rotator cuff tear cases and controls in a tertiary care medical center at UTSW and to evaluate the algorithm performance.

Methods

Study Population

Patients older than 40 years of age with a shoulder Magnetic Resonance Image (MRI) met the eligibility criteria for enrollment in an actively recruiting observational, case-control study for a GWAS at UTSW, which served as the gold-standard case-control classifications. Cases in this study were determined based on the presence of a shoulder MRI with evidence of an atraumatic rotator cuff tear (RCT) as documented in the patient's medical chart. Controls were patients with a shoulder MRI indicating a condition other than RCT, such as adhesive capsulitis, osteoarthritis, or shoulder instability. Trained research personnel recorded patient information and classification as case or as control in a web-based data collection tool (REDCap) as the gold standard for this study[24].

Processing the Gold Standard Dataset

1. Initially, we downloaded a de-identified dataset from REDCap, which included the current case or control classifications for 492 participants (405 cases and 87 controls) who were enrolled from 2021-2023. This dataset was maintained as our gold standard for subsequent analysis. Although this dataset lacked personal identifiers, each entry was associated with a unique, study-specific identifier that allowed us to align records accurately across datasets.

Applying the algorithm developed at VUMC to the UTSW EHR databases

2. Next, we applied the VUMC algorithm to all 492 participants in our epic databases, specifically: Caboodle and Clarity. The algorithm employed specific combinations of 18 CPT codes, 13 ICD-9-CM codes, and 39 ICD-10-CM codes. This ensured precise identification of participants with rotator cuff tears while distinguishing them from those with other shoulder conditions, such as adhesive capsulitis, glenohumeral osteoarthritis, or scapular dyskinesis.

   Additionally, the algorithm had frequency and temporality requirements: 1) To ensure accuracy, the codes needed to be mentioned more than once at separate time points in the medical record, and 2) codes had to satisfy temporal relationship requirements with other codes. For example, to become a case, a patient had to have a CPT code for a shoulder MRI followed by an ICD code for rotator cuff tear diagnosis within one year after the CPT code. Table 1a and 1b display the full algorithm criteria. Table 3 displays our full data dictionary.

Data Comparison and Verification Process

3. We utilized R (An open-source programming language for data analysis) to compare the algorithm's output classifications (cases or controls) with those in the gold-standard dataset, focusing on identifying discrepancies such as false positives, false negatives, and missing cases between the two sets. To assess the source of these differences, we performed a thorough manual review of each participant’s medical chart. This was an essential step to understand how to address the discrepancies and improve the algorithm. Lastly, we calculated the algorithm’s sensitivity, specificity, and accuracy. Figure 1 shows a visual representation of our methodology.

Results

Initially, the algorithm identified 398 (80.9%) patients correctly as cases or controls (371 true positive (TP) cases and, 27 true negative (TN) controls). There were 60 false positives (FP), and 34 false negatives (FN). We examined the 94 discrepancies (60 FP and 34 FN) between the algorithm's outcomes and the existing case-control determinations based on the GWAS study in REDCap (Figure 2). Through a manual review of the medical records, including image impressions, procedures, and clinical notes, we discovered that only 11 of the 60 FP cases (18.3%) were truly false positives. The remaining 49 records (81.7%) were mislabeled in our gold standard database in REDCap. Of these 49 records, 42 (85.7%) had conflicting diagnoses recorded with radiologists identifying a rotator cuff tear (RCT) based on imaging, while treating physicians labeled these cases as tendinitis or dyskinesis. Additionally, in six cases research staff made data entry errors. A single patient had two diagnoses, including RCT and glenohumeral osteoarthritis (GHOA). For the 34 FN cases, we found that only 26 (76.5%) were true misclassifications by the algorithm. The remaining eight records were mislabeled in our gold standard in REDCap, with six being data entry errors and two having conflicting diagnoses where radiologists did not diagnose RCT, but the treating physicians did. Figure 3 illustrates all discrepancies with the gold standard identified for the false positive and negative cases. Specifically, it shows 44 cases with conflicting diagnoses, 12 data entry errors, and one case with a dual diagnosis.

After this thorough review, we reclassified the records and determined that the algorithm produced 420 TP, 26 FN, 11 FP, and 35 TN. Lastly, metrics were recalculated, resulting in a sensitivity of 94%, specificity of 76%, and accuracy of 92 %. Ultimately, the true number of discrepancies was 37 (11 FP and 26 FN). Table 2 shows a matrix with our results adjusted for errors in our gold standard.

Discussion

We implemented an external algorithm that classified cases and controls for an atraumatic rotator cuff tear study in our EHR and faced several challenges: 1) the initial extraction process failed to identify 33 patients out of the 492 participants due to differences in usage of CPT codes between the organization where the algorithm was originally developed (VUMC) and the organization where the algorithm was applied (UTSW). For example, the procedure for the "repair of ruptured musculotendinous cuff," which was coded as 23412 in one EHR system and 23410 in the other. These differences extended beyond individual procedures. We observed that some ICD and CPT codes were not included in our initial data dictionary because they were represented by different codes in other institutions. Additionally, we identified the need to account for patients whose imaging studies were performed externally and thus required the inclusion of specific CPT codes associated with these external images. To address these discrepancies, we expanded the algorithm's data dictionary to include additional local CPT and ICD-9 codes that were unique to UTSW Medical Center. Figure 4 shows the percentage of additional codes that were unique to UTSW Medical Center (11%), the percentage of codes that were unique to VUMC Medical center (14%) and the percentage of codes that were shared between institutions (75%). Note that Table 3 displays all shared codes between both organizations.

Following our modifications, the algorithm successfully identified most cases and controls, demonstrating the effectiveness of the updated data dictionary and coding practices in harmonizing patient records across different institutional EHR systems. While this reconciliation process was labor-intensive, it provided significant insights into the variability of coding practices between different EHR systems. For example, the identification of locally defined codes as well a small percentage of procedures coded differently across EHRs highlights the importance of meticulous code verification and standardization in multicenter studies to ensure data integrity and comparability. Metadata sharing prior to data collection for such multicenter studies could emphasize potential coding discrepancies and decrease time consuming tasks such as manual EHR review.

Additionally, we found 94 discrepancies between the algorithm's outcomes and the existing classifications in our gold standard, which prompted us to perform a thorough manual review of these records during which we found a significant number of mislabeled patients in our gold standard database reducing the true number of discrepancies to only 37 (11 FP and 26 FN). The implementation of the VUMC algorithm allowed us to improve the quality of our gold standard enhancing the accuracy and reliability of patient identification and classification in our institution.

An important aspect to consider is the very definition of the "gold standard" against which algorithms and clinical judgments are compared. The observed discrepancies in our findings largely stem from differences in provider interpretations, particularly between radiologists and other specialists such as orthopedic surgeons and physiatrists. This raises critical questions about the role of disciplinary perspectives in clinical decision-making. Notably, the algorithm appears to align most closely with radiologists' determinations, likely because it is designed around radiology report impressions. This observation highlights the nuanced nature of algorithmic performance, which may be influenced by the specific clinical lens through which evidence is interpreted.

We anticipate that the implementation of the modified algorithm in other performance research sites would likely show further coding discrepancies, but the return would likely be diminished for each additional institution resulting in an algorithm that could be applied in other tertiary medical centers using ICD and CPT codes.

Ensuring data consistency and integrity is paramount for producing valid and reproducible research outcomes[25]. By addressing the diverging coding practices and harmonizing them, we improved the robustness of our dataset, which is essential for drawing meaningful conclusions in clinical studies. Moreover, this implementation highlighted the need for standardized coding systems and meticulous data verification processes, ultimately contributing to the advancement of data interoperability and quality in multicenter research.

Limitations

One limitation of this study is the inherent variability in coding practices across different medical centers, which impacted the initial performance of the VUMC algorithm when applied to our patient population. Another limitation is that the algorithm was only tested at a single institution, which limits the generalizability of the findings. Testing the algorithm in different organizations could reveal additional coding discrepancies and further affect its performance. This emphasizes the importance of validating such algorithms across diverse settings to ensure their robustness and adaptability in multicenter research studies.

Conclusions

Implementing and validating the VUMC algorithm at UTSW, an institution with its own patient population and health system, suggests that this tool can perform reliably outside its original development environment. While coding discrepancies need to be addressed, we showed that a rule-based algorithm could be a potential alternative to better identify and validate multi-site patient cohorts. Additionally, the algorithm allowed us to pinpoint 12 data entry errors in our gold standard and gave us an opportunity to validate our classifications.

Clinical Relevance

The study highlights the critical importance of harmonizing CPT and ICD codes across institutions to ensure accurate patient classification in multicenter studies. Practitioners should be aware that algorithm performance may vary depending on coding practices and the clinical interpretation lens. Addressing coding discrepancies improves data quality, ultimately enhancing the reliability of research outcomes and patient care.

Acknowledgments

Research reported in this publication was supported by the National Center for Advancing Translational Sciences of the National Institute of Health under award Number UL1TR003163. As well as the National Institute of Arthritis and Musculoskeletal and Skin Diseases of the National Institute of Health under Award Number R01AR074989. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

In the spirit of fostering scientific collaboration, the SQL query and code used in this study will be made publicly available upon publication of this and related articles in review. Researchers interested in accessing the code for further analysis or replication of the study findings are encouraged to visit https://github.com/Estefanie-Rapp/Rule_based_algorithm-

Conflict of Interest Statement:

The authors declare that there are no conflicts of interest in the research. No financial, personal, or professional affiliations have influenced the study, data interpretation, or content presented in this work.

Human Subjects Protection and Ethical Considerations

Our study received approval from the Institutional Review Board center STU-2020-0689. Only patients who provided informed consent at UTSW were included in the data query. To ensure confidentiality, all patient information was de-identified and securely managed.

Multiple Choice Questions

1. Which challenges were faced during the algorithm implementation for the rotator cuff tear study?

A) Lack of patient consent

B) Differences in CPT code usage across organizations

C) Insufficient sample size

D) Limited imaging availability

Correct Answer: B) Differences in CPT code usage across organizations

2. What was identified as a necessary modification to improve the algorithm's performance?

A) Reducing the patient sample size

B) Changing the software used for data analysis

C) Increasing the number of healthcare providers involved

D) Expanding the algorithm’s data dictionary to include additional CPT and ICD codes

Correct Answer: D) Expanding the algorithm’s data dictionary to include additional CPT and ICD codes

3. What criteria were used to classify patients as cases in the study?

A) Patients older than 50 years with shoulder pain

B) Patients with a shoulder MRI indicating adhesive capsulitis

C) Patients with a shoulder MRI showing evidence of an atraumatic rotator cuff tear (RCT)

D) Patients with any shoulder-related condition documented in their medical chart

Correct Answer: C) Patients with a shoulder MRI showing evidence of an atraumatic rotator cuff tear (RCT)

References

1. Adler-Milstein J, Holmgren AJ, Kralovec P, Worzala C, Searcy T, Patel V. Electronic health record adoption in US hospitals: the emergence of a digital “advanced use” divide. Journal of the American Medical Informatics Association. 2017;24(6):1142-1148. doi:10.1093/jamia/ocx080
2. Henke E, Zoch M, Peng Y, Reinecke I, Sedlmayr M, Bathelt F. Conceptual design of a generic data harmonization process for OMOP common data model. BMC Med Inform Decis Mak. Feb 26 2024;24(1):58. doi:10.1186/s12911-024-02458-7
3. Kiourtis A, Nifakos S, Mavrogiorgou A, Kyriazis D. Aggregating the syntactic and semantic similarity of healthcare data towards their transformation to HL7 FHIR through ontology matching. Int J Med Inform. Dec 2019;132:104002. doi:10.1016/j.ijmedinf.2019.104002
4. Garza M, Del Fiol G, Tenenbaum J, Walden A, Zozus MN. Evaluating common data models for use with a longitudinal community registry. J Biomed Inform. Dec 2016;64:333-341. doi:10.1016/j.jbi.2016.10.016
5. Kumar G, Basri S, Imam AA, Khowaja SA, Capretz LF, Balogun AO. Data Harmonization for Heterogeneous Datasets: A Systematic Literature Review. Applied Sciences. 2021;11(17):8275.

6. Sedlakova J, Daniore P, Horn Wintsch A, et al. Challenges and best practices for digital unstructured data enrichment in health research: A systematic narrative review. PLOS Digit Health. Oct 2023;2(10):e0000347. doi:10.1371/journal.pdig.0000347
7. Peng Y, Henke E, Reinecke I, Zoch M, Sedlmayr M, Bathelt F. An ETL-process design for data harmonization to participate in international research with German real-world data based on FHIR and OMOP CDM. Int J Med Inform. Jan 2023;169:104925. doi:10.1016/j.ijmedinf.2022.104925
8. Rosenbloom ST, Carroll RJ, Warner JL, Matheny ME, Denny JC. Representing Knowledge Consistently Across Health Systems. Yearb Med Inform. Aug 2017;26(1):139-147. doi:10.15265/iy-2017-018
9. Bick AG, Metcalf GA, Mayo KR, et al. Genomic data in the All of Us Research Program. Nature. 2024/03/01 2024;627(8003):340-346. doi:10.1038/s41586-023-06957-x
10. Durbin RM, Altshuler D, Durbin RM, et al. A map of human genome variation from population-scale sequencing. Nature. 2010/10/01 2010;467(7319):1061-1073. doi:10.1038/nature09534
11. Marees AT, de Kluiver H, Stringer S, et al. A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. International Journal of Methods in Psychiatric Research. 2018;27(2):e1608. doi:https://doi.org/10.1002/mpr.1608
12. Tashjian RZ, Kim SK, Roche MD, Jones KB, Teerlink CC. Genetic variants associated with rotator cuff tearing utilizing multiple population-based genetic resources. Journal of Shoulder and Elbow Surgery. 2021;30(3):520-531. doi:10.1016/j.jse.2020.06.036
13. Castro VM, Apperson WK, Gainer VS, et al. Evaluation of matched control algorithms in EHR-based phenotyping studies: A case study of inflammatory bowel disease comorbidities. Journal of Biomedical Informatics. 2014/12/01/ 2014;52:105-111. doi:https://doi.org/10.1016/j.jbi.2014.08.012
14. Thomas SV, Suresh K, Suresh G. Design and data analysis case-controlled study in clinical research. Annals of Indian Academy of Neurology. 2013;16(4):483-487. doi:10.4103/0972-2327.120429
15. Bucalo M, Gabetta M, Chiudinelli L, et al. i2b2 to Optimize Patients Enrollment. Stud Health Technol Inform. May 27 2021;281:506-507. doi:10.3233/shti210217
16. Prebay ZJ, Ostrovsky AM, Buck M, Chung PH. A TriNetX Registry Analysis of the Need for Second Procedures following Index Anterior and Posterior Urethroplasty. J Clin Med. Mar 5 2023;12(5)doi:10.3390/jcm12052055
17. Chamberlin SR, Bedrick SD, Cohen AM, et al. Evaluation of patient-level retrieval from electronic health record data for a cohort discovery task. JAMIA Open. Oct 2020;3(3):395-404. doi:10.1093/jamiaopen/ooaa026
18. Mallik S, Zhao Z. Graph- and rule-based learning algorithms: a comprehensive review of their applications for cancer type classification and prognosis using genomic data. Brief Bioinform. Mar 23 2020;21(2):368-394. doi:10.1093/bib/bby120
19. Teixeira PL, Wei WQ, Cronin RM, et al. Evaluating electronic health record data sources and algorithmic approaches to identify hypertensive individuals. J Am Med Inform Assoc. Jan 2017;24(1):162-171. doi:10.1093/jamia/ocw071
20. Ganz DA, Esserman D, Latham NK, et al. Validation of a Rule-Based ICD-10-CM Algorithm to Detect Fall Injuries in Medicare Data. The Journals of Gerontology: Series A. 2024;79(7)doi:10.1093/gerona/glae096
21. Yu J, Pacheco JA, Ghosh AS, et al. Under-specification as the source of ambiguity and vagueness in narrative phenotype algorithm definitions. BMC Med Inform Decis Mak. Jan 28 2022;22(1):23. doi:10.1186/s12911-022-01759-z
22. Hruby GW, Boland MR, Cimino JJ, et al. Characterization of the biomedical query mediation process. AMIA Jt Summits Transl Sci Proc. 2013;2013:89-93.

23. Herzberg S, Garduno-Rapp NE, Ong H, et al. Standardizing Phenotypic Algorithms for the Classification of Degenerative Rotator Cuff Tear from Electronic Health Record Systems. medRxiv. 2024:2024.09.29.24314565. doi:10.1101/2024.09.29.24314565
24. Harris PA, Taylor R, Minor BL, et al. The REDCap consortium: Building an international community of software platform partners. Journal of Biomedical Informatics. 2019/07/01/ 2019;95:103208. doi:https://doi.org/10.1016/j.jbi.2019.103208
25. Shewade HD, Vidhubala E, Subramani DP, et al. Open access tools for quality-assured and efficient data entry in a large, state-wide tobacco survey in India. Global Health Action. 2017/01/01 2017;10(1):1394763. doi:10.1080/16549716.2017.1394763

Table1a. Algorithm Criteria for cases

Table 1a			
	Case Definition	**Description**	**Criteria and Boolean Logic**
1	Specific Surgical Inclusion	The diagnosis (Dx) date is determined by the earliest date associated with a specific surgical procedure.	a) rct_cpt_surg_spec_include
2	Non-Specific Surgical/ICD Inclusion	The Dx date is the earliest date associated with non-specific surgical procedures or ICD codes, with additional criteria for diagnosis within a year.	**a)** rct_cpt_surg_nonspec_include **b) AND (** rct_icd9_diag_include **OR** rct_icd10_diag_include within 1 year after**)**
3	Imaging and Diagnosis	The Dx date is based on imaging CPT/ICD codes, with diagnosis codes within a year and exclusion criteria applied afterward.	**a) (**rct_cpt_image_include **OR** RCT_icd9_image_include**)** b) **AND (**rct_icd9_diag_include **OR** rct_icd10_diag_include within 1 year after**)** **c) NOT (**rct_icd9_exclusions OR rct_icd10_exclusions after CPT/ICD image include codes**)**
4 A	Multiple ICD Inclusions (3 Visits)	The Dx date is determined by the 3rd unique ICD code, ensuring there are 3 visits with the relevant diagnosis without exclusion codes.	a) >=3 unique visits with mentions of rct_icd9_diag_include **OR** rct_icd10_diag_include b) **NOT** rct_icd9_exclusions OR rct_icd10_exclusions (After ICD inclusion codes)
4 B	Multiple ICD Inclusions (4 V isits)	The Dx date is set by the 4th unique ICD code, ensuring there are at least 4 mentions of the relevant diagnosis without exclusion codes.	a) >=4 mentions of rct_icd9_diag_include **OR** >=4 mentions of rct_icd10_diag_include b) **NOT** rct_icd9_exclusions **OR** rct_icd10_exclusions (After ICD diag include)

Table 1b. Algorithm Criteria for Controls

Table 1b			
	Control Definition	**Description**	**Criteria and Boolean Logic**
1	Any non-case	Any non-case patients from CPT/ICD list.	A) **NOT** case status FROM CPT/ICD codes
2	CPT/ICD only with Imaging code confirmation for in-tact rotator cuff	All patients with CPT codes for imaging. All patients with ICD-9 codes for imaging. All patients with ICD-10 codes for imaging and exclusion criteria applied afterward. If the patient meets criteria for being a case, they are excluded from the control group.	A) (rct_cpt_image_include **OR** RCT_icd9_image_include **OR** RCT_icd10_image_include) B) **NOT case** status from CPT/ICD codes C) **NOT** **(**rct_cpt_surg_spec_include **OR** rct_cpt_surg_nonspec_include **OR** RCT_icd9_diag_include **OR** RCT_icd10_diag_include**)**

Table 2. Performance Metrics

Table 2	**Actual Cases**	**Actual Controls**	**Performance Metric**
Labeled as Case	420	11	Sensitivity 94%
Labeled as Control	26	35	Specificity 76%
			Accuracy 92%

Table 3. Data Dictionary List

Data Dictionary List						
VARIABLE NAME	**TYPE**	**CODE**	**CODE NAME/ DESCRIPTION**	**UTSW only**	**VUMC only**	**Shared in both**
rct_cpt_surg_spec_include	ICD9CM	83.63	Rotator cuff repair	No	No	Yes
RCT_icd9_diag_include	ICD9CM	727.61	Complete rupture of rotator cuff	No	No	Yes
RCT_icd9_diag_include	ICD9CM	726.13	Partial tear of rotator cuff	No	No	Yes
RCT_icd9_diag_include	ICD9CM	83.63	Rotator cuff repair	No	No	Yes
RCT_icd9_nontraum	ICD9CM	727.6	Rupture of tendon nontraumatic	No	No	Yes
RCT_icd9_nontraum	ICD9CM	727.6	Nontraumatic rupture of unspecified tendon	No	No	Yes
RCT_icd9_exclusions	ICD9CM	840.3	Infraspinatus (muscle) (tendon) sprain	No	No	Yes
RCT_icd9_exclusions	ICD9CM	840.4	Rotator cuff (capsule) sprain	No	No	Yes
RCT_icd9_exclusions	ICD9CM	840.5	Subscapularis (muscle) sprain	No	No	Yes
RCT_icd9_exclusions	ICD9CM	840.6	Supraspinatus (muscle) (tendon) sprain	No	No	Yes
RCT_icd9_image_include	ICD9CM	88.94	Magnetic Resonance Imaging of Musculoskeletal	No	No	Yes
RCT_icd9_image_include	ICD9CM	88.32	Contrast arthrogram	No	No	Yes
RCT_icd9_image_include	ICD9CM	88.7	Diagnostic Ultrasound	No	No	Yes
RCT_icd10_diag_include	ICD10CM	M75.120	Complete rotator cuff tear or rupture of unspecified shoulder, not specified as traumatic	No	No	Yes
RCT_icd10_diag_include	ICD10CM	M75.121	Complete rotator cuff tear or rupture of right shoulder, not specified as traumatic	No	No	Yes

RCT_icd10_diag_include	ICD10CM	M75.122	Complete rotator cuff tear or rupture of left shoulder, not specified as traumatic	No	No	Yes
RCT_icd10_diag_include	ICD10CM	M75.110	Incomplete rotator cuff tear or rupture of unpsecified shoulder, not specified as traumatic	No	No	Yes
RCT_icd10_diag_include	ICD10CM	M75.111	Incomplete rotator cuff tear or rupture of right shoulder, not specified as traumatic	No	No	Yes
RCT_icd10_diag_include	ICD10CM	M75.112	Incomplete rotator cuff tear or rupture of left shoulder, not specified as traumatic	No	No	Yes
RCT_icd10_diag_include	ICD10CM	M75.100	Unspecified rotator cuff tear or rupture of unspecified shoulder, not specified as traumatic	No	No	Yes
RCT_icd10_diag_include	ICD10CM	M75.101	Unspecified rotator cuff tear or rupture of right shoulder, not specified as traumatic	No	No	Yes
RCT_icd10_diag_include	ICD10CM	M75.102	Unspecified rotator cuff tear or rupture of left shoulder, not specified as traumatic	No	No	Yes
rct_icd10_exclude	ICD10CM	S46.011A	Strain of muscle(s) and tendon(s) of the rotator cuff of right shoulder, initial encounter	No	No	Yes
rct_icd10_exclude	ICD10CM	S46.011D	Strain of muscle(s) and tendon(s) of the rotator cuff of right shoulder, subsequent encounter	No	No	Yes
rct_icd10_exclude	ICD10CM	S46.011S	Strain of muscle(s) and tendon(s) of the rotator cuff of right shoulder, sequela	No	No	Yes
rct_icd10_exclude	ICD10CM	S46.012A	Strain of muscle(s) and tendon(s) of the rotator cuff of left shoulder, initial encounter	No	No	Yes
rct_icd10_exclude	ICD10CM	S46.012D	Strain of muscle(s) and tendon(s) of the rotator cuff of left shoulder, sequential encounter	No	No	Yes

rct_icd10_exclude	ICD10CM	S46.012S	Strain of muscle(s) and tendon(s) of the rotator cuff of left shoulder, sequela	No	No	Yes
rct_icd10_exclude	ICD10CM	S46.011A	Strain of muscle(s) and tendon(s) of the rotator cuff of unspecified shoulder, initial encounter	No	No	Yes
rct_icd10_exclude	ICD10CM	S46.011D	Strain of muscle(s) and tendon(s) of the rotator cuff of unspecified shoulder, subsequent encounter	No	No	Yes
rct_icd10_exclude	ICD10CM	S46.011S	Strain of muscle(s) and tendon(s) of the rotator cuff of unspecified shoulder, sequela	No	No	Yes
rct_icd10_exclude	ICD10CM	S46.021A	Laceration of muscle(s) and tendon(s) of the rotator cuff of the right shoulder, initial encounter	No	No	Yes
rct_icd10_exclude	ICD10CM	S46.021D	Laceration of muscle(s) and tendon(s) of the rotator cuff of the right shoulder, Sequential encounter	No	No	Yes
rct_icd10_exclude	ICD10CM	S46.021S	Laceration of muscle(s) and tendon(s) of the rotator cuff of the right shoulder, sequela	No	No	Yes
rct_icd10_exclude	ICD10CM	S46.022A	Laceration of muscle(s) and tendon(s) of the rotator cuff of the left shoulder, initial encounter	No	No	Yes
rct_icd10_exclude	ICD10CM	S46.022D	Laceration of muscle(s) and tendon(s) of the rotator cuff of the left shoulder, Sequential encounter	No	No	Yes
rct_icd10_exclude	ICD10CM	S46.022S	Laceration of muscle(s) and tendon(s) of the rotator cuff of the left shoulder, sequela	No	No	Yes
rct_icd10_exclude	ICD10CM	S46.029A	Laceration of muscle(s) and tendon(s) of the rotator cuff of unspecified shoulder,	No	No	Yes

			initial encounter			
rct_icd10_exclude	ICD10CM	S46.029D	Laceration of muscle(s) and tendon(s) of the rotator cuff of unspecified shoulder, Sequential encounter	No	No	Yes
rct_icd10_exclude	ICD10CM	S46.029S	Laceration of muscle(s) and tendon(s) of the rotator cuff of unspecified shoulder, sequela	No	No	Yes
rct_icd10_exclude	ICD10CM	S43.421A	Sprain of muscle(s) and tendon(s) of the rotator cuff of right shoulder, initial encounter	No	No	Yes
rct_icd10_exclude	ICD10CM	S43.421D	Sprain of muscle(s) and tendon(s) of the rotator cuff of right shoulder, sequential encounter	No	No	Yes
rct_icd10_exclude	ICD10CM	S43.421S	Sprain of muscle(s) and tendon(s) of the rotator cuff of right shoulder, sequela	No	No	Yes
rct_icd10_exclude	ICD10CM	S43.422A	Sprain of muscle(s) and tendon(s) of the rotator cuff of left shoulder, initial encounter	No	No	Yes
rct_icd10_exclude	ICD10CM	S43.422D	Sprain of muscle(s) and tendon(s) of the rotator cuff of left shoulder, sequential encounter	No	No	Yes
rct_icd10_exclude	ICD10CM	S43.422S	Sprain of muscle(s) and tendon(s) of the rotator cuff of left shoulder, sequela	No	No	Yes
rct_icd10_exclude	ICD10CM	S43.429A	Sprain of muscle(s) and tendon(s) of the rotator cuff of unspecified shoulder, initial encounter	No	No	Yes
rct_icd10_exclude	ICD10CM	S43.429D	Sprain of muscle(s) and tendon(s) of the rotator cuff of unspecified shoulder, sequential encounter	No	No	Yes
rct_icd10_exclude	ICD10CM	S43.429S	Sprain of muscle(s) and tendon(s) of the rotator cuff of unspecified shoulder, sequela	No	No	Yes

rct_icd10_exclude	ICD10CM	M12.511	Traumatic Arthropathy, right shoulder	No	No	Yes
rct_icd10_exclude	ICD10CM	M12.512	Traumatic Arthropathy, left shoulder	No	No	Yes
rct_icd10_exclude	ICD10CM	M12.519	Traumatic Arthropathy, unspecified shoulder	No	No	Yes
rct_cpt_surg_spec_include	CPT	23412	Repair of ruptured musculotendinous cuff (eg. Rotator cuff) open; chronic	No	No	Yes
rct_cpt_surg_spec_include	CPT	23420	Reconstruction of complete shoulder (rotator) cuff avulsion; chronic	No	No	Yes
rct_cpt_surg_spec_include	CPT	29827	Arthroscopy, shoulder, surgical; with rotator cuff repair	No	No	Yes
rct_cpt_surg_nonspec_include	CPT	80.21	Arthroscopy, shoulder	Yes	No	No
rct_cpt_surg_nonspec_include	CPT	29826	Arthroscopy, shoulder, surgical; decompression of subacromial space with partial acromioplasty, with coracoacromial ligament (ie, arch) release, when performed	No	No	Yes
rct_cpt_surg_nonspec_include	CPT	29805	Arthroscopy, shoulder, diagnostic, with or without synovial biopsy (separate procedure)	Yes	No	No
rct_cpt_surg_nonspec_include	CPT	29822	Arthroscopy, shoulder, surgical; debridement, limited	No	No	Yes
rct_cpt_surg_nonspec_include	CPT	29823	Arthroscopy, shoulder, surgical; debridement, extensive	No	No	Yes
rct_cpt_surg_nonspec_include	CPT	01610	Anesthesia for all procedures on nerves, muscles, tendons, fascia, and bursae of shoulder and axilla	Yes	No	No
rct_cpt_surg_nonspec_include	CPT	01622	Anesthesia for diagnostic arthroscopic procedures of shoulder joint	Yes	No	No
rct_cpt_surg_	CPT	01630	Anesthesia for open or	Yes	No	No

nonspec_include			surgical arthroscopic procedures on humeral head and neck, sternoclavicular joint, acromioclavicular joint, and shoulder joint; not otherwise specified			
rct_cpt_surg_nonspec_include	CPT	01638	Anesthesia for open or surgical arthroscopic procedures on humeral head and neck, sternoclavicular joint, acromioclavicular joint, and shoulder joint; total shoulder replacement	Yes	No	No
rct_cpt_surg_nonspec_include	CPT	01710	Anesthesia for procedures on nerves, muscles, tendons, fascia, and bursae of upper arm and elbow; not otherwise specified	Yes	No	No
rct_cpt_image_include	CPT	23350	Injection procedure for shoulder arthrography or enhanced CT/MRI Shoulder arthrography	No	No	Yes
rct_cpt_image_include	CPT	73221	MRI Shoulder, Elbow, Wrist or Clavicle w/o contrast	No	No	Yes
rct_cpt_image_include	CPT	73223	MRI Shoulder, Elbow, Wrist or Clavicle w/wo Contrast	No	No	Yes
rct_cpt_image_include	CPT	73218	MRI Upper Extremity w/o contrast	No	No	Yes
rct_cpt_image_include	CPT	73220	MRI Upper Extremity w/ contrast involvement	No	No	Yes
rct_cpt_image_include	CPT	76140	CT/MR/MRA outside study	Yes	No	No
rct_cpt_surg_nonspec_include	CPT	23410	Repair of ruptured musculotendinous cuff (eg. Rotator cuff) open; chronic	No	Yes	No
rct_cpt_surg_nonspec_include	CPT	23397	Under Repair, Revision, and/or Reconstruction Procedures on the Shoulder	No	Yes	No
rct_cpt_surg_nonspec_include	CPT	29901	Under Endoscopy/Arthroscopy Procedures on the Musculoskeletal System	No	Yes	No

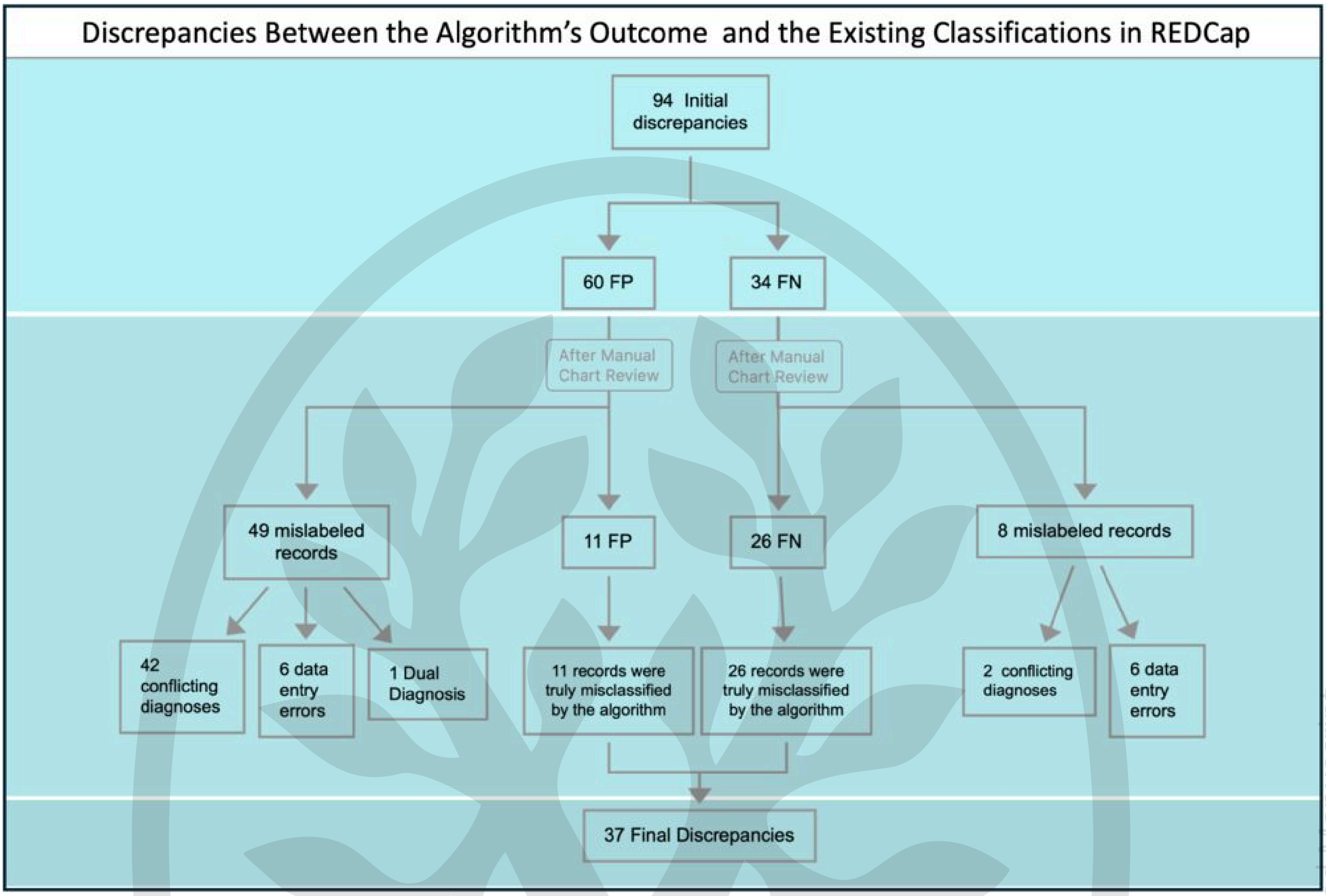
rct_cpt_surg_exclude	CPT	24341	repair, tendon or muscle, upper arm or elbow, each tendon or muscle, primary or secondary (excludes Rotator cuff)	No	Yes	No
rct_cpt_image_include	CPT	0055T	computer assisted musculoskeletal surgical navigational orthopedic procedure, with image guidance based on CT/MRI images (list separately in addition to code for primary procedure)	No	Yes	No
RCT_icd10_image_include	ICD10CM	BP3FYZZ	MRI upper extremity left, with contrast	No	Yes	No
RCT_icd10_image_include	ICD10CM	BP3EZZZ	MRI upper extremity left	No	Yes	No
rct_cpt_image_include	CPT	76880	[Expired] Ultrasound\, extremity\, nonvascular\, real time with image documentation	No	Yes	No
rct_cpt_image_include	CPT	78662	Ultrasound\, limited\, joint or other non-vascular extremity structure (i.e. joint space\, peri-articular tendon[s]\, muscle[s]\, nerve[s]\, other soft tissue structure[s]\, or soft tissue mass[es]) real time with image documentation	No	Yes	No
rct_cpt_image_include	CPT	78661	Ultrasound\, complete joint (ie. Joint space and peri-articular soft tissue structures) real time with image documentation	No	Yes	No

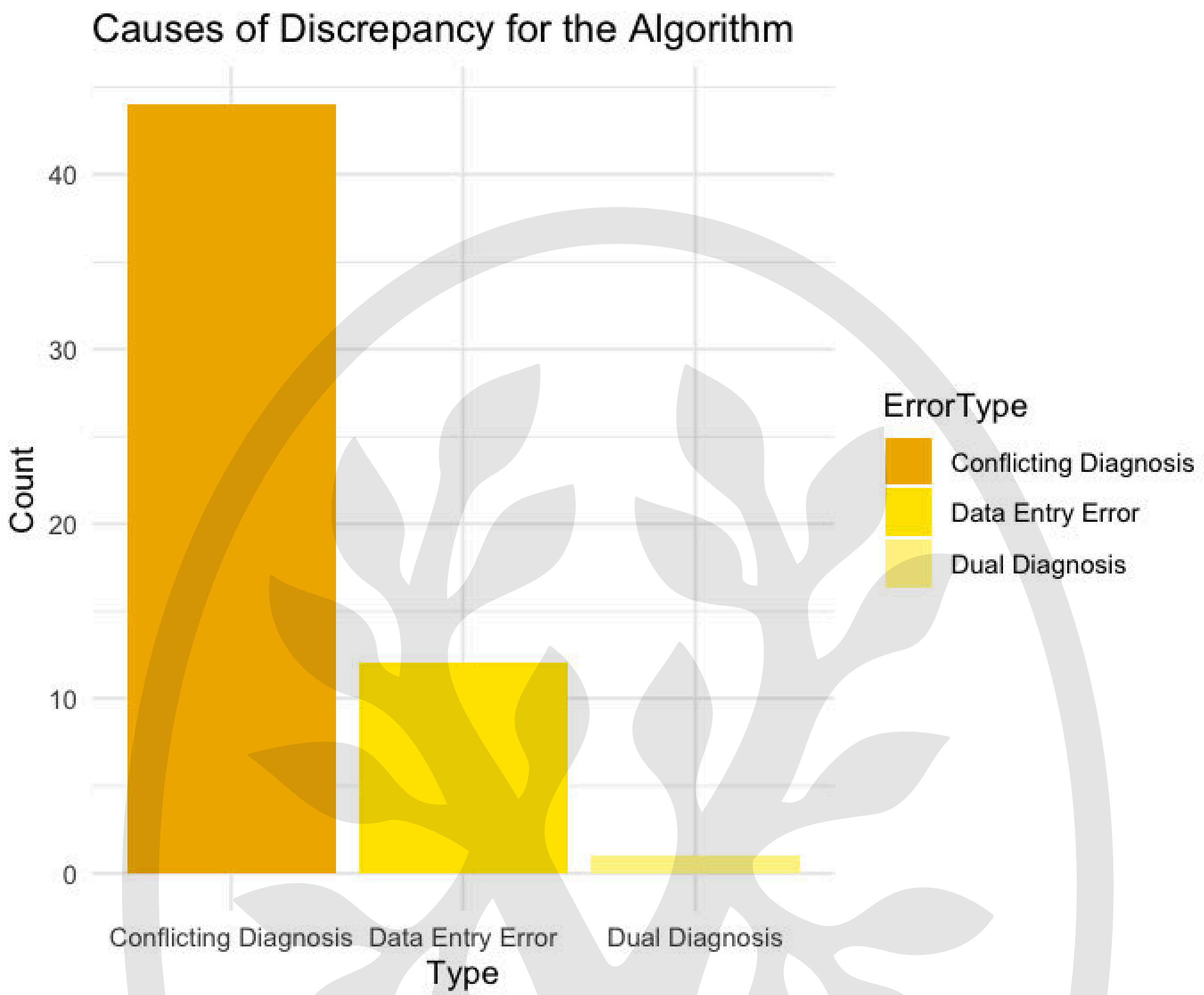
Figure 1. Visual representation of our methodology.

Figure 2. Discrepancies between Algorithm Outcomes and Existing Classifications in REDCap. The chart displays 94 discrepancies, categorized into 60 false positives (FP) and 34 false negatives (FN).

Figure 3. Causes of Discrepancy Within the Algorithm and the Gold Standard.

Figure 4. Distribution of Shared and Unique Codes Between Organizations. The blue section represents the percentage of additional ICD/CPT codes used exclusively at UTSW Medical Center, while the yellow section indicates the percentage of codes unique to the VUMC.The overlap between the blue and yellow areas represents the percentage of codes shared by both organizations.


Discrepancies Between the Algorithm's Outcome and the Existing Classifications in REDCap
94 Initial discrepancies
60 FP
34 FN
After Manual Chart Review
After Manual Chart Review
49 mislabeled records
11 FP
26 FN
8 mislabeled records
42 conflicting diagnoses
6 data entry errors
1 Dual Diagnosis
11 records were truly misclassified by the algorithm
26 records were truly misclassified by the algorithm
2 conflicting diagnoses
6 data entry errors
37 Final Discrepancies



Causes of Discrepancy for the Algorithm
Count
40
30
20
10
0
Conflicting Diagnosis
Data Entry Error
Dual Diagnosis
Type
ErrorType
Conflicting Diagnosis
Data Entry Error
Dual Diagnosis


11%
80.21
29805
01610
01622
01630
01638
01710
76140
75%
of the codes are shared between organizations
14%
23410
23397
29901
24341
0055T
BP3FYZZ
BP3EZZZ
76880
78662
78661