Diagnostic algorithms to study post-concussion syndrome using electronic health records: validating a method to capture an important patient population

Jessica Dennis[1,2], Aaron M. Yengo-Kahn[3,4], Paul Kirby[3], Gary S. Solomon[3,4], Nancy J. Cox[1,2], Scott L. Zuckerman[3,4]

[1]Division of Genetic Medicine, Department of Medicine, Vanderbilt University Medical Center Nashville, TN; [2]Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, TN; [3]Vanderbilt Sports Concussion Center, Vanderbilt University School of Medicine, Nashville, TN; [4]Department of Neurological Surgery, Vanderbilt University School of Medicine, Nashville, TN

Keywords:

Post-concussion syndrome, electronic health records, diagnostic algorithm, genetics

Corresponding Authors:

Jessica Dennis, MSc PhD
Division of Genetic Medicine
Vanderbilt University Medical Center
511J Light Hall
Nashville, TN 37232
Email: Jessica.dennis@vumc.org

Scott L. Zuckerman, MD MPH
Department of Neurological Surgery
Vanderbilt University School of Medicine
Medical Center North T-4224
Nashville, TN 37212
Email: scott.zuckerman@vumc.org

Abstract

Introduction: Post-concussion syndrome (PCS) is characterized by persistent cognitive, somatic, and emotional symptoms after a mild traumatic brain injury (mTBI). Genetic and other biological variables may contribute to PCS etiology, and the emergence of biobanks linked to electronic health records (EHR) offers new opportunities for research on PCS. We sought to validate the use of EHR data of PCS patients by comparing two diagnostic algorithms.

Methods: Vanderbilt University Medical Center curates a de-identified database of 2.8 million patient EHR. We developed two EHR-based algorithmic approaches that identified individuals with PCS by: (i) natural language processing (NLP) of narrative text in the EHR combined with structured demographic, diagnostic, and encounter data; or (ii) coded billing and procedure data. The predictive value of each algorithm was assessed, and cases and controls identified by each approach were compared on demographic and medical characteristics.

Results: First, the NLP algorithm identified 507 cases and 10,857 controls. The positive predictive value (PPV) in the cases was 82% and the negative predictive value in the controls was 78%. Second, the coded algorithm identified 1,142 patients with two or more PCS billing codes and had a PPV of 76%. Comparisons of PCS controls to both case groups recovered known epidemiology of PCS: cases were more likely than controls to be female and to have pre-morbid diagnoses of anxiety, migraine, and PTSD. In contrast, controls and cases were equally likely to have ADHD and learning disabilities, in accordance with the findings of recent systematic reviews of PCS risk factors.

Conclusions: EHR are a valuable research tool for PCS. Ascertainment based on coded data alone had a predictive value comparable to an NLP algorithm, recovered known PCS risk factors, and maximized the number of included patients.

Introduction

More than 1.7 million Americans visit an emergency department or are hospitalized for a traumatic brain injury (TBI) each year, a number thought to be an underestimation due to those who do not seek medical care.[1] TBI disproportionately affects the young and the old, and while motor vehicle collisions are a leading cause of TBI,[1,2] TBIs sustained in sport and military combat are increasingly recognized.[3,4]

Up to 90% of TBIs are classified as mild (mTBI), defined as positive brain imaging with Glasgow Coma Scale (GCS) scores of 13-15, or concussion, defined as negative brain imaging with a similar GCS score range.[2,5] Symptoms of an mTBI may include temporary confusion, loss of consciousness, amnesia, or other neurological abnormality such as seizure,[6,7] and these symptoms typically resolve spontaneously or within seven to 10 days.[5,8-10] Up to 30% of patients, however, do not recover in this temporal window and are unable to return to work, school, or other pre-injury activities for up to several months.[8,11,12] These patients complain of persistent somatic, cognitive, and emotional symptoms, known clinically as post-concussion syndrome (PCS).

The determinants of PCS are poorly understood. Recent studies show that females and those with low educational attainment endure a longer recovery, as do patients with a personal and/or family history of headaches, migraines, and mood disorders.[11-16] The literature is mixed, however, with regards to the effects of age, attention deficit hyperactive disorder (ADHD), and learning disability, though these factors may increase the risk of injury.[17,18] Initial injury severity and symptom burden also contribute to PCS risk, fueling the view that PCS is the result of both trauma to the brain and underlying psychological factors.[17,19]

Despite this knowledge, PCS prediction tools explain at most 30% of the variability in mTBI outcome,[11-16] spurring research into the role of other biological factors in recovery, including genetics.[20] Preliminary studies in small selected samples have found associations between candidate genes and short- and long-term TBI outcomes, but aside from the *APOE* e4 allele, which, in a meta-analysis of 14 studies[Zhou et al. 2008] increased the risk of poor 6 month outcomes by 36%, results have yet to be replicated in larger, representative populations. Moreover, there has been no agnostic search for genetic determinants of TBI recovery, as in a genome-wide association study (GWAS). Such hypothesis-free approaches have revealed novel biology in many rare and common diseases,[21] but require: 1) very large numbers of cases and controls to be adequately powered; and 2) replication of significant findings in an independent dataset to avoid false positives.[22]

Electronic health records (EHR) linked to biorepositories such as DNA databanks offer new opportunities to study the biology of TBI recovery and PCS. EHR are a rich data source, including information on pre-injury medical conditions, time-of-injury procedures, diagnoses, and post-injury follow-up care. The sample sizes afforded by EHR are unprecedented, and since these data are collected in the course of routine clinical care, EHR studies are less costly and time-consuming than clinically recruited samples or prospective trials. EHR-linked biobanks are also growing in number and scale worldwide. Within the United States, 10 institutions have joined the National Human Genome Research Institute-funded Electronic Medical Records and Genomics (eMERGE) network (https://emerge.mc. vanderbilt.edu) since 2007, with at least 126,000 samples genotyped to date. More ambitiously, the National Institutes of Health Precision Medicine Initiative *All of US* program (https://allofus.nih.gov) aims to recruit and follow over one million Americans, linking their EHR to physical assessments, participant surveys, mobile health technologies, and biospecimens. Similar data are already available from the first 500,000 participants enrolled in the UK Biobank (http://www.ukbiobank.ac.uk) as well as the first 500,000 enrolled in the US

Department of Veterans Affairs' Million Veteran Program (http://www.research.va.gov/mvp), and studies using these data across a range of disorders and research questions are proving the value of EHR-recruited samples.[23-25]

Defining valid study populations is a necessary first step to using EHR data. Manual review of EHR to identify a gold standard case-control set is time consuming, labor intensive, and impractical for large datasets. Algorithms are more efficient and usually take one of two forms.[23] Code-based algorithms apply filters to structured, coded data such as demographics, diagnostic and procedure codes, medications, and laboratory results. Algorithms based on natural language processing (NLP), on the other hand, use keyword searches and text mining software to extract data from unstructured text in the EHR. Both approaches have demonstrated merit,[23] and NLP-based algorithms are typically used in conjunction with filters on coded data. An algorithm's performance is usually judged by its positive predictive value (PPV), the probability that a case identified by the algorithm is a true case, and its negative predictive value (NPV), the probability that a control identified by the algorithm is free of the disease under study.

Given the far-reaching public health implications of mTBI and the need for research, we sought to study mTBI recovery using large patient datasets. The objective of our study was to evaluate the validity of two algorithms for identifying PCS cases and controls from EHR data: (i) an NLP algorithm with additional filters applied to coded data; and (ii) a coded algorithm, based on only a limited set of filters applied to coded data. Our evaluation proceeded in three stages. First, we manually reviewed a subset of records identified by the NLP and coded algorithms and calculated positive and negative predictive values. Second, we compared all identified cases and controls on known PCS risk factors to determine the

epidemiology of PCS in an EHR-recruited sample. Third, we evaluated the validity of billing code data on cause of injury and PCS symptoms to determine to guide future algorithm development efforts.

Methods

Data Source

Vanderbilt University Medical Center (VUMC) is a tertiary care hospital in Nashville, Tennessee, with affiliated clinics throughout the region. EHR have been used at VUMC since the early 1990s, and these data are compiled in a de-identified database known as the Synthetic Derivative (SD).[26] The SD contains more than 2.8 million patient EHR, and available data include basic demographics, billing codes (the majority of billing codes are International Classification of Diseases, Ninth Revision (ICD-9), with VUMC transitioning to Tenth Revision (ICD-10) in 2015), procedure codes, clinical documents, medications, laboratory values, and dates of inpatient stays and outpatient visits. The resource is continuously updating, and data extracted for this project were current to January 13, 2017. The Vanderbilt University Institutional Review Board approved this project (IRB #151116).

Gold Standard PCS Definition

PCS was defined as one or more post-concussion symptoms (Supplementary Table 1) that persisted beyond 14 days of an mTBI. The TBI was considered moderate / severe, and subsequently excluded, if it was: (i) penetrating; (ii) accompanied by brain surgery; (iii) associated with a GCS score <13; or (iv) accompanied by a hospitalization >5 days. A strict concussion definition was also considered, in which the TBI was ineligible in the presence of positive brain imaging.[5] Additional exclusion criteria were <5 years of age, or any confounding neurological disease diagnosed before the mTBI (brain tumor, stroke, seizures, meningitis, intracranial abscess, multiple sclerosis, Alzheimer disease, Parkinson disease, or other cerebral degeneration).

Sample Selection

(i) Natural Language Processing (NLP) Algorithm

We deployed a multi-step NLP algorithm, augmented by filters on coded data, to identify PCS cases and controls (Figure 1). The algorithm filtered based on billing codes, procedure codes, inpatient hospitalizations, outpatient visits, as well as keyword searches and NLP of clinical notes. In step 1, patients with mTBI were identified using a list of ICD-9 and ICD-10 mTBI codes (Supplementary Table 2). In step 2, we excluded patients with evidence of severe TBI, defined as *any* TBI code (Supplementary Table 3) followed by a neurosurgical procedure code (ICD-9 01* and 02*; ICD-10 00*, 0N*, R40.242, R40.243; Current Procedural Terminology 61000-62258) within 7 days. These neurosurgical procedure codes were previously found to be 97% sensitive and 94% specific for severe TBI.[27] In step 3, we identified the most recent eligible mTBI using discharge windows as a proxy for TBI severity, and excluded mTBI codes that fell within a discharge window of >5 days. If a patient had no eligible mTBI codes, they were excluded. In step 4, patients meeting the following criteria were excluded: <5 years of age at the mTBI; a moderate or severe TBI (defined as any TBI code occurring in a discharge window >5 days) within 365 days of the mTBI; or a history of confounding neurological disease diagnosed before the mTBI (Supplementary Table 4).

In step 5, patients were classified cases and controls based on the occurrence of PCS billing codes (ICD-9 310.2, ICD-10 F07.81) or keywords, including misspellings, parsed from clinical documents. Controls had no evidence of PCS anywhere in their record, and additionally had to have visited VUMC as an outpatient at least once in the 365 days before and once in the 365 days after the mTBI (i.e., using

VUMC as their "medical home"). Cases had to have a PCS billing code or keyword on the same day as at least one PCS symptom billing code or keyword (Supplementary Table 1), and the PCS code or keyword needed to occur at least 14 days and up to 365 days after the mTBI. Patients identified by this algorithm are hereafter called NLP cases and NLP controls.

(ii) Coded Algorithm

We selected a second case group entirely independent of the NLP algorithm, based on only filters applied to coded data. Cases were required to have two or more instances of the PCS billing code on different days,[24,28] and we excluded patients with any neurosurgical procedure code in their record, who were <5 years of age at their first PCS code, and those with confounding neurological histories diagnosed before their first PCS billing code. Patients identified by this algorithm are hereafter called coded cases.

Variable Definitions

Structured data were used to describe the injury and symptom characteristics of patients identified by the NLP and coded algorithms. TBI morbidity group was derived by mapping ICD codes to categories of skull fracture, contusion, hemorrhage, concussion, and other/unspecified (Supplementary Table 3). Brain imaging was captured by ICD billing and procedure codes (Supplementary Table 5). TBI cause was assigned by mapping ICD external cause codes to categories of motor vehicle collision, fall, struck by and against, and assault (Supplementary Table 6). Finally, data on PCS symptoms were extracted from ICD codes (Supplementary Table 1).

Manual Review

The validity of both algorithms was evaluated by manual chart review of 50 NLP cases, 50 NLP controls, and 50 coded cases. Records were randomly selected by JD, ensuring that the NLP and coded case groups were non-overlapping. We extracted data on TBI severity, care sought, cause of injury, and subsequent PCS symptoms. A REDCap data collection instrument was piloted on 20 records, 10 of which were reviewed in triplicate (JD, PK, and AYK or SLZ), with high inter-rater agreement (Cohen's kappa=0.80). The remaining 140 records were reviewed by PK, with all disagreements resolved by consensus.

Statistical Analyses

The NPV of the NLP algorithm was the proportion of manually-reviewed controls with no PCS symptoms persisting beyond 14 days of an eligible mTBI or concussion. The PPV of the NLP and coded algorithms were the proportion of manually-reviewed cases that met the broad mTBI and strict concussion definitions of PCS.

Each case group was compared to controls on age, sex, and pre-morbid diagnoses that have previously been associated with PCS risk: ADHD, anxiety, depression, learning disability, migraine, and post-traumatic stress disorder (PTSD). Diagnoses were captured by billing codes (Supplementary Table 7), and we only considered those that occurred before the eligible mTBI in NLP patients, and before the first PCS code in coded patients. An age difference was detected by a t-test while categorical variables were evaluated by chi-square tests. Statistical tests were two-sided, and all analyses were conducted in R version 3.3.[29]

Sport is a common cause of TBI, but no ICD external cause code exists specifically for sport. We therefore investigated how sport injuries were captured by external cause codes in the EHR in the 150 manually reviewed records. Likewise, the billing codes that characterize PCS symptoms are not specific to PCS, but may still recapitulate PCS symptoms. We therefore calculated the sensitivity and specificity of symptom billing codes in all true positive PCS patients identified by manual review. Sensitivity was the probability of being assigned a symptom billing code, given the presence of the symptom; specificity was the probability of not being assigned a symptom billing code, given the absence of the symptom.

Results

Sample Characteristics

The NLP algorithm identified 507 cases and 10,857 controls in the SD. Cases were on average two years younger than controls (Table 1), and were more likely to be female and white. More than half of mTBI codes in cases and controls mapped to the Concussion / loss of consciousness TBI morbidity group, while the remainder mostly fell into the Other and unspecified head injury category. Motor vehicle collision was the most common cause of injury in both cases and controls. Nearly one fifth of patients were identified by PCS keywords as opposed to billing codes, and only 46.4% had two or more PCS billing codes. Memory difficulties (18.1%), dizziness (17.9%), and headache (15.2%) were the most common PCS symptoms captured by billing codes in the year following the mTBI in cases.

The coded algorithm identified 5039 patients with at least one PCS code and 1372 patients with at least two PCS codes. Of these, we excluded 164 patients with a history of neurological disease when the first PCS code was assigned, 30 who were less than five years of age, and 36 who had a neurosurgical procedure code, for a final sample of 1142 coded cases (266 patients were identified as cases by both algorithms). Exactly half of cases were female. A TBI code was found in 61.8% of cases, suggesting that many PCS patients either did not seek care for the TBI, or sought care outside of VUMC. Most TBI codes mapped to the milder categories of Concussion / loss of consciousness and Other and unspecified head injury. Nearly half of patients had brain imaging, and motor vehicle collision was again the most common cause of injury. The top three PCS symptoms were headache (31.9%), memory difficulties (14.3%), and dizziness (13.0%).

Algorithm Predictive Values

The NPV of the NLP algorithm was 78% and its PPV was 82% (Table 2). Most controls were excluded because the TBI was either of unclear severity or was moderate/severe while most cases were excluded because PCS symptoms resolved within 14 days of the TBI. A single case was excluded for other reasons (a prior pituitary tumor). When we applied the strict concussion definition, the NPV of the algorithm was unchanged while the PPV dropped to 76%.

The PPV of the coded algorithm was 76% (Table 2). The top reasons for exclusion were an absence of symptoms and evidence of moderate/severe TBI. Applying the strict concussion definition reduced the PPV to 66%.

Epidemiology of PCS in EHR

Cases identified by both approaches were more likely than controls to be female (Table 1; p=3.25x10-7 for NLP cases and p=2.25x10-7 for coded cases). The association with age, however, was discrepant across case groups; compared to controls, NLP cases were two years younger on average (p=1.58x10-3) while coded cases were nearly four years older (p=4.15x10-11). Pre-morbid anxiety, migraine, and PTSD were more frequent in both case groups versus controls (Table 3). A higher prevalence of depression in cases was also found, but the increase was only statistically significant in the larger sample of coded cases. In contrast, diagnoses with uncertain associations with PCS, ADHD and learning disabilities, were equally common in our cases and controls, suggesting that the enrichment of pre-morbid anxiety, migraine, and PTSD in our cases captured true associations as opposed to a systematic under-reporting of diagnoses in controls.

Evaluation of Structured Data on TBI Cause and PCS Symptoms

Cause of injury and symptom billing codes captured some, but not all of the data in the EHR. Of the 150 manually reviewed subjects, 64 were injured in a motor vehicle collision, but only 36 had a motor vehicle collision external cause code (Table 4). Sports was the next leading cause of injury (35/150), but 28 of these patients were coded as having an unknown cause of injury. The sensitivity of billing codes that characterized PCS symptoms ranged from 0 to 50% in the sample of 79 true PCS cases (Table 5), but the codes were very specific, so their absence could be useful in defining future control populations without PCS symptoms.

Discussion

TBIs are sustained by more than 1 in 120 Americans annually,[1] and adverse outcomes affect many. In this first attempt to capture patients with mTBI and PCS from an EHR, we report moderate predictive values for both an NLP-based algorithm, and a simpler coded algorithm that used only structured data. We also show that including all algorithm-identified patients in case-control comparisons recovered the known epidemiology of PCS, suggesting that when the sample size is large, true associations can be detected despite imperfect patient classification. Female gender, prior mood disorders, and headache history are some of the strongest predictors of poor recovery after an mTBI.[11-16] We used only EHR data to capture pre-injury medical histories and found very statistically significant associations with these variables. Our results indicate that EHR linked to biorepositories could be used to identify novel biological determinants of recovery, ultimately leading to early identification and intervention for patients at risk.

The NLP algorithm had the best PPV and detected patients by PCS keywords that were not captured by the coded algorithm. Nonetheless, the number of cases captured by the NLP algorithm was less than one tenth of the patients in the SD with a single PCS billing code, and requiring coded documentation of the TBI prior to the PCS diagnosis excluded many potential cases. The algorithm was also time-consuming to develop and test, and complicated to implement. These strengths and limitations of NLP algorithms are well recognized.[23] EHR algorithms for autism,[30] bipolar disorder,[31] and major depression[23] had the best PPV (~85%) when narrative text was included, and a systematic review of 19 algorithms across a range of chronic and infectious conditions found that incorporating narrative EHR text improved case detection from 61.7% to 78.1%.[32] Nonetheless, fewer patients have sufficient EMR data to be classified by NLP algorithms, leading to considerable reductions in sample size.[23,33]

The coded algorithm had a lower PPV, but captured twice as many cases. Diagnostic and procedure codes can be "noisy", yet they have been used repeatedly in EHR research to validate known environmental and genetic associations, and for novel discovery.[24,25,33-36] In the largest proof-of-concept experiment, 77 robust disease-gene associations were tested for replication using disease diagnoses derived from ICD-9 billing codes in 13,835 patients.[37] Sixty-six percent of the associations replicated.

Imperfect codes may also better represent the true clinical spectrum of disease. We excluded patients with moderate / severe TBI and subsequent PCS symptoms, but these patients may be valuable to include if the biology that underlies TBI recovery is insensitive to TBI severity. Large prospective registry-based studies support this hypothesis. They have found an increased incidence of adverse psychological, cognitive, emotional, and social outcomes across the TBI severity spectrum, just to a lesser extent in mTBI patients.[38-41] Future EHR-based studies of PCS could therefore test all PCS patients, regardless of TBI severity, in sensitivity analyses.

We compared our algorithms' performance to a gold standard PCS definition, but diagnostic criteria for PCS are contentious.[42] The term PCS was first used in the 1930s. In the ICD-9 manual, PCS was described as a constellation of symptoms, and in the DSM-IV and ICD-10 manuals, explicit diagnostic criteria were provided.[43] The agreement between DSM-IV and ICD-10 criteria, however, is poor,[43] the DSM-5 has removed the PCS diagnosis entirely in favor of defining mild or major neurocognitive disorder following TBI,[44] and few clinicians even adhere to DSM or ICD guidelines in practice.[45] Large-scale EHR data offer new opportunities to resolve these diagnostic discrepancies by revealing how TBI patients actually cluster and recover clinically. Data-driven approaches have already uncovered novel autism spectrum disorder subgroups using the EHR,[46] and machine learning algorithms applied to EHR data

have accurately predicted suicide attempts up to two years in advance.[47] Our study is an important first step towards applying similar tools to the EHR of TBI patients.

Limitations and Strengths

Algorithms were developed and tested at only a single institution. Also, the PPVs were modest, but in line with the PPVs of EHR algorithms published for other psychiatric diseases.[23,30,31] The NPV of the NLP algorithm, on the other hand, was lower than expected. Insufficient data to classify TBI severity was a leading reason for excluding controls, emphasizing the challenges of using EHR data for research when care is fractured across providers. Nonetheless, this study demonstrates the feasibility of ascertaining large samples of PCS cases and controls from the EHR, and is a valuable starting place for future algorithm development efforts.

Conclusions

EHR linked to biorepositories hold great promise for understanding TBI sequelae. They will accelerate investigations of novel genetic and other biomarkers of TBI recovery, and could enable data-driven discoveries of disease subtypes and prognosis. Two strategies were employed in the current study to ascertain patients with mTBI and PCS from the EHR, and each strategy had tradeoffs. On balance, however, the coded algorithm had a reasonable PPV, recovered known PCS associations and maximized the number of included patients. No single approach will be perfect, and future studies should continue to compare different diagnostic algorithms. EHRs are a vast, growing resource available to many independent research groups, and the use of EHR-linked biobanks in TBI research will have far-reaching public health benefits.

Acknowledgements:

The authors thank Lea K. Davis and Rebecca T. Levinson for comments on earlier iterations of the manuscript, and Doug Conway for help implementing the algorithms and navigating access to the EHR data.

Jessica Dennis is supported by the Canadian Institutes of Health Research (award MFE-142936). Support for Nancy J. Cox was provided by R01 MH113362, U54MD010722, and U01HG009086. This project was supported by a Vanderbilt Institute for Clinical and Translational Research micro-grant (VR20299) and was conducted in part using the resources of the Advanced Computing Center for Research and Education at Vanderbilt University, Nashville, TN. The datasets used for this project were obtained from Vanderbilt University Medical Center's Synthetic Derivative, which is supported by numerous sources: institutional funding, private agencies, and federal grants. These include the NIH funded Shared Instrumentation Grant S10RR025141; and CTSA grants UL1TR002243, UL1TR000445, and UL1RR024975. The REDCap tool used in this study was supported by grants UL1 TR000445 from NCATS/NIH.

Author Disclosure Statement:

Gary Solomon is a consultant for the Nashville Predators, Tennessee Titans, and the athletic departments of Tennessee Tech University and the University of Tennessee, fees paid to institution. He is also a consultant to the National Football League Department of Health and Safety.

The other authors have no competing financial interests.

References

1. Faul M, XU L, Wald MM, Coronado VG. Traumatic Brain Injury in the United States: Emergency Department Visits, Hospitalizations and Deaths 2002-2006. Atlanta (GA); 2010.

2. Cassidy JD, Carroll LJ, Peloso PM, Borg J, von Holst H, Holm L, Kraus J, Coronado VG, Injury WHOCCTFoMTB. Incidence, risk factors and prevention of mild traumatic brain injury: results of the WHO Collaborating Centre Task Force on Mild Traumatic Brain Injury. *J Rehabil Med*. 2004(43 Suppl):28-60.

3. Wojcik BE, Stein CR, Bagg K, Humphrey RJ, Orosco J. Traumatic brain injury hospitalizations of U.S. army soldiers deployed to Afghanistan and Iraq. *Am J Prev Med*. 2010;38(1 Suppl):S108-116.

4. Zuckerman SL, Kerr ZY, Yengo-Kahn A, Wasserman E, Covassin T, Solomon GS. Epidemiology of Sports-Related Concussion in NCAA Athletes From 2009-2010 to 2013-2014: Incidence, Recurrence, and Mechanisms. *Am J Sports Med*. 2015;43(11):2654-2662.

5. McCrory P, Meeuwisse W, Dvorak J, Aubry M, Bailes J, Broglio S, Cantu RC, Cassidy D, Echemendia RJ, Castellani RJ, Davis GA, Ellenbogen R, Emery C, Engebretsen L, Feddermann-Demont N, Giza CC, Guskiewicz KM, Herring S, Iverson GL, Johnston KM, Kissick J, Kutcher J, Leddy JJ, Maddocks D, Makdissi M, Manley GT, McCrea M, Meehan WP, Nagahiro S, Patricios J, Putukian M, Schneider KJ, Sills A, Tator CH, Turner M, Vos PE. Consensus statement on concussion in sport-the 5th international conference on concussion in sport held in Berlin, October 2016. *Br J Sports Med*. 2017.

6. Carroll LJ, Cassidy JD, Holm L, Kraus J, Coronado VG, Injury WHOCCTFoMTB. Methodological issues and research recommendations for mild traumatic brain injury: the WHO Collaborating Centre Task Force on Mild Traumatic Brain Injury. *J Rehabil Med*. 2004(43 Suppl):113-125.

7. Jagoda AS, Bazarian JJ, Bruns JJ, Jr., Cantrill SV, Gean AD, Howard PK, Ghajar J, Riggio S, Wright DW, Wears RL, Bakshy A, Burgess P, Wald MM, Whitson RR, American College of Emergency P, Centers

for Disease C, Prevention. Clinical policy: neuroimaging and decisionmaking in adult mild traumatic brain injury in the acute setting. *Ann Emerg Med*. 2008;52(6):714-748.

8. Cancelliere C, Kristman VL, Cassidy JD, Hincapie CA, Cote P, Boyle E, Carroll LJ, Stalnacke BM, Nygren-de Boussard C, Borg J. Systematic review of return to work after mild traumatic brain injury: results of the International Collaboration on Mild Traumatic Brain Injury Prognosis. *Arch Phys Med Rehabil*. 2014;95(3 Suppl):S201-209.

9. Carroll LJ, Cassidy JD, Cancelliere C, Cote P, Hincapie CA, Kristman VL, Holm LW, Borg J, Nygren-de Boussard C, Hartvigsen J. Systematic review of the prognosis after mild traumatic brain injury in adults: cognitive, psychiatric, and mortality outcomes: results of the International Collaboration on Mild Traumatic Brain Injury Prognosis. *Arch Phys Med Rehabil*. 2014;95(3 Suppl):S152-173.

10. Kamins J, Bigler E, Covassin T, Henry L, Kemp S, Leddy JJ, Mayer A, McCrea M, Prins M, Schneider KJ, Valovich McLeod TC, Zemek R, Giza CC. What is the physiological time to recovery after concussion? A systematic review. *Br J Sports Med*. 2017;51(12):935-940.

11. van der Naalt J, Timmerman ME, de Koning ME, van der Horn HJ, Scheenen ME, Jacobs B, Hageman G, Yilmaz T, Roks G, Spikman JM. Early predictors of outcome after mild traumatic brain injury (UPFRONT): an observational cohort study. *Lancet Neurol*. 2017;16(7):532-540.

12. Zemek R, Barrowman N, Freedman SB, Gravel J, Gagnon I, McGahern C, Aglipay M, Sangha G, Boutis K, Beer D, Craig W, Burns E, Farion KJ, Mikrogianakis A, Barlow K, Dubrovsky AS, Meeuwisse W, Gioia G, Meehan WP, 3rd, Beauchamp MH, Kamil Y, Grool AM, Hoshizaki B, Anderson P, Brooks BL, Yeates KO, Vassilyadi M, Klassen T, Keightley M, Richer L, DeMatteo C, Osmond MH, Pediatric Emergency Research Canada Concussion T. Clinical Risk Score for Persistent Postconcussion Symptoms Among Children With Acute Concussion in the ED. *JAMA*. 2016;315(10):1014-1025.

13. Stulemeijer M, van der Werf S, Borm GF, Vos PE. Early prediction of favourable recovery 6 months after mild traumatic brain injury. *J Neurol Neurosurg Psychiatry*. 2008;79(8):936-942.

14. Jacobs B, Beems T, Stulemeijer M, van Vugt AB, van der Vliet TM, Borm GF, Vos PE. Outcome prediction in mild traumatic brain injury: age and clinical variables are stronger predictors than CT abnormalities. *J Neurotrauma*. 2010;27(4):655-668.

15. Lingsma HF, Yue JK, Maas AI, Steyerberg EW, Manley GT, Investigators T-T. Outcome prediction after mild and complicated mild traumatic brain injury: external validation of existing models and identification of new predictors using the TRACK-TBI pilot study. *J Neurotrauma*. 2015;32(2):83-94.

16. Cnossen MC, Winkler EA, Yue JK, Okonkwo DO, Valadka A, Steyerberg EW, Lingsma H, Manley GTMDPD. Development of a Prediction Model for Post-Concussive Symptoms following Mild Traumatic Brain Injury: A TRACK-TBI Pilot Study. *J Neurotrauma*. 2017.

17. Iverson GL, Gardner AJ, Terry DP, Ponsford JL, Sills AK, Broshek DK, Solomon GS. Predictors of clinical recovery from concussion: a systematic review. *Br J Sports Med*. 2017;51(12):941-948.

18. Silverberg ND, Gardner AJ, Brubacher JR, Panenka WJ, Li JJ, Iverson GL. Systematic review of multivariable prognostic models for mild traumatic brain injury. *J Neurotrauma*. 2015;32(8):517-526.

19. Bryant RA. Disentangling mild traumatic brain injury and stress reactions. *N Engl J Med*. 2008;358(5):525-527.

20. McCrea M, Meier T, Huber D, Ptito A, Bigler E, Debert CT, Manley G, Menon D, Chen JK, Wall R, Schneider KJ, McAllister T. Role of advanced neuroimaging, fluid biomarkers and genetic testing in the assessment of sport-related concussion: a systematic review. *Br J Sports Med*. 2017;51(12):919-929.

21. MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, Junkins H, McMahon A, Milano A, Morales J, Pendlington ZM, Welter D, Burdett T, Hindorff L, Flicek P, Cunningham F, Parkinson H. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res*. 2017;45(D1):D896-D901.

22. Burton PR, Hansell AL, Fortier I, Manolio TA, Khoury MJ, Little J, Elliott P. Size matters: just how big is BIG?: Quantifying realistic sample size requirements for human genome epidemiology. *Int J Epidemiol*. 2009;38(1):263-273.

23. Smoller JW. The use of electronic health records for psychiatric phenotyping and genomics. *Am J Med Genet B Neuropsychiatr Genet*. 2017.

24. Denny JC, Bastarache L, Roden DM. Phenome-Wide Association Studies as a Tool to Advance Precision Medicine. *Annu Rev Genomics Hum Genet*. 2016;17:353-373.

25. Wei WQ, Denny JC. Extracting research-quality phenotypes from electronic health records to support precision medicine. *Genome Med*. 2015;7(1):41.

26. Roden DM, Pulley JM, Basford MA, Bernard GR, Clayton EW, Balser JR, Masys DR. Development of a large-scale de-identified DNA biobank to enable personalized medicine. *Clin Pharmacol Ther*. 2008;84(3):362-369.

27. Carroll CP, Cochran JA, Guse CE, Wang MC. Are we underestimating the burden of traumatic brain injury? Surveillance of severe traumatic brain injury using centers for disease control International classification of disease, ninth revision, clinical modification, traumatic brain injury codes. *Neurosurgery*. 2012;71(6):1064-1070; discussion 1070.

28. Wei WQ, Teixeira PL, Mo H, Cronin RM, Warner JL, Denny JC. Combining billing codes, clinical notes, and medications from electronic health records provides superior phenotyping performance. *J Am Med Inform Assoc*. 2016;23(e1):e20-27.

29. R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2017.

30. Lingren T, Chen P, Bochenek J, Doshi-Velez F, Manning-Courtney P, Bickel J, Wildenger Welchons L, Reinhold J, Bing N, Ni Y, Barbaresi W, Mentch F, Basford M, Denny J, Vazquez L, Perry C, Namjou B, Qiu H, Connolly J, Abrams D, Holm IA, Cobb BA, Lingren N, Solti I, Hakonarson H, Kohane IS, Harley J,

Savova G. Electronic Health Record Based Algorithm to Identify Patients with Autism Spectrum Disorder. *PLoS One*. 2016;11(7):e0159621.

31. Castro VM, Minnier J, Murphy SN, Kohane I, Churchill SE, Gainer V, Cai T, Hoffnagle AG, Dai Y, Block S, Weill SR, Nadal-Vicens M, Pollastri AR, Rosenquist JN, Goryachev S, Ongur D, Sklar P, Perlis RH, Smoller JW, International Cohort Collection for Bipolar Disorder C. Validation of electronic health record phenotyping of bipolar disorder cases and controls. *Am J Psychiatry*. 2015;172(4):363-372.

32. Ford E, Carroll JA, Smith HE, Scott D, Cassell JA. Extracting information from the text of electronic medical records to improve case detection: a systematic review. *J Am Med Inform Assoc*. 2016;23(5):1007-1015.

33. Chen CY, Lee PH, Castro VM, Minnier J, Charney AW, Stahl EA, Ruderfer DM, Murphy SN, Gainer V, Cai T, Jones I, Pato CN, Pato MT, Landen M, Sklar P, Perlis RH, Smoller JW. Genetic validation of bipolar disorder identified by automated phenotyping using electronic health records. *Transl Psychiatry*. 2018;8(1):86.

34. Ritchie MD, Denny JC, Crawford DC, Ramirez AH, Weiner JB, Pulley JM, Basford MA, Brown-Gentry K, Balser JR, Masys DR, Haines JL, Roden DM. Robust replication of genotype-phenotype associations across multiple diseases in an electronic medical record. *Am J Hum Genet*. 2010;86(4):560-572.

35. Denny JC, Ritchie MD, Basford MA, Pulley JM, Bastarache L, Brown-Gentry K, Wang D, Masys DR, Roden DM, Crawford DC. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics*. 2010;26(9):1205-1210.

36. Newton KM, Peissig PL, Kho AN, Bielinski SJ, Berg RL, Choudhary V, Basford M, Chute CG, Kullo IJ, Li R, Pacheco JA, Rasmussen LV, Spangler L, Denny JC. Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network. *J Am Med Inform Assoc*. 2013;20(e1):e147-154.

37. Denny JC, Bastarache L, Ritchie MD, Carroll RJ, Zink R, Mosley JD, Field JR, Pulley JM, Ramirez AH, Bowton E, Basford MA, Carrell DS, Peissig PL, Kho AN, Pacheco JA, Rasmussen LV, Crosslin DR, Crane PK, Pathak J, Bielinski SJ, Pendergrass SA, Xu H, Hindorff LA, Li R, Manolio TA, Chute CG, Chisholm RL, Larson EB, Jarvik GP, Brilliant MH, McCarty CA, Kullo IJ, Haines JL, Crawford DC, Masys DR, Roden DM. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat Biotechnol*. 2013;31(12):1102-1110.

38. Orlovska S, Pedersen MS, Benros ME, Mortensen PB, Agerbo E, Nordentoft M. Head injury as risk factor for psychiatric disorders: a nationwide register-based follow-up study of 113,906 persons with head injury. *Am J Psychiatry*. 2014;171(4):463-469.

39. Sariaslan A, Sharp DJ, D'Onofrio BM, Larsson H, Fazel S. Long-Term Outcomes Associated with Traumatic Brain Injury in Childhood and Adolescence: A Nationwide Swedish Cohort Study of a Wide Range of Medical and Social Outcomes. *PLoS Med*. 2016;13(8):e1002103.

40. Gardner RC, Burke JF, Nettiksimmons J, Kaup A, Barnes DE, Yaffe K. Dementia risk after traumatic brain injury vs nonbrain trauma: the role of age and severity. *JAMA Neurol*. 2014;71(12):1490-1497.

41. Gardner RC, Burke JF, Nettiksimmons J, Goldman S, Tanner CM, Yaffe K. Traumatic brain injury in later life increases risk for Parkinson disease. *Ann Neurol*. 2015;77(6):987-995.

42. Cassidy JD, Cancelliere C, Carroll LJ, Cote P, Hincapie CA, Holm LW, Hartvigsen J, Donovan J, Nygren-de Boussard C, Kristman VL, Borg J. Systematic review of self-reported prognosis in adults after mild traumatic brain injury: results of the International Collaboration on Mild Traumatic Brain Injury Prognosis. *Arch Phys Med Rehabil*. 2014;95(3 Suppl):S132-151.

43. Boake C, McCauley SR, Levin HS, Contant CF, Song JX, Brown SA, Goodman HS, Brundage SI, Diaz-Marchan PJ, Merritt SG. Limited agreement between criteria-based diagnoses of postconcussional syndrome. *J Neuropsychiatry Clin Neurosci*. 2004;16(4):493-499.

44. Wortzel HS, Arciniegas DB. The DSM-5 approach to the evaluation of traumatic brain injury and its neuropsychiatric sequelae. *NeuroRehabilitation*. 2014;34(4):613-623.

45. Rose SC, Fischer AN, Heyer GL. How long is too long? The lack of consensus regarding the post-concussion syndrome diagnosis. *Brain Inj*. 2015;29(7-8):798-803.

46. Doshi-Velez F, Ge Y, Kohane I. Comorbidity clusters in autism spectrum disorders: an electronic health record time-series analysis. *Pediatrics*. 2014;133(1):e54-63.

47. Walsh CG, Ribeiro JD, Franklin JC. Predicting Risk of Suicide Attempts Over Time Through Machine Learning. *Clinical Psychological Science*. 2017;5(3):457-469.

Figure Legend

Figure 1. Inclusion and exclusion criteria used by the natural language processing algorithm to select mild traumatic brain injury (mTBI) patients from the Vanderbilt University Medical Center Synthetic Derivative (VUMC SD), with and without subsequent post-concussion syndrome (PCS).

Table 1. Characteristics of patients identified by the NLP and coded algorithms.

	NLP controls (N=10857)	NLP cases (N=507)	Coded cases (N=1142)
Female gender, N (%)	4259 (39.2)	257 (50.7%)	571 (50.0)
Age, mean (SD)[a]	30.29 (18.54)	28.25 (17.15)	34.09 (18.36)
Pediatric (<18 years of age), N (%)[a]	3842 (35.4)	222 (43.8)	294 (25.7)
Self-identified race, N (%)			
White	8325 (76.7)	407 (80.3)	924 (80.9)
African American	1483 (13.7)	55 (10.8)	111 (9.7)
Other / Unknown	1049 (9.7)	45 (8.9)	107 (9.4)
TBI code and morbidity group, N (%)[b]	10857 (100)	507 (100)	706 (61.8)
Skull fracture	13 (0.1)	1 (0.2)	94 (8.2)
Contusion	0 (0)	0 (0)	56 (4.9)
Hemorrhage	0 (0)	1 (0.2)	120 (10.5)
Concussion / loss of consciousness	6349 (58.5)	262 (51.7)	444 (38.9)
Other and unspecified head injury	4918 (45.3)	253 (49.9)	560 (49.0)
Brain imaging, N (%)	5714 (52.6)	157 (31.0)	544 (47.6)
TBI cause, N (%)			
Motor vehicle collision	3877 (35.7)	90 (17.8)	246 (21.5)
Fall	1156 (10.6)	38 (7.5)	92 (8.1)
Struck by and against	684 (6.3)	47 (9.3)	63 (5.5)
Assault	330 (3.0)	6 (1.2)	20 (1.8)
Unknown	4810 (44.3)	326 (64.3)	721 (63.1)
Any PCS code	0	420 (82.8)	1142 (100)
Two or more PCS codes	0	235 (46.4)	1142 (100)
PCS symptom codes			
Any	799 (7.4)	208 (41.0)	582 (51.0)
Anxiety	121 (1.1)	21 (4.1)	61 (5.3)
Apathy	0	0	0 (0)
Depression	325 (3.0)	37 (7.3)	39 (3.4)
Concentration difficulties	7 (0.1)	20 (3.9)	13 (1.1)
Dizziness	232 (2.1)	91 (17.9)	148 (13.0)
Emotional lability	0	0	0 (0)
Fatigue	109 (1.0)	19 (3.7)	11 (1.0)
Headache	102 (0.9)	77 (15.2)	364 (31.9)
Irritable	1 (0)	0	1 (0.1)
Impulsive	0	0	0 (0)
Memory difficulties	164 (1.5)	92 (18.1)	163 (14.3)
Sleep disturbances	91 (0.8)	16 (3.2)	32 (2.8)
Other	34 (0.3)	7 (1.4)	31 (2.7)

N denotes number; NLP, natural language processing; PCS, post-concussion syndrome; SD, standard deviation; and TBI, traumatic brain injury.

[a]Age was determined at the time of the most recent eligible mTBI code in NLP patients and at the time of the first PCS billing code in coded patients.

[b]Some NLP patients received multiple TBI codes on the date of the qualifying event, and some coded cases had multiple TBI codes within the window of 365 days before and 30 days after the first PCS billing code. Each TBI code was mapped to TBI type categories, which is why the row counts exceed the number of individuals in each group.

Table 2. Negative and positive predictive values of the algorithms determined in randomly selected sample of 150 patients.

	NLP controls (N=50)	NLP cases (N=50)	Coded cases (N=50)
Symptom-based exclusion criteria applied to controls			
PCS symptoms persist beyond 14 days of the index TBI and are first reported <365 days after the index TBI	3	NA	NA
Symptom-based exclusion criteria applied to cases			
No PCS symptoms after the index TBI	NA	2	4
PCS symptoms resolve within 14 days of the index TBI	NA	4	0
PCS symptoms are first reported >365 days after the index TBI	NA	0	3
TBI-based exclusion criteria applied to controls and cases			
Index TBI was of unclear severity	4	0	1
Index TBI was moderate/severe (categories are not mutually exclusive)	4	3	4
Penetrating head injury	0	0	0
Brain surgery within 7 days of index TBI	0	0	0
Glasgow coma scale score ≤13	1	0	3
Hospitalization >5 days	3	2	2
Other reasons for exclusion	0	1	0
Patients remaining	39	41	38
Predictive value	**NPV=78%**	**PPV=82%**	**PPV=76%**
Positive imaging	0	3	5
Patients remaining	39	38	33
Predictive value (strict concussion definition)	**NPV=78%**	**PPV=76%**	**PPV=66%**

N denotes number; NPV, negative predictive value; PCS, post-concussion syndrome; PPV, positive predictive value; and TBI, traumatic brain injury.

Table 3. Pre-injury diagnoses in controls and cases. Diagnoses (Supplementary Table 7) had to occur before the eligible mTBI in NLP patients, and before the first PCS billing code in coded patients.

Diagnosis	NLP controls (N=10857)	NLP cases (N=507)	p-value vs. controls	Coded cases (N=1142)	p-value vs. controls
ADHD, N (%)	227 (2.1)	14 (2.8)	0.39	24 (2.1)	1
Anxiety, N (%)	560 (5.2)	39 (7.7)	0.02	112 (9.8)	1.26E-10
Depression, N (%)	581 (5.4)	36 (7.1)	0.11	101 (8.8)	1.74E-06
Learning disability, N (%)	129 (1.2)	10 (2.0)	0.17	19 (1.7)	0.21
Migraine, N (%)	246 (2.3)	35 (6.9)	1.31E-10	117 (10.2)	4.16E-50
PTSD, N (%)	100 (0.9)	14 (2.8)	1.25E-04	27 (2.4)	1.18E-05

ADHD denotes attention deficit hyperactive disorder; N, number; NLP, natural language processing; and PTSD, post-traumatic stress disorder.

Table 4. Causes of injury extracted manually versus captured by ICD external cause codes (mappings in Supplementary Table 6) in the sample of 150 manually reviewed subjects.

Manual extraction	ICD external cause code groupings					Total (%)
	Motor vehicle collision	Fall	Struck by and against	Assault	Unknown	
Motor vehicle collision	36	0	1	0	27	64 (42.7)
Fall	0	11	1	1	18	31 (20.7)
Sport	0	0	7	0	28	35 (23.3)
Assault	0	0	0	1	0	1 (0.7)
Other	2	0	5	0	12	19 (12.7)
Total	38	11	14	2	85	150 (100)

Table 5. Sensitivity and specificity of billing codes that characterize PCS symptoms, assessed in the sample of 79 true positive PCS patients.

PCS symptoms after index TBI	N identified by manual abstraction	Sensitivity of billing codes	Specificity of billing codes
Anxiety	6	33.3%	98.6%
Apathy	0	NA	100.0%
Depression	8	37.5%	98.6%
Concentration difficulties	28	3.6%	96.1%
Dizziness	38	50.0%	97.6%
Emotional lability	7	0.0%	100.0%
Fatigue	13	15.4%	100.0%
Headache	65	30.8%	100.0%
Irritable	8	0.0%	100.0%
Impulsive	2	0.0%	100.0%
Memory difficulties	43	44.2%	100.0%
Sleep disturbances	12	8.3%	97.0%
Other	22	0.0%	98.2%

PCS denotes post-concussion syndrome and TBI, traumatic brain injury.

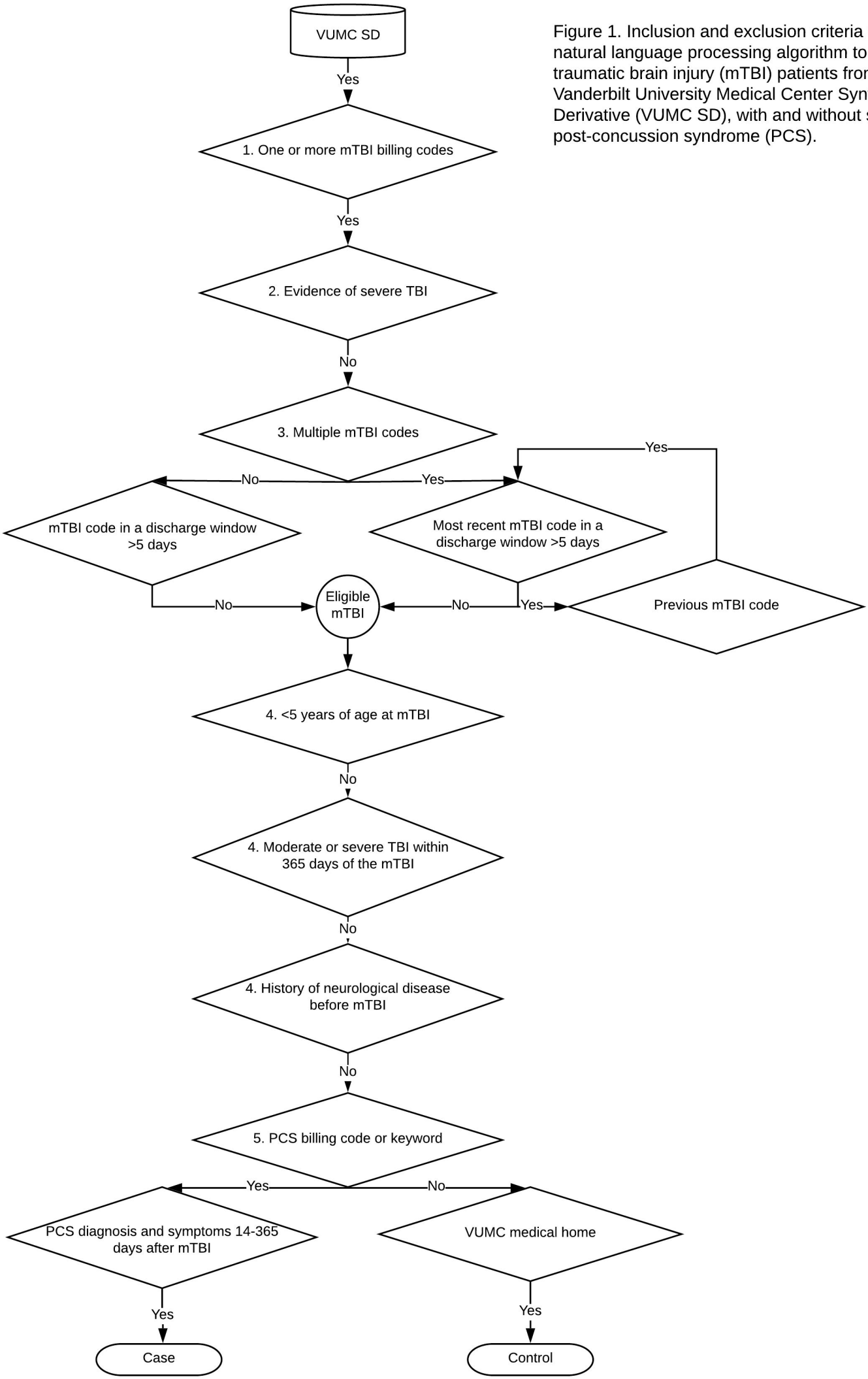

Figure 1. Inclusion and exclusion criteria used by the natural language processing algorithm to select mild traumatic brain injury (mTBI) patients from the Vanderbilt University Medical Center Synthetic Derivative (VUMC SD), with and without subsequent post-concussion syndrome (PCS).