Identification of Transgender and Gender-Diverse Individuals in Electronic Health Records

Yining Hua[1,2], Liqin Wang[2], Vi Nguyen[2], David W. Bates[2], Dinah Foer, MD[2*], Li Zhou, MD, PhD[1,2*]

[1]Department of Biomedical Informatics, [2]Harvard Medical School; Division of General Internal Medicine and Primary Care, Brigham and Women's Hospital}, Boston, MA, USA

Background

Problem: Identification of **transgender and gender diverse (TGD)** patients from electronic health record (EHR) data remains a central challenge.

What is already known: Recent studies that propose using retrospective EHR data to identify patient gender identity primarily **rely on structured data** and **rule-based algorithms** using limited sets of medical codes and related keywords, which has low accuracy due to missing information in the structured EHR data and strict rules of exact matches.

Aim: To build a deep learning-based gender identification model leveraging free-text notes and structured data.

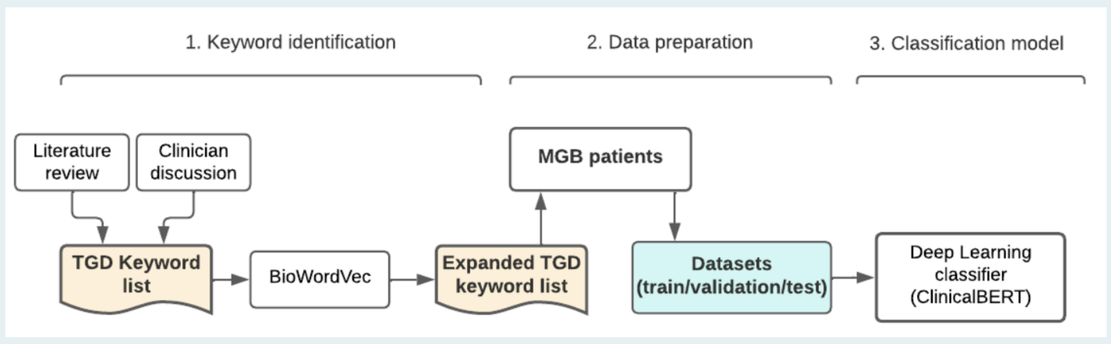
Methods

Setting: This study was conducted at the **Mass General Brigham (MGB) health system** with the Research Patient Data Registry and the Enterprise Data Warehouse.

Inclusion criteria: Patients who had ≥1 encounter between April 1, 2017 and April 1, 2022, and were adults (18 years old and above) by the start time of the study.

Data types: (1) sex and gender demographics fields; (2) patient diagnoses from the International Classification of Diseases (ICD), Clinical Modification codes; (3) patient procedures from the ICD Procedures codes; (4) problem lists; (5) free-text notes, including procedure, admit, progress, visit, discharge, and H&P.

Key steps in pipeline development



1. Keyword Identification

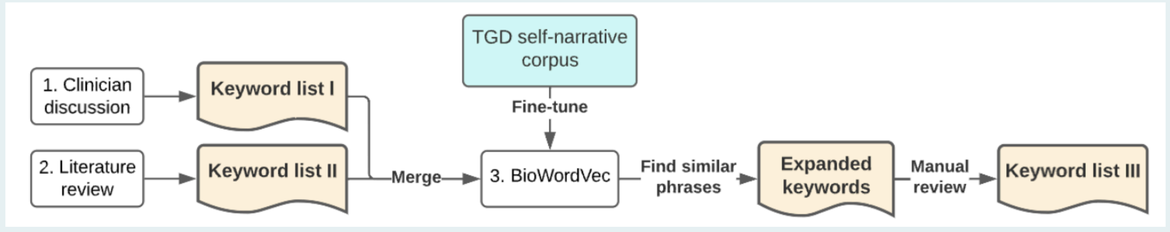


Figure 2. A BioWordVec [1] method for generating comprehensive TGD keywords.

2. Data Preparation

- We matched code names in structured data and used templates to concatenate them with clinical notes in the following order: (1) diagnoses and procedures, (2) sex and gender demographics, (3) note sentences.
- We split Dataset I into a training set and Test set I, manually reviewed 200 patients from Dataset II to be Test set III, and created Testset II and Testset IV by excluding patients with explicit sex and gender demographics.

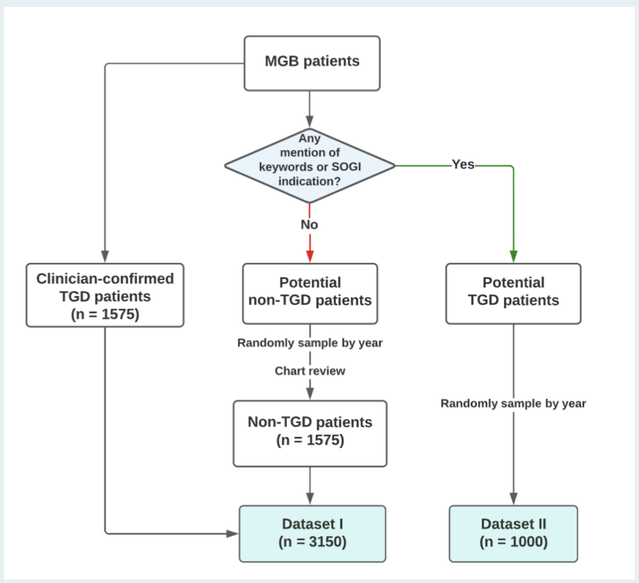


Figure 3. Data retrieval and data set curation.

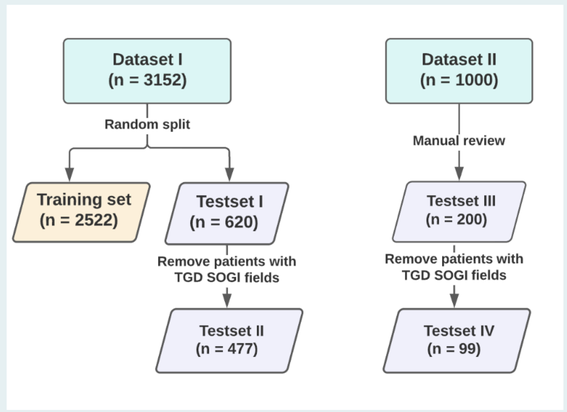


Figure 4. Building training and test sets.

3. Classification Models

- **The deep learning model:** We implemented the classification model (*Bio_ClinicalBERT_TGD*), with **Bio_ClinicalBERT [3]** as the prototype. We fine-tuned the model on a binary classification task of assigning whether a patient is TGD or not.
- **Baselines:** We used the best identification pipeline in the literature: **Guo et al.'s** best single-rule algorithm (2 diagnosis codes + 1 keyword) and best combined-rule algorithm [(gender field indicates transgender) OR (>= 1 diagnosis code AND >= 1 TGD keyword)] [2].
- **Other models:** Considering the hardware requirement of BERT models, we also experimented with the traditional statistical machine learning methods. Instead of BERT, we used **TF-IDF** to encode the texts We used **Support Vector Machine (SVM)**, **Random Forest**, and **Logistic Regression** to classify the encoded texts.

Results

Keywords The final keyword list has 109 keywords (58 new) with at least one appearance.

Model performance on Testset I & II (Tables 1&2)

- All machine learning algorithms (random forest, logistic regression, support vector machine, and *Bio_ClinicalBERT_TGD*) outperformed the rule baselines on both Testset I and Testset II. Overall, the performances were **stable** both within and cross the test sets.
- *Bio_ClinicalBERT_TGD* had the best overall performance.
- Logistics Regression on TF-IDF had comparable performance.

	Sensitivity	Specificity	PPV	NPV	F1-score	Accuracy
Guo et al. (best single rule)	75.08	94.59	92.80	80.36	83.01	85.20
Guo et al. (best combined rule)	78.96	94.59	93.13	82.89	85.46	87.07
Random Forest	82.00 ± 0.01	98.27 ± 0.01	97.95 ± 0.01	84.41 ± 0.01	89.26 ± 0.01	90.10 ± 0.00
Support Vector Machine	83.37 ± 0.02	98.55 ± 0.00	98.38 ± 0.01	84.76 ± 0.01	90.25 ± 0.01	90.72 ± 0.01
Logistic Regression	81.68 ± 0.03	**98.77 ± 0.01**	**98.56 ± 0.01**	83.92 ± 0.02	89.31 ± 0.02	90.08 ± 0.01
Bio_ClinicalBERT	**84.12 ± 0.01**	98.08 ± 0.01	98.05 ± 0.02	**86.45 ± 0.01**	**90.68 ± 0.02**	**91.59 ± 0.02**

Table 1. Performance of TGD identification algorithms on Testset I.

	Sensitivity	Specificity	PPV	NPV	F1-score	Accuracy
Guo et al. (best single rule)	67.38	95.12	88.73	83.65	76.60	85.05
Guo et al. (best combined rule)	70.59	95.12	89.19	85.01	78.81	86.21
Random Forest	82.47 ± 0.01	97.07 ± 0.02	97.22 ± 0.01	81.85 ± 0.07	89.23 ± 0.00	89.24 ± 0.02
Support Vector Machine	81.24 ± 0.01	98.22 ± 0.01	98.15 ± 0.01	81.11 ± 0.06	88.90 ± 0.01	89.07 ± 0.01
Logistic Regression	82.45 ± 0.01	**98.45 ± 0.00**	**98.19 ± 0.01**	**84.50 ± 0.01**	89.63 ± 0.01	**90.33 ± 0.01**
Bio_ClinicalBERT	**84.36 ± 0.01**	97.96 ± 0.02	97.96 ± 0.02	84.11 ± 0.02	**91.26 ± 0.01**	89.10 ± 0.02

Table 2. Performance of TGD identification algorithms on Testset II.

Model performance on Testset III & IV (Table 4) *Bio_ClinicalBERT_TGD* had F1 scores greater than 95%, regardless of determinable gender demographics.

	Sensitivity	Specificity	PPV	NPV	F1-score	Accuracy
Testset III	98.37	93.75	99.45	83.33	**98.91**	98.00
Testset IV	96.39	93.75	98.77	83.33	**97.56**	95.96

Table 3. Performance of Bio_ClinicalBERT-TGD on Testsets III & IV.

Discussion & Conclusion

NLP and clinical notes can help to identify patients' gender identities despite missing sex and gender demographics fields.

Bio_ClinicalBERT identifies patients more accurately than other approaches.

Random forest classifiers based on TF-IDF- encoded features provide **a faster prediction speed with comparable accuracy**.

Completeness of gender identity fields does not impact the availability of relevant information in the documentation of diagnoses, procedures, and notes.

Keyword augmentation using **publicly available cross-domain datasets helps identify relevant information**.

Some features in previous studies should be reconsidered: Previous literature generally considers information in both sexual orientation and gender identity fields. Since sexual orientation is independent of gender identity, we suggest excluding data from the sexual orientation demographics in future research. Also, phrases such as "gender identity" should not be used because they are often section names rather than meaningful keywords directly related to TGD.

Our pipeline **successfully categorizes patient gender at the patient level**, a complicated prediction task.

Future directions: Our pipeline **could be applied to note, section-level, or even sentence-level** predictions, requiring less text preprocessing but more labeling work.

References

[1] Zhang Y, Chen Q, Yang Z, Lin H, Lu Z. BioWordVec, improving biomedical word embeddings with subword information and MeSH. Sci Data. 2019;6(1):52. doi:10.1038/s41597-019-0055-0

[2] Guo Y, He X, Lyu T, et al. Developing and Validating a Computable Phenotype for the Identification of Transgender and Gender Nonconforming Individuals and Subgroups. *AMIA Annu Symp Proc AMIA Symp*. 2020;2020:514-523.

[3] Alsentzer E, Murphy J, Boag W, et al. Publicly Available Clinical BERT Embeddings. In: *Proceedings of the 2nd Clinical Natural Language Processing Workshop*.

Funding

This study was funded by a research award from CRICO (PI: D. Foer).