HHS Public Access

Author manuscript

Intell Based Med. Author manuscript; available in PMC 2024 January 12.

Published in final edited form as:

Intell Based Med. 2023 ; 8: . doi:10.1016/j.ibmed.2023.100118.

Integrating unsupervised and supervised learning techniques to predict traumatic brain injury: A population-based study

Suvd Zulbayar[a,e], **Tatyana Mollayeva**[a,b,c,d], **Angela Colantonio**[a,b,c,d,e,f], **Vincy Chan**[b,c,d,e], **Michael Escobar**[a,*]

[a]Dalla Lana School of Public Health, University of Toronto, Toronto, ON M5T 3M7, Canada

[b]Rehabilitation Sciences Institute, Temerty Faculty of Medicine, University of Toronto, Toronto, ON M5G 1V7, Canada

[c]Acquired Brain Injury Research Lab, Department of Occupational Science and Occupational Therapy, University of Toronto, Toronto, ON M5G 1V7, Canada

[d]KITE Research Institute, Toronto Rehabilitation Institute-University Health Network, Toronto, ON M5G 2A2, Canada

[e]Institute of Health and Policy, Management and Evaluation, University of Toronto, M5T 3M6, Canada

[f]ICES, Toronto, ON, M4N 3M5, Canada

Abstract

This work aimed to identify pre-existing health conditions of patients with traumatic brain injury (TBI) and develop predictive models for the first TBI event and its external causes by employing a combination of unsupervised and supervised learning algorithms. We acquired up to five years of pre-injury diagnoses for 488,107 patients with TBI and 488,107 matched control patients who entered the emergency department or acute care hospitals between April 1st, 2002, and March

*Corresponding author. Dalla Lana School of Public Health, Health Sciences Building, 155 College Street, 6th Floor, Toronto, ON, M5T 3M7, Canada. m.escobar@utoronto.ca (M. Escobar).

Author contributions

A.C., T.M., V.C. and M.E. conceived the original concept and initiated the work. M.E. designed the statistical analysis for this work and S.Z. carried out and optimized the analyses with M.E.'s support. S.Z. wrote the manuscript and optimized reporting. A.C., T.M., M.E., contributed to the writing of the manuscript. All authors discussed the results biweekly and further steps for analyses and interpretation. All authors read the final version of the manuscript and commented on the text.

Ethical approval and informed consent

Approval: The study protocol was approved by the ethics committees at the clinical (Toronto Rehabilitation Institute-University Health Network) and academic (University of Toronto) institutions. Accordance: All methods were carried out in accordance with the relevant guidelines and regulations.

Informed consent: This research utilized encrypted administrative health data authorized under Section 45 of Ontario's Personal Health Information Protection Act. The data are housed at ICES, an independent, non-profit research institute, whose legal status under Ontario's health information privacy law allowed it to collect and analyze healthcare and patient characteristics data, without individual patient consent, for health system evaluation and improvement.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ibmed.2023.100118.

31st, 2020. Diagnoses were obtained from the Ontario Health Insurance Plan (OHIP) database which contains province-wide claims data by physicians in Ontario, Canada for inpatient and outpatient services. A screening process was conducted on the OHIP diagnostic codes to limit the subsequent analysis to codes that were predictive of TBI, which concluded that 314 codes were significantly associated with TBI. The Latent Dirichlet Allocation (LDA) model was applied to the diagnostic codes and generated an optimal number of 19 topics that concur with published literature but also suggest other unexplored areas. Estimated word-topic probabilities from the LDA model helped us detect pre-morbid conditions among patients with TBI by uncovering the underlying patterns of diagnoses, meanwhile estimated document-topic probabilities were utilized in variable creation as form of a dimension reduction. We created 19 topic scores for each patient in the cohort which were utilized along with socio-demographic factors for Random Forest binary classifier models. Test set performances evaluated using area under the receiver operating characteristic curve (AUC) were: TBI event (AUC = 0.85), external cause of injury: falls (AUC = 0.85), struck by/against (AUC = 0.83), cyclist collision (AUC = 0.76), motor vehicle collision (AUC = 0.83). Our analysis successfully demonstrated the feasibility of using machine learning to predict TBI due to various external causes and identified the most important factors that contribute to this prediction.

Keywords

Cause of injury; Diagnostic data; Latent Dirichlet allocation; Random forest; Topic modelling; Topic score

1. Introduction

Traumatic brain injury (TBI) is a leading cause of death and disability globally [1]. In Canada, approximately 22.6% of injury deaths between 2002 and 2016 were associated with a TBI diagnosis [2]. It is imperative that we exhaust all possible data sources and analytic tools available to inform prevention for TBI as the consequences of this injury can be various disorders and diseases [3,4] with a significant impact on morbidity and mortality [4,5]. In efforts to improve injury surveillance and guide primary prevention decisions, it is critical to identify risk factors that are important in the prediction of TBI event.

Previous quantitative studies utilizing machine learning (ML) techniques have shown promising results for predicting post-injury short-term and long-term outcomes [6-13] to prevent from adverse consequences, such as in-hospital morbidity and mortality, disability, need for supervision, and productivity outcomes [6,14,15]. However, their results were restricted to small clinical samples with limited generalizability across diverse cohorts of patients who sustained TBI through various mechanisms, representing one of the central barriers to the implementation of emerged algorithms into preventive medicine and public health initiatives. Our objective was to overcome this critical issue by utilizing administrative healthcare data from population-based sources in a publicly funded healthcare system. In TBI research, the implementation of ML algorithms on large scale high-quality datasets would improve robustness and enable low-cost identification of complex associations between pre-morbid conditions and TBI. Indeed, recent studies developed data

mining algorithms [16,17] to detect health status markers associated with TBI [5] and its severity [18]. However, there still exists a gap in knowledge to develop prediction models for TBI sustained due to various external causes to prevent from the injury event itself.

Predicting the occurrence of TBI is a great challenge due to the spontaneous nature of the injury and the difficulty in recognising pre-injury risk factors. It is evident that certain behavioral patterns and external hazards increase the susceptibility of someone to a head injury, such as, taking part in sports-related activities, drug and alcohol usage, self-harming, and exposure to violence [5,19-21]. These exposures are often hard to measure or not reported prior to injury, making it difficult to gather retrospective data. Research findings suggest that many disorders and diseases are significant risk factors of TBI and are often reflected in the cause of injury and severity [5,17]. Hence, it is critical to design primary prevention strategies [22] that capture the complexity of this injury on an individual level by considering all necessary clinical, physiological, social, demographic, and lifestyle factors.

To fill the gap in knowledge, our study proposes to predict TBI by exploiting the results of a topic modelling approach, namely the Latent Dirichlet Allocation (LDA) model. In addition to well-known risk factors (age, sex, place of residence, income) of TBI [5], our study looked at diagnostic history of patients with TBI during their pre-injury period of up to five years, where the diagnoses were made by physicians for inpatient, outpatient, and long-term care services. This two-step modelling approach would allow us to directly identify pre-existent health conditions that increase the risk of TBI and recognize diagnostic patterns that may reveal habits and behavior that are common amongst TBI patients.

The LDA model is a generative probabilistic model designed to extract topics from collections of discrete data and is mainly adopted as a flexible approach for unsupervised classification of a text corpora [23]. In the healthcare field, many large and unstructured textual datasets utilize this approach to mine topics, including, but not limited to, clinical notes by nurses and physicians [24,25], and survey data about patient feedbacks and experiences [26-31]. Additionally, LDA was successfully applied to International Classification of Disease (ICD) based diagnostic codes to explore associations amongst diagnostic groups [32], discover patterns of care [33], and facilitate information retrieval for clinicians [34]. However, the popularity of LDA comes from its wide range of applications, such as being frequently used as a dimensionality reduction technique to reduce computational costs for predictive modelling [23,35,36]. Since we have a high number of attributes in the diagnostic data, LDA model can be utilized to create new variables on a reduced feature space for predictive modelling. The fitted LDA model also offers an additional insight into common pre-morbid conditions of TBI patients that would assist in designing primary prevention frameworks and improving injury surveillance by primary care providers and policymakers.

In the context of TBI, the underlying cause (mechanism) of injury emphasizes the complex nature of this disease process [3], hence, identifying main drivers of each mechanism may be helpful in informing preventive decisions for those who are more susceptible to certain external cause(s) of injury. Therefore, the main goal of this study is to build ML algorithms to predict the initial TBI event as well as its four leading causes of injury (e.g., falls,

struck by/against, cyclist collision, motor vehicle collision [18]) using pre-injury physician diagnoses and socio-demographic information.

2. Materials and methods

2.1. Data sources

We accessed the ICES (formerly Institute of Clinical Evaluative Sciences) [37] data repository to collect health insurance claims data for all residents of Ontario, Canada who received physician services between the fiscal years of 2002 and 2020 from the Ontario Health Insurance Plan (OHIP) database. The ICES data repository stores population-wide de-identified, linkable health data sets for all publicly funded services provided to the residents of Ontario. As such, we were able to link patient-level data between the OHIP database with the records for emergency department (ED) visits, obtained from National Ambulatory Care Reporting System (NACRS), and acute care visits, obtained from Discharge Abstract Database (DAD).

The NACRS and DAD datasets contained primary and secondary diagnoses recorded using International Classification of Diseases, Tenth Revision, Canada (ICD-10-CA) codes [38] in addition to various clinical, demographic, and socio-economic variables. We identified 488,107 patients with TBI, i.e., cases, who were discharged from their first TBI-related hospitalization in the ED or acute care (defined as “TBI event”) between April 1st, 2002, and March 31st, 2020 and diagnosed with a TBI-specific code (ICD-10-CA codes: S020, S021, S023, S027, S028, S029, S040, S071, S06) [5]. For patients with TBI (referred to as TBI patients), their cause(s) of injury [39] were grouped into five categories according to their ICD-10-CA external cause of injury codes: falls, struck by/against, cyclist collision, motor vehicle collision, and other (Supplementary Table 1 for the full list of ICD-10-CA codes). From the NACRS and DAD datasets, we also gathered information on *age at injury* (continuous), *sex* (binary), *rurality indicator* (binary), and *income quintile* (categorical), and individually matched the TBI patients to one control patient based on these socio-demographic factors as well as fiscal year of admission to the ED or acute care. Control patients were identified from those without a TBI-related visit to the ED or acute care between April 1st, 2002, and March 31st, 2020. Patients were aged from 0 to 95 years old at the time of their injury.

The OHIP database contains billing information from primary care physicians in Ontario, including *date of service, diagnostic code, age at service,* and other fee-related information. For billing purposes, there is only one diagnosis assigned per row, so one physician service could result in more than one observation in the dataset. It was found that 30-days preceding the TBI event (i.e., TBI index date) is part of the TBI event window [17], hence, our study included OHIP diagnoses made between 5 years and 30 days to30days before the first TBI event. This window was defined as “pre-injury phase”. Since control patients have no TBI-related visit, we used the admission date of a visit to the ED or acute care that happened closest to their index date (index date for control patients does not correspond to an event) as a “pseudo event” date to determine their pre-injury phase. For each patient, diagnoses were filtered according to the OHIP *date of service* and all OHIP diagnostic code(s) assigned during the pre-injury phase were collected. We then selected the OHIP diagnostic codes

according to the OHIP claims data dictionary provided by ICES, where diagnostic codes are either in a three-digit numeric format or alphanumeric, starting with the alphabetic letter "C" followed by two digits [40]. We only included diagnostic codes that were in the data dictionary and removed diagnostic codes "999 = Undiagnosed" and "100 = Monitored".

Our cohort consisted of 488,107 matched pairs who were hospitalized in the ED/acute care where some patients have no history of a publicly funded physician service made during their pre-injury phase, hence, not every patient has OHIP diagnosis. After applying the inclusion criteria, there were a total of 23,563,833 OHIP diagnoses for 468,313 TBI patients, and the number of distinct diagnostic codes per patient ranged from 1 to 95. For controls, there were 17,439,627 OHIP diagnoses for 459,374 patients, with a range of 1 to80 distinct diagnostic codes claimed per patient (see Supplementary data file for the list of unique diagnostic codes and their prevalence in cases/controls).

The entire cohort was randomly split into 244,064 matched pairs for training, 122,018 matched pairs for validation, and 122,025 matched pairs for testing (Table 1 for summary statistics). The number of patients without any OHIP diagnoses were: 9,899 cases and 14,286 controls in the training set; 4,881 cases and 7,266 controls in the validation set; 5,014 cases and 7,181 controls in the test set.

2.2. Proposed approach

2.2.1. Diagnostic code selection—We began with a diagnostic code screening process on the training set to limit the subsequent analysis to OHIP codes that were associated with TBI event and to ensure that the diagnostic codes are likely to have some "signal" when grouping TBI patients using LDA. We only assessed the presence of a diagnostic code in a patient's diagnostic history rather than the frequency. McNemar's test for matched pair data [41] was conducted for each unique diagnostic code in the training set to detect diagnostic codes that were significantly associated with TBI event, and if the number of discordant pairs were less than 25, Fisher's exact test was used instead [42]. Benjamini-Yekutieli (BY) correction was used to control for the false discovery rate (FDR) [16,43]. This was followed by a calculation of the Phi-coefficient for binary data [44] to calculate the correlation between the significant codes (resulting from McNemar's test with BY adjustment) and TBI event, where any diagnostic codes that had negative or zero correlation were eliminated from further analysis, and diagnostic codes that were positively correlated with TBI event were retained as the final list of "selected diagnostic codes".

2.2.2. Topic modelling with LDA—We employed a two-step ML approach to develop predictive models for TBI and its causes of injury by first creating new features using LDA. The LDA method can be utilized as a dimensionality reduction technique on our highly sparse and high dimensional OHIP diagnoses dataset by collapsing together terms that have the same semantics [35], thus enabling us to create new variables called "topic scores". Hence, some concepts have a slightly different meaning in our context:

- a *word* or a *term* is equivalent to a *diagnostic code*,
- a *document* is equivalent to a *sequence of diagnostic code(s) for a patient*

- a *corpus* is a *collection of patients' sequence of diagnostic code(s)*
- a *topic* is a *distribution of diagnostic codes*
- a *vocabulary* is a *list of distinct diagnostic codes in a corpus* (bounded by the selected diagnostic codes)

We used R packages "topicmodels" [45] and "lda" [46] for topic modelling with LDA, and "ldatuning" package for tuning the number of topics [47] on the TBI-training set. Each split set of cases and controls makes a "corpus", however, only the TBI-training set corpus was used to train the LDA model, where the resulting fitted model was used to estimate parameters of the other five corpora, i.e., TBI-validation, TBI-test, control-training, control-validation, and control-test. After subsetting to the selected diagnostic codes as per Section 2.2.1, each corpus was expressed as document-term matrices (DTM) based on the number of occurrences of diagnoses. Therefore, the dimension of DTM is equals to the number of patients in the corpus by the number of unique diagnostic codes in the corpus. The entries in the matrix are the number of times a patient has been diagnosed with the corresponding diagnostic code during the pre-injury phase.

The number of topics have to be set a priori to training the LDA model. This is the only hyperparameter that needs to be tuned for the LDA model. The TBI-training set was used to perform a grid-search over 2 to 50 topics by optimizing four metrics; Griffiths2004, Deveaud2014, Arun2010, and CaoJuan2009 [48-51] using the Variational Exception Maximization (VEM) algorithm [52] for estimation. The most parsimonious number of topics, K, would have to maximize the first two metrics while minimizing the last two metrics. The LDA model was then trained on the same corpus using the optimal number of topics, K, and VEM algorithm to infer topics.

The fitted model was then utilized in two ways:

1. Detect pre-morbidities: The model estimates the probability of a word given topic denoted by posterior beta (β), which is the probability of each diagnostic code belonging to a specific group.

$$\beta_{i,k} = P(word_i \mid topic_k) = P(diagnosis\ code_i \mid group_K)$$
$$where\ i = 1, \ldots, length\ of\ vocabulary\ in\ a\ corpus\ and$$
$$k = 1, \ldots, K$$

This can be utilized as an unsupervised learning approach to group and characterize TBI patients according to their OHIP diagnoses made during the pre-injury phase and detect any pre-existent health conditions of TBI patients and common co-occurring diagnoses.

2. Topic score creation: The model also estimates the probability of a topic given document denoted by posterior gamma (Γ), which is the probability of a patient belonging to each group.

$$\Gamma_{k,j} = number\ of\ diagnoses_j \times P(group_k \mid patient_j)$$
$$where\ k = 1, \ldots, K\ and$$
$$j = 1, \ldots, number\ of\ patients\ in\ a\ corpus$$

This probability multiplied by each patient's total number of OHIP diagnoses made during the five-year pre-injury period results in weighted topic scores which can then be used as features for predictive modelling.

$$Topic\ Score_{k,j} = document\ length_j \times \Gamma_{kj}$$
$$where\ k = 1, \ldots, K\ and$$
$$j = 1, \ldots, number\ of\ patients\ in\ a\ corpus \quad (1)$$

The fitted LDA model can be used to obtain the posterior gammas of the TBI-training corpus, as well as infer topic distributions of unseen documents, which would allow us to estimate gammas for the other five corpora. Patients with no OHIP diagnosis during the pre-injury phase or those who were eliminated due to the diagnostic screening process were assigned a topic score of zero (TS= 0) for all topics.

Number of visits made to a physician is associated with the health of a patient, hence, the weighted document-topic probability score, i.e., topic score, reflects the patient's health status. The number of diagnoses has a large variability across patient populations as some patients experience more recurrent visits to a physician's office due to chronic health problems, making the topic scores highly skewed to the right. We performed rank-normalization on the topic scores within each group of the split sets using "RankNorm" function which applies rank-based inverse normal transformation [53].

2.2.3. Predicting TBI event and causes of injury—The primary prediction task of our study is to develop a classification model for TBI event by utilizing the normalized topic scores (created in Section 2.2.2) along with the patient socio-demographics (*age at injury* as a continuous variable, binary *sex* (male, female), *rurality indicator* (yes, no), and *income quintile* (levels from lowest to highest: 1, 2, 3, 4, 5)) as candidate predictors. The secondary task includes developing individual binary classifier models for each of the four main external causes of injury; falls, struck by/against, cyclist collisions, and motor vehicle collisions. External causes were modelled individually as they are not mutually exclusive events (i.e., TBI could be sustained due to one or more cause(s) of injury occurring at the same time). We considered the same candidate predictors as TBI prediction, however, for external cause prediction, we limited the analysis to the patients that had the cause of injury of interest and their matched controls, unlike TBI prediction where we considered the entire cohort. Table 1 shows the number of patients who sustained TBI due to each cause of injury, i.e., the number of cases for each cause of injury prediction task.

Random forest is an ensemble ML algorithm that consists of many decorrelated decision trees and has a very powerful predictive performance due to the process of bagging and feature randomness that are involved in building each tree [54]. This algorithm is widely

used as a variable selection method as it naturally identifies informative features with the most predictive power by calculating the importance of a feature [55]. There are several measures for variable importance (VIMP) calculation, we've used "anti-VIMP" measure offered by the "random-ForestSRC" package in R [56]. It simply calculates the effect of each variable by comparing the out-of-bag error that results from assigning the variable of interest to the opposite node whenever there is a node split on that variable. This would allow us to see the main drivers of TBI as well as each external cause of injury.

For each prediction task, the training set was used to train the RF model, and the validation set was used to tune the hyperparameters and select the best model, whereas the test set was used to assess the performance of the selected model. There are several hyperparameters in RF that need to be optimized, including the number of random splits used for splitting, the number of trees to grow, minimum terminal node size, and the number of features randomly sampled as candidates at each split. We tuned the number of trees starting from 100 and iteratively increased by 100 trees, until there was no further improvement in the validation set area under the receiver operating characteristic curve (AUC). For a given number of trees, the terminal node size and the number of randomly sampled features were optimized by minimizing the out-of-bag error. Number of random splits used for splitting was kept at 1 by default. Test set performance was assessed for the selected model using AUC.

3. Results

3.1. Diagnostic code selection

There were 578 distinct diagnostic codes discovered in the TBI-training set, and 575 for control-training set, with a total of 580 unique codes in the combined training set. After conducting McNemar's test on each of the 580 distinct codes, 397 codes were found significant with *P, which reduced to 353 after controlling for the FDR. Furthermore, a total of 39 diagnostic codes were screened out after the calculation of Phi-coefficient, in which 12 of them had a negative correlation with TBI status and 27 had a zero correlation. Consequently, 314 OHIP diagnostic codes (Supplementary Table 2 for odds ratios) made it to the final list of selected diagnostic codes.

3.2. Topic modelling with LDA

After subsetting to the selected diagnostic codes in Section 3.1, TBI-training set had 9,786,264 observations for 231,594 patients, which was used to tune the number of topics for the LDA model and train the LDA model using that optimal number of topics. As seen in Fig. 1, a grid search performed over 2 to 50 topics showed that Griffiths2004 (maximize) metric failed to converge, CaoJuan2009 (minimize) reaches a plateau after 19 topics, and Arun2010 (minimize) metric is reasonably small at k = 19, meanwhile Deveaud2014 (maximize) has a local maxima at 19 topics and a global maxima at 48 topics. A manual inspection of high-ranking terms in each of the 48 topics revealed multiple topics with similar themes (i.e., same term(s) representing several topics) and any additional topic over 19 topics did not provide us with a distinct clinically relevant theme. For our study purposes, we wanted to ensure diversity across topics to assist clinicians set up distinct surveillance measures for each risk factor. Therefore, through the recommendation of clinical experts,

19 topics were preferred over 48 topics as it provides us with the most diverse range of topics with an added benefit of reduced computational costs for the subsequent predictive modelling. In addition, smaller number of groups would allow us to distinguish diagnoses that are most likely to co-occur in a topic and yield more insightful results for inferential purposes.

To get a better understanding of the pre-morbid conditions that often co-occur together in TBI patients, we selected the top terms for each topic using the following criteria: if there is a term with $\beta > 0.8$, only retain that term; else keep the highest-ranking terms until the sum of β reach 0.8 or 15 terms are selected. This criterion was applied strictly for exploratory purposes only and to see the diagnostic patterns within each topic as well as to gather representative diagnoses of each topic. As a result, 132 terms were selected for 19 topics, with 119 unique terms. Table 2 shows topic names assigned for each group based on the diagnoses that are most likely to belong to the topic of interest, see Supplementary Fig. 1 and Supplementary Table 3 for the list of top diagnostic codes and their corresponding probabilities. Some topics, such as, Topic 17 has a dominant term with $\beta = 0.9888$ corresponding to "304 = Drug dependence, drug addiction", which means that TBI patients with this pre-injury diagnosis most likely don't have any other pre-existent health conditions. On the other hand, some topics are composed of multiple diagnoses with small word-topic probabilities, for instance, Topic 1 is distributed over many diagnoses with $\beta < 0.2$ that correspond to various conditions related to heart disease and these diagnoses tend to co-occur in a patient during the five years preceding TBI event.

3.3. Predicting TBI event and causes of injury

Table 1 provides the following information for each split set: the number of TBI patients, and the number of patients who sustained TBI due to each external cause of injury, i.e., the number of cases for each prediction task. Each split set was composed of matched pairs only.

As per Table 4, the best model for TBI event prediction had 300 trees, terminal node size of 1, and 22 variables were randomly sampled at each node split. This trained model yielded an AUC of 0.908 on the validation set. Test set performance was AUC = 0.8474 for 122,025 TBI patients and 122,025 without TBI (Supplementary Tables 4 and 5 for other performance metrics and confusion matrices calculated using a threshold of 0.5 for all prediction tasks). According to Table 5, the top 5 variables for TBI prediction were, listed from most important to least important, "*Topic 17* – Drug dependence and addiction", "*Topic 14* – Newborn/infant care and ill-defined infections", "*Topic 5* – Injuries to extremities, disorders of musculoskeletal system, overexertion", "*Topic 4* – Kidney failure, sepsis, anemia", and "*Topic 2* – Microbia and allergic skin, eye, and upper airway manifestations". See Table 5 for the full list of VIMP measurements for each predictor.

According to the highest validation AUC of 0.8988 in predicting TBI due to falls, the optimal hyperparameters were 400 trees, node size of 1, and 22 randomly sampled features. The best model was able to discriminate with AUC = 0.8536 on the test set with 54,703 matched pairs. The most important predictors were "*Topic 14* – Newborn/infant care and

ill-defined infections", "*Topic 17* – Drug dependence and addiction", "*Topic 5* – Injuries to extremities, disorders of musculoskeletal system, overexertion", "*Topic 2* – Microbial and allergic skin, eye, and upper airway manifestations", and "*Topic 4* – Kidney failure, sepsis, anemia".

As for predicting TBI due to being struck by/against, the optimal hyperparameters were 200 trees, terminal node size of 1, and 6 randomly sampled variables, which resulted in an AUC of 0.8969 on the validation set. Test set performance was AUC = 0.8269 for 42,012 TBI patients who got struck and 42,012 matched controls. The strongest predictors were "*Topic 17* – Drug dependence and addiction", "*Topic 5* – Injuries to extremities, disorders of musculoskeletal system, overexertion", "*Topic 14* – Newborn/infant care and ill-defined infections", "*Topic 4* – Kidney failure, sepsis, anemia", and "*Topic 2* – Microbial and allergic skin, eye, and upper airway manifestations".

Hyperparameters for predicting TBI due to cyclist collision were tuned to be 400 trees, terminal node size of 2, and 22 randomly sampled features. The validation set AUC of this model was 0.8664. Test set performance was AUC = 0.7608 for 3,046 matched pairs. The most important features were "*Topic 17* – Drug dependence and addiction", *age at injury*, "*Topic 14* – Newborn/infant care and ill-defined infections", "*Topic 4* – Kidney failure, sepsis, anemia", and "*Topic 5* – Injuries to extremities, disorders of musculoskeletal system, overexertion".

The selected model for predicting TBI due to motor vehicle collision had a validation AUC of 0.8997 for a model trained with terminal node size of 1 and 18 randomly sampled variables at each node split for 600 trees. The final model performance on the test set was evaluated to be AUC = 0.8312 for 13,137 cases who had motor vehicle collision and 13,137 matched controls. The top 5 variables were "*Topic 17* – Drug dependence and addiction", "*Topic 14* – Newborn/infant care and ill-defined infections", "*Topic 4* – Kidney failure, sepsis, anemia", "*Topic 5* – Injuries to extremities disorders of musculoskeletal system, overexertion", and *age at injury*.

4. Discussion

In this study, we aimed to facilitate the development of TBI event prediction by incorporating unsupervised and supervised ML algorithms on a province-wide population-based dataset and explored the patterns of pre-injury physician diagnoses of TBI patients using LDA. The study has demonstrated how administrative healthcare data could be used to predict the occurrence of TBI, including its external causes of injury. To the best of our knowledge, this is the largest study in predicting TBI event at the population level in terms of the number of patients in the study, the number of attributes related to pre-morbidities and social circumstances, and the length of period covered in the study.

Stemming from this ML research, and in agreement with earlier TBI studies [5,17-19,21,57,58] three top strategies to the planning of preventive programmes, across external causes of injury, are to focus on drug dependence and addiction, musculoskeletal injuries and overexertion, and ill-defined conditions [59]. These roughly correspond to

strategies that would use, respectively: (i) direct communication to influence the knowledge, attitudes and beliefs of the individual persons, community, and population as a whole concerning drug use/dependence-brain injury relationship; (ii) ergonomic, engineering, physical rehabilitation and legal developments that would enable risk reduction associated with musculoskeletal injuries and overexertion; and (iii) public-private resource and coalition developments through policy, social organizations and other agencies that care about growing concern of the ill-defined conditions and declining population health, to reinforce and reward for health-directed lifestyle. Since OHIP diagnoses are associated with services made by primary care physicians, they are often less serious than diagnoses made during hospitalization. This enabled us to capture risk factors that were previously not recognized [5, 17], such as symptoms associated with cold and allergic reactions, kidney and renal failure, newborn/infant care and conditions. On the other hand, some health conditions that are known to be associated with TBI and certain cause(s) of injury were not discovered in our study such as Alzheimer's disease and related dementia, which are recognized as significantly linked to TBI and falls [5,17,60]. However, many of our topics encompass diagnostics codes that tend to become more prevalent with age and are associated with Alzheimer's diseases and other forms of dementia [17], e.g., Topic 1 - Heart disorders and symptoms and Topic 3 - Stroke and excessive involuntary movement disorders (i.e., conditions like hypertension, coronary artery disease, heart failure, and stroke become more common with age, and for people with dementia).

The importance of "Topic 2 – Microbia and allergic skin, eye, and upper airway manifestations" in predicting TBI and its external causes of injury is a novel finding in literature. We have two hypotheses that require further study: (i) infections and allergies, and upper airway manifestations may increase the person's susceptibility to brain injury through altered immune system; and (ii) side effects caused by medications that are used to treat these symptoms and diagnoses may also increase their vulnerability to brain injury. To elaborate, infections and allergic reactions can trigger inflammatory responses in the body which further impact the central nervous system therefore influencing the severity of symptoms even due to a minor force to the head/body, leading to a concussion or a mild TBI. It is also possible that the medications used to treat infections or allergen exposures/upper airway manifestations, e.g., corticosteroids, antihistamines, may cause side effects including muscle weakness, sleep disturbances, drowsiness/blurred vision, and low blood pressure, putting the person at risk of brain injury due to falls and motor-vehicle collisions. Table 5 also shows that age is one of the most important variables in predicting TBI and its causes. Our results are consistent with prior research [61-64] on the topic, where the nature of age relationship observed here can be attributed to a combination of biological, behavioral, and multimorbidity factors. Babies and children have disproportionally larger heads as compared to the body and weaker muscles of the neck, making them more susceptible to brain injury caused by falls or assault. In addition, their protective mechanisms of the developing brain are not fully developed and in an event of an injury this can lead to more severe presentations, resulting in a greater likelihood of seeking care. In the older age, brain atrophy causes less cushioning between it and the skull, making it more susceptible to injury from falls or other impacts. This can also lead to more persistent and severe symptoms, and therefore care-seeking. Likewise, vascular wall

becomes less elastic and more prone to rupture in an impact injury to the head, causing intracranial hemorrhage and secondary TBI. Older people have number of chronic medical conditions, such as, cardiovascular diseases, diabetes, and liver disorders, making them more susceptible to TBI due to falls. Hence, younger, and older people (due to changes in cognition/dementia) may be less aware of potential dangers and may engage in activities that put them at risk for TBI. Another finding was that "Topic 14 – Newborn/infant care and ill-defined infections" was one of the leading predictors in Table 5. In addition to aforementioned factors that increase the risk of TBI in newborns and children, while our study does not concern cause-effect, the following ideas prompt further hypotheses on the relevance of Topic 14. Newborns/infants and children with ill-defined infections might have compromised immune system and weaker coordination, making them more vulnerable to a concussion caused by minor acceleration-deceleration forces applied to the body (even in a gentle rocking, swaying). If a child has an ill-defined infection at the time of a concussion, determining the primary cause of their symptoms might be challenging because the symptoms of infection and concussion can overlap (e.g., fever, weakness, disturbed sleep, changes in behavior). Further, infants or young children who may have conditions that result in prolonged crying for instance may be at risk of abusive head trauma [65]. The most common cause of severe TBI in children less than two years of age is abusive head trauma [66].

Compared to previous TBI-related prediction studies [7,67], we were able to achieve a significantly higher prediction performance of AUC = 0.85 for TBI event prediction due to RF's ability to explain complex relationships within a large set of attributes by capturing non-linear and higher order interaction effects of features [68]. OHIP dataset fits this criterion as it is composed of a highly sparse, noisy, and interrelated features. However, RF is often referred to as a "black box" model as its predictive power comes at the expense of interpretability. Unlike regression models, it is difficult to replicate the results of this model and make direct inferences, such as calculating odds ratio (OR) or estimated coefficients. In Supplementary Table 7, we calculated the matched ORs for all topics by converting the untransformed topic scores to binary topic score: $BTS_i = 1$ *if* $TS_i > 1$ *for* $i = 1,\ldots, 19$, which concluded that all topics are significant with OR > 1. To prevent from overfitting, we used validation sets to tune the hyperparameters, however, overfitting can still be observed in all our selected RF models. It is more evident in models trained with smaller datasets, such as, prediction model for TBI due to cyclist collision was trained on 5956 matched pairs and yielded a validation AUC of 0.8664 and test AUC was dropped to 0.7608. Future work may focus on using cross-validation technique to optimize the hyperparameters or introduce more data to avoid overfitting.

The effectiveness of LDA model in topic modelling of textual data has been proven in various languages, yet the application of this method on other types of discrete data have not been extensively studied. Even though diagnostic codes have no direct meaning like textual data, the model was able to group TBI patients successfully as underlying health conditions and lifestyle habits are discovered within the selected terms for each topic. For instance, diagnoses that are most associated with "Topic 5 – Injuries to extremities, disorders of musculoskeletal system, overexertion" are common sports-related injuries that athletes or active people are more susceptible to. Since Topic 5 is relatively uniformly distributed,

consisting of many diagnostic codes with low β, the diagnoses are likely to co-occur in a patient's diagnostic history with a small frequency, revealing a behavioral pattern of athletes. Meanwhile, there are only two diagnostic codes selected for "Topic 15 -Unspecified anxiety, neurotic, and associated states" where the dominant term ($\beta = 0.73$) corresponds to mental health disorders and the other term is its associated symptoms. "Topic 14 – Newborn/infant care and ill-defined infections" is composed of multiple diagnoses typically made for children which can be confirmed as the patients with the topic ($TS_{14} > 1$) are young children with 25th quantile = 2, Median = 5, 75th quantile = 11 years old at injury (Supplementary Tables 6a-c). The model is able to capture such high-level information because LDA is a "bag-of-words" model, meaning it ignores the order of the terms or words and tends to assign words that co-occur together in a document a higher chance of belonging to the same topic [69]. In our case, the order and the time of patient diagnoses are ignored making the dataset simpler to analyze and explore without the issue of missing data. For this reason, LDA has been previously applied as a data mining approach to ICD codes for exploratory purposes [32-34], but the results were never utilized for predictive modelling. Text data based LDA and RF combined approach was proven to yield powerful predictions [70-72], yet our study was the first to show that diagnostic code based LDA model results can be exploited for predictive modelling with RF. We applied this approach to reduce the features from 314 binary diagnostic codes to 19 continuous rank-normalized topic scores, which gave us computational benefits with a minimal loss of information. However, our models were very sensitive to patients who were not diagnosed with any of the selected diagnostic codes or with no diagnostic history, as misclassification and overfitting were observed due to the manual assignment of $TS = 0$. For example, the selected model for TBI prediction task misclassified 38,944 patients in the test set, among them was all the patients who were assigned $TS = 0$ (6,246 TBI patients and 9,588 controls). It is naturally challenging to predict outcomes for people with no previous history as the model did not have enough examples to learn from, limiting the application of our method to patients with the selected diagnostic codes for better prediction results.

Although, the population-based dataset we used is significantly larger than all other TBI-related ML studies, several considerations need to be taken when applying our results to different populations. To begin with, we need to address the potential risk of bias that was introduced due to cohort selection. Since our cohort was composed of patients who were hospitalized in ED or acute care, they may have more serious underlying health problems than the general population. In addition, there is an extreme underrepresentation of patients without TBI in the dataset as our sample prevalence of TBI is much higher than the population prevalence [73] due to 1:1 matching. However, even if we were to have the same population prevalence (0.5% in Canada) [73] in the sample, it is challenging to deploy ML classification models with a reliable strong accuracy when there is such an extreme class imbalance because the classifiers often tend to be biased towards the majority class [74,75]. Nonetheless, we were able to identify pre-existing health conditions and lifestyle factors that distinguish patients with TBI from their controls. Previous non population-based studies indicated that low socioeconomic status, male sex, and place of residence were risk factors of TBI and significant predictors for TBI outcome [76-79], which do not agree with our results as *sex, income quintile,* and *rurality indicator* were

the least important variables for all prediction tasks, this is because we matched cases and controls based on these variables. Secondly, the accuracy and the quality of codes assigned is overall questionable as we eliminated a total of 1,351,089 observations from cases, and 1,442,646 from controls due to typos in their OHIP diagnostic codes as per ICES data dictionary [40]. Since Electronic Health Records (EHR) are often misreported, it is not feasible to determine which codes are inaccurate. Instances of inaccurate code assignments can be seen in “916 = Well baby care” where 20,854 out of 528,000 OHIP diagnoses with this code were for patients with age at OHIP service date of over two years old. Thirdly, we might have overlooked important OHIP diagnoses due to the screening process or absence in the diagnostic history of patients in the training set, as well as many known clinically relevant predictors were not available in our dataset [80-88], such as race/ethnicity, education, and occupation; predictive performance might be improved with the introduction of these variables. Lastly, one could further explore other common causes of injury, such as assaults, firearms, which were grouped under “other” category and were not individually predicted due to lack of representation in our cohort. We also did not investigate the severity of injury and its relationship with pre-injury health conditions due to missingness (unknown for 45.7% of TBI patients).

5. Conclusion

Our study results demonstrate the feasibility of using LDA model on diagnostic data to predict TBI event as well as its external causes. Similar approaches can be replicated in other public health research to retrieve information from diagnostic codes and to gain computational benefits for predictive modelling. An introduction of meaningful clinical and physiological parameters, additional observations from a longer period, and a larger cohort (both cases and controls) may improve the performance and the robustness of our proposed approach. Future studies could take advantage of the longitudinal nature of the dataset and integrate the time and the order of diagnoses to boost prediction performance, however, this needs to be carefully studied as there would be a problem of missing data. We could do so by employing Recurrent Neural Networks for predictive modelling or utilize Latent Semantic Analysis [89] to discover associations among diagnoses based on closeness since physician diagnoses that happen close in time may be connected to each other.

Since this research was conducted on a retrospective data, the results need to be applied with caution due to ever-changing circumstances. The world has been greatly affected by the novel coronavirus (COVID-19) pandemic since 2019, and the virus’ and the vaccines’ long-term health impacts in relation to TBI are largely unknown. In addition, acceleration in the development of auto-pilot cars may eventually prevent from all road traffic injuries, including TBI. An improved access to information and better education in the society might reduce incidences of assault-related TBI.

The study results show that prevention of initial TBI could be possible by early detecting pre-morbid conditions that increases the risk of TBI and targeting individuals based on their lifestyle habits, age, and other socio-demographic factors to improve precision medicine. Furthermore, safety interventions can be put in place for an individual based on their behavior that potentially exposes them to certain types of external causes of injury.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The research reported in this publication was supported by the National Institute of Neurological Disorders and Stroke of the National Institutes of Health under Award Number R01NS117921. The content is solely the authors' responsibility and does not necessarily represent the official views of the National Institutes of Health. During the work under the study researchers T.M. and A.C. were supported by Canada Research Chair programs (CRC-2021-00074, CRC-2019-00019).

This study contracted ICES Data & Analytic Services (DAS) and used de-identified data from the ICES Data Repository, which is managed by ICES with support from its funders and partners: Canada's Strategy for Patient-Oriented Research (SPOR), the Ontario SPOR Support Unit, the Canadian Institutes of Health Research, and the Government of Ontario. The opinions, results and conclusions reported are those of the authors. No endorsement by ICES or any of its funders or partners is intended or should be inferred. Parts of this material are based on data and information compiled and provided by CIHI. However, the analyses, conclusions, opinions, and statements expressed herein are those of the author, and not necessarily those of CIHI.

Data statement

ICES is an independent, non-profit research institute funded by an annual grant from the Ontario Ministry of Health and Long-Term Care (MOHLTC). As a prescribed entity under Ontario's privacy legislation, ICES is authorized to collect and use health care data for the purposes of health system analysis, evaluation, and decision support. Secure access to these data is governed by policies and procedures that are approved by the Information and Privacy Commissioner of Ontario. The dataset from this study is held securely in coded form at the Institute for Clinical Evaluative Sciences (ICES). While data sharing agreements prohibit ICES from making the dataset publicly available, access may be granted to those who meet pre-specified criteria for confidential access, available at www.ices.on.ca/DAS. The full dataset creation plan and underlying analytic code are available from the authors upon request, understanding that the computer programs may rely upon coding templates or macros that are unique to ICES and are therefore either inaccessible or may require modification.

References

[1]. Johnson WD, Griswold DP. Traumatic brain injury: a global challenge. Lancet Neurol 2017;16:949–50. 10.1016/S1474-4422(17)30362-9 [PubMed: 29122521]

[2]. Release notice - injury in review, 2020 edition: spotlight on traumatic brain injuries across the life course. Health Promot Chronic Dis Prev Can 2020;40. 10.24095/hpcdp.40.9.05. 294–294. [PubMed: 32909940]

[3]. Bramlett HM, Dietrich WD. Long-term consequences of traumatic brain injury: current status of potential mechanisms of injury and neurological outcomes. J Neurotrauma 2015;32:1834–48. 10.1089/neu.2014.3352 [PubMed: 25158206]

[4]. Masel BE, DeWitt DS. Traumatic brain injury: a disease process, not an event. J Neurotrauma 2010;27:1529–40. 10.1089/neu.2010.1358. [PubMed: 20504161]

[5]. Mollayeva T, Sutton M, Chan V, Colantonio A, Sayantee J, Escobar M. Data mining to understand health status preceding traumatic brain injury. Sci Rep 2019. 10.1038/s41598-019-41916-5.

[6]. Matzuo K, Aihara H, Nakai T, Morishita A, Tohma Y, Eiji K. Machine learning to predict in-hospital morbidity and mortality after traumatic brain injury. J Neurotrauma 2020;37:202–10. 10.1089/neu.2018.6276. [PubMed: 31359814]

[7]. Gravesteijn BY, Nieboer D, Ercole A, Lingsma HF, Nelson D, van Calster B, et al. Machine learning algorithms performed no better than regression models for prognostication in traumatic brain injury. J Clin Epidemiol 2020;122:95–107. 10.1016/j.jclinepi.2020.03.005. [PubMed: 32201256]

[8]. Hernandes Rocha TA, Elahi C, Cristina da Silva N, Sakita FM, Fuller A, Mmbaga BT, et al. A traumatic brain injury prognostic model to support in-hospital triage in a low-income country: a machine learning–based approach. J Neurosurg 2020;132:1961–9. 10.3171/2019.2.JNS182098.

[9]. Folweiler KA, Sandsmark DK, Diaz-Arrastia R, Cohen AS, Masino AJ. Unsupervised machine learning reveals novel traumatic brain injury patient phenotypes with distinct acute injury profiles and long-term outcomes. J Neurotrauma 2020;37:1431–44. 10.1089/neu.2019.6705. [PubMed: 32008422]

[10]. Asgari S, Adams H, Kasprowicz M, Czosnyka M, Smielewski P, Ercole A. Feasibility of hidden markov models for the description of time-varying physiologic state after severe traumatic brain injury. Crit Care Med 2019;47:e880–5. 10.1097/CCM.0000000000003966. [PubMed: 31517697]

[11]. Mitra J, Shen K, Ghose S, Bourgeat P, Fripp J, Salvado O, et al. Statistical machine learning to identify traumatic brain injury (TBI) from structural disconnections of white matter networks. Neuroimage 2016;129:247–59. 10.1016/j.neuroimage.2016.01.056. [PubMed: 26827816]

[12]. Raj R, Luostarinen T, Pursiainen E, Posti JP, Takala RSK, Bendel S, et al. Machine learning-based dynamic mortality prediction after traumatic brain injury. Sci Rep 2019;9:17672. 10.1038/s41598-019-53889-6. [PubMed: 31776366]

[13]. Rau C-S, Kuo P-J, Chien P-C, Huang C-Y, Hsieh H-Y, Hsieh C-H. Mortality prediction in patients with isolated moderate and severe traumatic brain injury using machine learning models. PLoS One 2018;13:e0207192. 10.1371/journal.pone.0207192. [PubMed: 30412613]

[14]. Brown A, Malec J, McClelland R, Diehl N, Englander J, Cifu D. Clinical elements that predict outcome after traumatic brain injury A prospective multicenter recursive partitioning (Decision-Tree) analysis. J Neurotrauma 2005;22:1040–51. 10.1089/neu.2005.22.1040. [PubMed: 16238482]

[15]. Rizoli S, Peterson A, Bugler E, Coimbra R, Kerby J, Minei J, et al. Early prediction of outcome after severe traumatic brain injury a simple and practical model. BMC Emerg Med 2016. 10.1186/s12873-016-0098-x.

[16]. Jana S, Sutton M, Mollayeva T, Chan V, Colantonio A, Escobar MD. Application of multiple testing procedures for identifying relevant comorbidities, from a large set, in traumatic brain injury for research applications utilizing big health-administrative data. Front Big Data 2022;5:793606. 10.3389/fdata.2022.793606. [PubMed: 36247970]

[17]. Mollayeva T, Tran A, Chan V, Colantonio A, Sutton M, Escobar M. Decoding health status transitions of over 200 000 patients with traumatic brain injury from preceding injury to the injury event. Sci Rep 2022. 10.1038/s41598-022-08782-0.

[18]. Mollayeva T, Tran A, Chan V, Colantonio A, Escobar MD. Sex-specific analysis of traumatic brain injury events: applying computational and data visualization techniques to inform prevention and management. BMC Med Res Methodol 2022;22:30. 10.1186/s12874-021-01493-6. [PubMed: 35094688]

[19]. Sahler CS, Greenwald BD. Traumatic brain injury in sports: a review, rehabilitation research and practice (2012. 2012. p. 1–10. 10.1155/2012/659652.

[20]. Winqvist S, Luukinen H, Jokelainen J, Lehtilahti M, Näyhä S, Hillbom M. Recurrent traumatic brain injury is predicted by the index injury occurring under the influence of alcohol. Brain Inj 2008;22:780–5. 10.1080/02699050802339397. [PubMed: 18787988]

[21]. Winqvist S, Jokelainen J, Luukinen H, Hillbom M. Adolescents' drinking habits predict later occurrence of traumatic brain injury: 35-year follow-up of the northern Finland 1966 birth cohort. J Adolesc Health 2006;39. 10.1016/j.jadohealth.2005.12.019. 275.e1–275.e7.

[22]. Viano D, von Holst H, Gordon E. Serious brain injury from traffic-related causes: priorities for primary prevention. Accid Anal Prev 1997;29:811–6. 10.1016/S0001-4575(97)00050-X. [PubMed: 9370017]

[23]. Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. J Mach Learn Res 2003;3:993–1022.

[24]. Cohen R, Aviram I, Elhadad M, Elhadad N. Redundancy-aware topic modeling for patient record notes. PLoS One 2014;9:e87555. 10.1371/journal.pone.0087555. [PubMed: 24551060]

[25]. Gangavarapu T, Jayasimha A, Krishnan G, Kamath S. Predicting ICD-9 code groups with fuzzy similarity based supervised multi-label classification of unstructured clinical nursing notes. Knowl Base Syst 2020;190. 10.1016/j.knosys.2019.105321.

[26]. Dzubur E, Khalil C, Almario CV, Noah B, Minhas D, Ishimori M, et al. Patient concerns and perceptions regarding biologic therapies in ankylosing spondylitis: insights from a large-scale survey of social media platforms. Arthritis Care Res 2019;71:323–30. 10.1002/acr.23600.

[27]. Fairie P, Zhang Z, D'Souza A, Walsh T, Quan H, Santana M. Categorising patient concerns using natural language processing techniques. BMJ Health Care Inf 2021;28. 10.1136/bmjhci-2020-100274.

[28]. Li Y, Rapkin B, Atkinson TM, Schofield E, Bochner BH. Leveraging Latent Dirichlet Allocation in processing free-text personal goals among patients undergoing bladder cancer surgery. Qual Life Res 2019;28:1441–55. 10.1007/s11136-019-02132-w. [PubMed: 30798421]

[29]. Okon E, Rachakonda V, Hong HJ, Callison-Burch C, Lipoff JB. Natural language processing of Reddit data to evaluate dermatology patient experiences and therapeutics. J Am Acad Dermatol 2020;83:803–8. 10.1016/j.jaad.2019.07.014. [PubMed: 31306722]

[30]. Wagland R, Recio-Saucedo A, Simon M, Bracher M, Hunt K, Foster C, et al. Development and testing of a text-mining approach to analyse patients' comments on their experiences of colorectal cancer care. BMJ Qual Saf 2016;25:604–14. 10.1136/bmjqs-2015-004063.

[31]. Yang M, Kiang M, Shang W. Filtering big data from social media – building an early warning system for adverse drug reactions. J Biomed Inf 2015;54:230–40. 10.1016/j.jbi.2015.01.011.

[32]. Li DC, Thermeau T, Christopher Chute, Liu H. Discovering associations among diagnosis groups using topic modeling. AMIA Jt Summits Transl Sci Proc 2014. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4419765/pdf/1860052.pdf.

[33]. Chiudinelli L, Dagliati A, Tibollo V, Albasini S, Geifman N, Peek N, et al. Mining post-surgical care processes in breast cancer patients. Artif Intell Med 2020. 10.1016/j.artmed.2020.101855.

[34]. Shao Y, Morris R, Bray B, Zeng-Treitler Q. Topic modeling based on ICD codes for clinical documents. In: Intelligent systems and applications; 2022. p. 184–98. 10.1007/978-3-030-82196-8_14.

[35]. Aggarwal CC, Zhai C, editors. Mining text data. Boston, MA: Springer US; 2012. 10.1007/978-1-4614-3223-4.

[36]. Campbell JC, Hindle A, Stroulia E. Latent dirichlet allocation. In: The art and science of analyzing software data. Elsevier; 2015. p. 139–59. 10.1016/B978-0-12-411519-4.00006-9.

[37]. Iron K, Manuel DG. Quality assessment of administrative data (QuAAD): an opportunity for enhancing ontario's health data. 2007. https://www.ices.on.ca/flip-publication/quality-assessment-of-administrative/files/assets/basic-html/toc.html (accessed October 1, 2021).

[38]. Canadian Institute for Health Information. International statistical classification of diseases and related health problems, Tenth revision, Canada (ICD-10-CA): alphabetical index and tabular list. 2022.

[39]. Fegan S. Recommended ICD-10-CA codes for injury core indicators. Association of Public Health Epidemiologists in Ontario; 2013. http://core.apheo.ca/index.php?pid=306.

[40]. Ontario Health Insurance Plan Claims - OHIP Diagnosis Code, ICES Data Dictionary. (n.d.). https://datadictionary.ices.on.ca/Applications/DataDictionary/Variables.aspx?LibName=OHIP&MemName=&Variable=DXCODE (accessed November 1, 2022).

[41]. McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika 1947;12:153–7. 10.1007/BF02295996. [PubMed: 20254758]

[42]. Yarnold P. UniODA vs. Mcnemar's test for correlated proportions: diagnosis of disease before vs. After treatment. Optimal Data Anal 2015;4:24–6.

[43]. Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. Ann Stat 2001;29. 10.1214/aos/1013699998.

[44]. Chedzoy OB. Phi-coefficient. In: Kotz S, Read CB, Balakrishnan N, Vidakovic B, Johnson NL, editors. Encyclopedia of statistical Sciences. Hoboken, NJ, USA: John Wiley & Sons, Inc.; 2006. 10.1002/0471667196.ess1960.pub2. ess1960.pub2.

[45]. Grün B, Hornik K. Topic models. 2022. https://cran.r-project.org/web/packages/topicmodels/topicmodels.pdf (accessed November 3, 2022).

[46]. Chang J. Collapsed Gibbs sampling methods for topic models. 2015. https://cran.r-project.org/web/packages/lda/lda.pdf (accessed November 3, 2022).

[47]. Nikita M, Chaney N. Tuning of the latent dirichlet allocation models parameters. 2020. https://cran.r-project.org/web/packages/ldatuning/index.html (accessed November 3, 2022).

[48]. Arun R, Suresh V, Veni Madhavan CE, Narasimha Murthy MN. On finding the natural number of topics with latent dirichlet allocation: some observations. In: Zaki MJ, Yu JX, Ravindran B, Pudi V, editors. Advances in knowledge discovery and data mining. Berlin, Heidelberg: Springer Berlin Heidelberg; 2010. p. 391–402. 10.1007/978-3-642-13657-3_43.

[49]. Cao J, Xia T, Li J, Zhang Y, Tang S. A density-based method for adaptive LDA model selection. Neurocomputing 2009;72:1775–81. 10.1016/j.neucom.2008.06.011.

[50]. Deveaud R, SanJuan E, Bellot P. Accurate and effective latent concept modeling for ad hoc information retrieval. Doc Numér (Paris) 2014;17:61–84. 10.3166/dn.17.1.61-84.

[51]. Griffiths TL, Steyvers M. Finding scientific topics. Proc Natl Acad Sci USA 2004;101:5228–35. 10.1073/pnas.0307752101. [PubMed: 14872004]

[52]. El Assaad H, Samé A, Govaert G, Aknin P. A variational Expectation–Maximization algorithm for temporal data clustering. Comput Stat Data Anal 2016;103:206–28. 10.1016/j.csda.2016.05.007.

[53]. McCaw Z. Rank normal transformation omnibus test. 2022. https://cran.r-project.org/web/packages/RNOmni/index.html (accessed November 5, 2022).

[54]. Breiman L. Random forests. Mach Learn 2001;45:5–32. 10.1023/A:1010933404324.

[55]. Hapfelmeier A, Ulm K. A new variable selection approach using Random Forests. Comput Stat Data Anal 2013;60:50–69. 10.1016/j.csda.2012.09.020.

[56]. Ishwaran H, Kogalur UB. Fast unified random forests for survival, regression, and classification (RF-SRC). 2022. https://cran.r-project.org/web/packages/randomForestSRC/randomForestSRC.pdf (accessed November 10, 2022).

[57]. Chan V, Sutton M, Mollayeva T, Escobar MD, Hurst M, Colantonio A. Data mining to understand how health status preceding traumatic brain injury affects functional outcome: a population-based sex-stratified study. Arch Phys Med Rehabil 2020;101:1523–31. 10.1016/j.apmr.2020.05.017. [PubMed: 32544398]

[58]. Maas AIR, Menon DK, Adelson PD, Andelic N, Bell MJ, Belli A, et al. Traumatic brain injury: integrated approaches to improve prevention, clinical care, and research. Lancet Neurol 2017;16:987–1048. 10.1016/S1474-4422(17)30371-X. [PubMed: 29122524]

[59]. Ill-defined diseases, (n.d.). https://platform.who.int/mortality/themes/theme-details/MDB/ill-defined-diseases (accessed August 20, 2023).

[60]. Chan V, Zagorski B, Parsons D, Colantonio A. Older adults with acquired brain injury: a population based study. BMC Geriatr 2013;13:97. 10.1186/1471-2318-13-97. [PubMed: 24060144]

[61]. Peterson AB, Surveillance report of traumatic brain injury–related hospitalizations and deaths by age group, sex, and mechanism of injury—United States, 2016-2017, Centers for Disease Control and Prevention., n.d. https://www.cdc.gov/traumaticbraininjury/pdf/TBI-surveillance-report-2016-2017-508.pdf (accessed March 4, 2023).

[62]. Ewing-Cobbs L, Barnes MA, Fletcher JM. Early brain injury in children: development and reorganization of cognitive function. Dev Neuropsychol 2003;24:669–704. 10.1080/87565641.2003.9651915. [PubMed: 14561566]

[63]. Amgalan A, Maher AS, Ghosh S, Chui HC, Bogdan P, Irimia A. Brain age estimation reveals older adults' accelerated senescence after traumatic brain injury. Geroscience 2022;44:2509–25. 10.1007/s11357-022-00597-1. [PubMed: 35792961]

[64]. Tremblay S, Desjardins M, Bermudez P, Iturria-Medina Y, Evans AC, Jolicoeur P, et al. Mild traumatic brain injury: the effect of age at trauma onset on brain structure integrity. Neuroimage Clin 2019;23:101907. 10.1016/j.nicl.2019.101907. [PubMed: 31233955]

[65]. Violence Prevention. Centers for disease control and prevention. https://www.cdc.gov/violenceprevention/childabuseandneglect/Abusive-Head-Trauma.html; 2022.

[66]. Davies FC, Coats TJ, Fisher R, Lawrence T, Lecky FE. A profile of suspected child abuse as a subgroup of major trauma patients. Emerg Med J 2015;32:921–5. 10.1136/emermed-2015-205285. [PubMed: 26598630]

[67]. Bruschetta R, Tartarisco G, Lucca LF, Leto E, Ursino M, Tonin P, et al. Predicting outcome of traumatic brain injury: is machine learning the best way? Biomedicines 2022;10:686. 10.3390/biomedicines10030686. [PubMed: 35327488]

[68]. Chong S-L, Liu N, Barbier S, Ong MEH. Predictive modeling in pediatric traumatic brain injury using machine learning. BMC Med Res Methodol 2015;15:22. 10.1186/s12874-015-0015-0. [PubMed: 25886156]

[69]. Rahimi M, Zahedi M, Mashayekhi H. A probabilistic topic model based on short distance Co-occurrences. Expert Syst Appl 2022;193:116518. 10.1016/j.eswa.2022.116518.

[70]. Pérez J, Pérez A, Casillas A, Gojenola K. Cardiology record multi-label classification using latent Dirichlet allocation. Comput Methods Progr Biomed 2018;164:111–9. 10.1016/j.cmpb.2018.07.002.

[71]. Geletta S, Follett L, Laugerman M. Latent Dirichlet Allocation in predicting clinical trial terminations. BMC Med Inf Decis Making 2019;19:242. 10.1186/s12911-019-0973-y.

[72]. Sayadi K, Bui QV, Bui M. Multilayer classification of web pages using random forest and semi-supervised latent dirichlet allocation. In: 2015 15th international conference on innovations for community services (I4CS). Nuremberg, Germany: IEEE; 2015. p. 1–7. 10.1109/I4CS.2015.7294479.

[73]. Statistics on Brain Injury, Brain Injury Canada. (n.d.). https://braininjurycanada.ca/en/statistics/ (accessed January 10, 2023).

[74]. Fernández A, García S, Herrera F. Addressing the classification with imbalanced data: open problems and new challenges on class distribution. In: Corchado E, Kurzyński M, Woźniak M, editors. Hybrid artificial intelligent systems. Berlin, Heidelberg: Springer Berlin Heidelberg; 2011. p. 1–10. 10.1007/978-3-642-21219-2_1.

[75]. López V, Fernández A, García S, Palade V, Herrera F. An insight into classification with imbalanced data: empirical results and current trends on using data intrinsic characteristics. Inf Sci 2013;250:113–41. 10.1016/j.ins.2013.07.007.

[76]. Karwat ID, Krupa S, Gorczyca R. Causes and consequences of head injuries among rural inhabitants hospitalised in a Multi-organ Injury Ward. II. Circumstances, types and consequences of head injuries. Ann Agric Environ Med 2009;16:23–9. [PubMed: 19572474]

[77]. Kumar RG, Juengst SB, Wang Z, Dams-O'Connor K, Dikmen SS, O'Neil-Pirozzi TM, et al. Epidemiology of comorbid conditions among adults 50 Years and older with traumatic brain injury. J Head Trauma Rehabil 2018;33:15–24. 10.1097/HTR.0000000000000273. [PubMed: 28060201]

[78]. Osborn AJ, Mathias JL, Fairweather-Schmidt AK, Anstey KJ. Anxiety and comorbid depression following traumatic brain injury in a community-based sample of young, middle-aged and older adults. J Affect Disord 2017;213:214–21. 10.1016/j.jad.2016.09.045. [PubMed: 27919428]

[79]. Pugh MJ, Finley EP, Wang C-P, Copeland LA, Jaramillo CA, Swan AA, et al. A retrospective cohort study of comorbidity trajectories associated with traumatic brain injury in veterans of the Iraq and Afghanistan wars. Brain Inj 2016;30:1481–90. 10.1080/02699052.2016.1219055. [PubMed: 27834535]

[80]. Daugherty J, Waltzman D, Sarmiento K, Xu L. Traumatic brain injury–related deaths by race/ethnicity, sex, intent, and mechanism of injury — United States, 2000–2017. MMWR Morb Mortal Wkly Rep 2019;68:1050–6. 10.15585/mmwr.mm6846a2. [PubMed: 31751321]

[81]. Kucukboyaci NE, Long C, Smith M, Rath JF, Bushnik T. Cluster analysis of vulnerable groups in acute traumatic brain injury rehabilitation. Arch Phys Med Rehabil 2018;99:2365–9. 10.1016/j.apmr.2017.11.016. [PubMed: 29317223]

[82]. Chang P-FJ, Ostir GV, Kuo Y-F, Granger CV, Ottenbacher KJ. Ethnic differences in Discharge destination among older patients with traumatic brain injury. Arch Phys Med Rehabil 2008;89:231–6. 10.1016/j.apmr.2007.08.143. [PubMed: 18226645]

[83]. Gao S, Kumar RG, Wisniewski SR, Fabio A. Disparities in health care utilization of adults with traumatic brain injuries are related to insurance, race, and ethnicity: a systematic review. J Head Trauma Rehabil 2018;33:E40–50. 10.1097/HTR.0000000000000338. [PubMed: 28926481]

[84]. Chang VC, Ruseckaite R, Collie A, Colantonio A. Examining the epidemiology of work-related traumatic brain injury through a sex/gender lens: analysis of workers' compensation claims in Victoria, Australia. Occup Environ Med 2014;71:695–703. 10.1136/oemed-2014-102097. [PubMed: 25052083]

[85]. Colantonio A, Mroczek D, Patel J, Lewko J, Fergenbaum J, Brison R. Examining occupational traumatic brain injury in Ontario. Can J Public Health 2010;101:S58–62. 10.1007/BF03403848. [PubMed: 20629449]

[86]. Chang VC, Guerriero EN, Colantonio A. Epidemiology of work-related traumatic brain injury: a systematic review: work-Related Traumatic Brain Injury. Am J Ind Med 2015;58:353–77. 10.1002/ajim.22418. [PubMed: 25731875]

[87]. Sumowski JF, Chiaravalloti N, Krch D, Paxton J, DeLuca J. Education attenuates the negative impact of traumatic brain injury on cognitive status. Arch Phys Med Rehabil 2013;94:2562–4. 10.1016/j.apmr.2013.07.023. [PubMed: 23932968]

[88]. Kesler SR, Adams HF, Blasey CM, Bigler ED. Premorbid intellectual functioning, education, and brain size in traumatic brain injury: an investigation of the cognitive reserve hypothesis. Appl Neuropsychol 2003;10:153–62. 10.1207/S15324826AN1003_04. [PubMed: 12890641]

[89]. Dumais ST, Furnas GW, Landauer TK, Deerwester S, Harshman R. Using latent semantic analysis to improve access to textual information. In: Proceedings of the SIGCHI conference on human factors in computing systems - CHI '88. Washington, D.C., United States: ACM Press; 1988. p. 281–5. 10.1145/57167.57214.

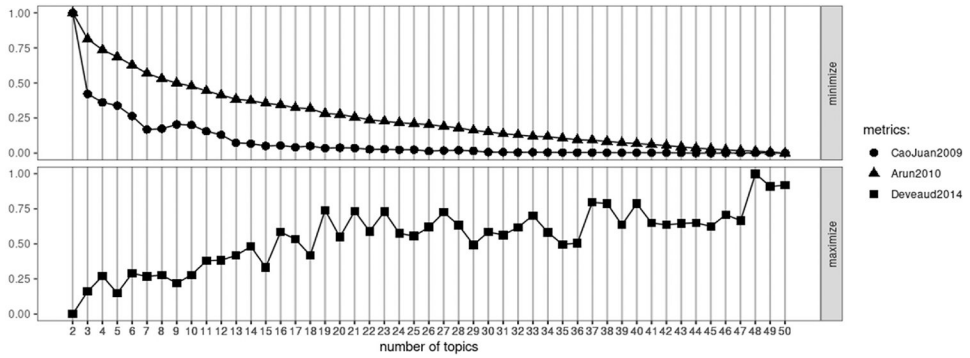

Fig. 1.

LDA tuning number of topics. Line graphs showing calculated metrics across different number of topics. *Note*. Griffiths2004 did not converge.

Table 1

Summary statistics by split sets. Abbreviations: SD = standard deviation; Pctl. = percentile; AIS = Abbreviated Injury Scale.

Variables	Training	Validation	Test
	(N = 488,128)	**(N = 244,036)**	**(N = 244,050)**
Socio-demographics			
Age at injury (years)			
Mean	35.1	35.1	35.1
SD	25.5	25.6	25.6
Pctl. 25	15	15	15
Median	27	27	27
Pctl. 75	54	54	54
Sex, n (%)			
Male	274,186 (56.2)	136,412 (55.9)	137,220 (56.2)
Rurality indicator, n (%)			
Rural	75,516 (15.5)	37,578 (15.4)	37,828 (15.5)
Income quintile, n (%)			
Q1 (lowest income)	99,290 (20.3)	49,216 (20.2)	49,390 (20.2)
Q2	94,866 (19.4)	47,638 (19.5)	47,518 (19.5)
Q3	96,042 (19.7)	48,174 (19.7)	47,708 (19.5)
Q4	98,886 (20.3)	49,880 (20.4)	49,914 (20.5)
Q5 (highest income)	99,044 (20.3)	49,128 (20.1)	49,520 (20.3)
OHIP-specific[a]			
Patients with a diagnosis, n (%)	463,943 (95.1)	231,889 (95.0)	231,855 (95.0)
Total number of diagnoses	20,441,616	10,290,796	10,271,548
TBI-specific[b]			
Patients with TBI	244,064	122,018	122,025
Injury severity, n (%)			
Unknown	111,387 (45.6)	55,869 (45.8)	55,958 (45.9)
Mild (AIS 1–2)	101,081 (41.4)	50,235 (41.2)	50,087 (41.0)
Moderate (AIS = 3)	6072 (2.5)	3076 (2.5)	3134 (2.6)
Severe (AIS >3)	25,524 (10.5)	12,838 (10.5)	12,846 (10.5)
Cause of injury, n (%)			
Falls	109,485 (44.9)	54,707 (44.8)	54,703 (44.8)
Struck by/against	84,382 (34.6)	42,351 (34.7)	42,012 (34.4)
Cyclist Collision	5956 (2.4)	3005 (2.5)	3046 (2.5)
Motor Vehicle Collision	25,742 (10.5)	12,675 (10.4)	13,137 (10.8)
Other	21,263 (8.7)	10,725 (8.8)	10,586 (8.7)

[a]Summary for patients with OHIP diagnosis.

[b]Summary for patients with TBI.

Table 2

Topic names. Names were assigned based on their top diagnostic codes (Supplementary Fig. 1 and Supplementary Table 3 for the full list of selected diagnostic codes).

There were 24,860 TBI patients and 38,284 control patients who were assigned $TS_i = 0$ *where* $i = 1, 2,\ldots,19$ due to not being diagnosed with any of the 314 selected diagnostic codes during their pre-injury phase. For the remainder of the patients, topic scores were computed for all 19 topics as per Equation (1), which were extremely positively skewed for each group, with the highest range being 0.0368 to 2796.025 for TBI-Validation set (Table 3 for statistics). Topic scores were then rank-normalized within each group of the split sets. The normalized topic scores, from here on referred to as $Topic_i$ *where* $i = 1, 2,\ldots, 19$, were then used along with socio-demographic variables (*age at injury, sex, rurality indicator, income quintile)* for prediction tasks.

Topic	Name
Topic 1	Heart disorders and symptoms
Topic 2	Microbial and allergic skin, eye, and upper airway manifestations
Topic 3	Stroke and excessive involuntary movement disorders
Topic 4	Kidney failure, sepsis, anemia
Topic 5	Injuries to extremities, disorders of musculoskeletal system, overexertion
Topic 6	Respiratory disorders, obstruction, and symptoms
Topic 7	Infectious/inflammatory diseases of skin and urinary system, and related concerns
Topic 8	Arthritis and manifestations of inflammatory, traumatic, and autoimmune disorders
Topic 9	Disorders of sense organs, trauma of upper extremities, and nutritional deficiencies
Topic 10	Inflammatory, traumatic multisystem disorders and mental manifestations
Topic 11	Lesions in anorectal area, venous and skin insufficiencies
Topic 12	General symptoms, migraine, and diseases of central nervous system
Topic 13	Unspecified back, depressive, and other diseases
Topic 14	Newborn/infant care and ill-defined infections
Topic 15	Unspecified anxiety, neurotic, and associated states
Topic 16	Symptoms of digestive, nervous, musculoskeletal, and excretion systems
Topic 17	Drug dependence and addiction
Topic 18	Unknown causes of morbidity and mortality
Topic 19	Disorders of glucose intolerance and complications

Table 3

Summary statistics for each split of cases and controls after subsetting to the selected diagnostic codes.

	TBI			Control		
	Training	**Validation**	**Test**	**Training**	**Validation**	**Test**
Number of patients	231,594	115,874	115,779	225,042	112,314	112,467
Male, n (%)	128,070 (55.3)	63,789 (55.1)	64,094 (55.4)	124,949 (55.5)	62,031 (55.2)	62,554 (55.6)
Total OHIP diagnoses	9,786,264	4,957,267	4,939,669	6,729,081	3,373,209	3,362,571
OHIP diagnoses per *patient*						
Min	1	1	1	1	1	1
Max	2586	2816	3068	3033	2013	2167
Mean	42.26	42.78	42.66	29.9	30.03	29.9
Median	22	22	22	16	16	16
Topic scores						
Min	0.04	0.04	0.04	0.04	0.04	0.04
Max	2426.89	2796.03	2687.77	2200.59	1716.55	2098.76
Mean	2.22	2.25	2.25	1.57	1.58	1.57
Median	0.11	0.11	0.11	0.11	0.11	0.11

Table 4

Random forest hyperparameter tuning for each prediction task.

Target variable		Number of Trees	Terminal Node Size	Sampled Variables	Validation AUC
TBI		100	4	22	0.906
		200	1	18	0.9068
		300	1	22	0.908
		400	1	15	0.9062
Cause of Injury	Falls	100	2	22	0.8954
		200	1	22	0.8975
		300	1	22	0.8987
		400	1	22	0.8988
		500	1	18	0.8988
	Struck by/against	100	1	6	0.892
		200	1	6	0.8969
		300	2	5	0.8939
	Cyclist Collision	100	1	18	0.856
		200	1	18	0.8589
		300	1	18	0.86
		400	2	22	0.8664
		500	2	12	0.8629
	Motor Vehicle Collision	100	4	22	0.8944
		200	2	22	0.8971
		300	2	18	0.8981
		400	2	23	0.8985
		500	2	18	0.8989
		600	1	18	0.8997
		700	1	18	0.8996

Table 5

Variable importance (VIMP) of the candidate predictors obtained from the final models. Age refers to *age at injury,* rural refers to *rurality indicator,* incquint refers to *income quintile,* Topic refers to normalized topic score variables.

Rank	TBI		Cause of Injury							
			Falls		Struck by/against		Cyclist Collision		Motor Vehicle Collision	
	Variable	VIMP	Variable	VIMP	Variable	VIMP	Variable	VIMP	Variable	VIMP
1	Topic 17	0.2320	Topic 14	0.2274	Topic 17	0.1991	Topic 17	0.1353	Topic 17	0.1744
2	Topic 14	0.2242	Topic 17	0.2155	Topic 5	0.1761	age	0.1245	Topic 14	0.1655
3	Topic 5	0.1782	Topic 5	0.1609	Topic 14	0.1610	Topic 14	0.1187	Topic 4	0.1235
4	Topic 4	0.1341	Topic 2	0.1400	Topic 4	0.1440	Topic 4	0.1124	Topic 5	0.1123
5	Topic 2	0.1277	Topic 4	0.1163	Topic 2	0.1404	Topic 5	0.1009	age	0.1112
6	Topic 3	0.1109	Topic 12	0.1065	Topic 19	0.1202	Topic 2	0.0999	Topic 2	0.0971
7	age	0.1034	Topic 3	0.1041	age	0.1110	Topic 19	0.0989	Topic 3	0.0894
8	Topic 19	0.0947	Topic 1	0.0931	Topic 8	0.1003	Topic 3	0.0780	Topic 15	0.0860
9	Topic 13	0.0936	Topic 15	0.0912	Topic 13	0.0971	Topic 15	0.0680	Topic 13	0.0788
10	Topic 1	0.0886	Topic 13	0.0875	Topic 3	0.0941	Topic 18	0.0664	Topic 12	0.0714
11	Topic 12	0.0862	age	0.0803	Topic 18	0.0911	Topic 13	0.0632	Topic 19	0.0659
12	Topic 15	0.0834	Topic 16	0.0712	Topic 1	0.0906	Topic 12	0.0582	Topic 18	0.0644
13	Topic 16	0.0761	Topic 18	0.0699	Topic 12	0.0885	Topic 1	0.0549	Topic 1	0.0630
14	Topic 9	0.0716	Topic 19	0.0695	Topic 9	0.0852	Topic 8	0.0533	Topic 16	0.0623
15	Topic 8	0.0714	Topic 9	0.0680	Topic 15	0.0843	Topic 9	0.0452	Topic 9	0.0621
16	Topic 18	0.0650	Topic 6	0.0544	Topic 16	0.0843	Topic 16	0.0448	Topic 8	0.0595
17	Topic 10	0.0625	Topic 8	0.0536	Topic 7	0.0772	Topic 7	0.0434	Topic 10	0.0558
18	Topic 6	0.0570	Topic 10	0.0465	Topic 6	0.0756	Topic 10	0.0428	Topic 6	0.0444
19	Topic 7	0.0492	Topic 11	0.0400	Topic 11	0.0738	Topic 6	0.0394	Topic 7	0.0429
20	Topic 11	0.0476	Topic 7	0.0368	Topic 10	0.0733	Topic 11	0.0373	Topic 11	0.0382
21	rural	0.0065	rural	0.0077	rural	0.0090	rural	0.0092	rural	0.0087
22	incquint	0.0034	incquint	0.0028	sex	0.0051	sex	0.0051	incquint	0.0027
23	sex	0.0016	sex	0.0007	incquint	0.0029	incquint	0.0027	sex	0.0015