Assessing the Feasibility of a Machine Learning-Based Computational Phenotype for Identifying Transgender and Gender Diverse Patients in the OMOP Common Data Model

William A. Baumgartner Jr.[1], Tyler Strickland[2], Danielle M. Kline[1], Abby M. Pribish[3], Molly McCallum[4], Amanuail Gebregzabheir[2], Dani Loeb[1,5], Lisa M. Schilling[1]

[1] Division of General Internal Medicine, University of Colorado School of Medicine, [2] Division of Endocrinology, Metabolism, and Diabetes, University of Colorado School of Medicine, [3] Division of Cardiology, University of Colorado School of Medicine, [4] University of Colorado School of Medicine, [5] Icahn School of Medicine at Mount Sinai

Background

It is well-documented that persons who identify as transgender and gender diverse (TGGD) do not receive equitable access to healthcare compared to cisgender persons due to factors such as stigma, lack of provider education and clinical competency, and health insurance disparities, and that these factors negatively affect their health outcomes[1,2]. Understanding the degree to which TGGD persons are disproportionately affected is itself a challenge, as the identification of TGGD patients in the electronic health record (EHR) can be difficult due to reasons such as a lack of systematic processes to query patients about their self-identified gender and to document that information in a structured manner[3].

Efforts to develop computational phenotypes (CPs), for identifying TGGD persons using data from electronic health records have become more prevalent in recent years[4]. Much of the identification of TGGD persons and cohorts for observational research and quality improvement has relied on ICD codes (Table 1), and to a lesser extent, use of certain medications not expected for the EHR recorded sex (i.e., estrogen use when sex=male), receipt of certain procedures (i.e., breast augmentation when sex=male), and keywords mentioned in clinical notes[4]. The work presented here uses a rule-based CP based on oft-used TGGD-related diagnostic codes as a baseline approach, and explores the use of machine learning (ML) techniques to classify patients as TGGD or non-TGGD. Specifically, we assess the feasibility of identifying TGGD persons within an OMOP CDM instance given the current constraints of the CDM which mandates the *PERSON.gender_concept_id* field not be used to store gender identity and instead recommends gender identity be stored in the *OBSERVATION* table, but only includes *Male* and *Female* Gender classes as standard concepts in the Gender domain.

It should be noted that in our contemporary socio-political climate, there are ethical considerations when pursuing research involving TGGD persons. There are potential risks that the CPs discussed here could be used for maleficent, non-research purposes. At the outset of this project, we worked closely with our TGGD community advisory board (CAB) to ensure that the goals of this project align with goals of the TGGD community. We will continue to work with our TGGD CAB as our research progresses.

Methods

This research is approved by the Colorado Multiple Institutional Review Board, Protocol 20-2302. Health Data Compass (HDC), the enterprise health data warehouse of the University of Colorado Anschutz Medical Campus (CUAMC), compiled an OMOP CDM v5.3 instance consisting of an intentionally inclusive patient pool designed to capture adult TGGD patients. HDC incorporated information from the Sexual Orientation and Gender Identity (SOGI) flowsheet used by the EHR vendor at CUAMC into the OMOP CDM as a separate database table. Among other relevant information, the SOGI data includes fields denoting a patient's self-reported gender identity and sex assigned at birth.

Vocabulary	Concept code|Concept label
Snomed	87991007|Gender identity disorder 407374003|Transsexual
MeSH	D063106|Transgender Persons
ICD10	F64.0|Transsexualism F64.1|Dual-role transvestism F64.2|Gender identity disorders of childhood F64.8|Other gender identity disorders F64.9|Gender identity disorder, unspecified Z87.890|Personal history of sex reassignment
ICD9	302.5|Trans-sexualism 302.50|Trans-sexualism with unspecified sexual history 302.51|Trans-sexualism with asexual history 302.52|Trans-sexualism with homosexual history 302.53|Trans-sexualism with heterosexual history 302.6|Gender identity disorder in children 302.85|Gender identity disorder in adolescents or adults

Table 1: TGGD diagnosis codes compiled from literature review of TGGD cohort definitions.

Diagnostic codes often used to create rule-based cohort definitions to identify TGGD patients in the EHR were compiled via literature review (Table 1) and augmented by analysis of our OMOP CDM instance. A gold standard (GS) corpus of TGGD and non-TGGD patients was derived from our OMOP CDM instance using both the compiled diagnostic codes as well as SOGI data. TGGD patients were selected for the GS based on having one of the diagnostic codes or if their TGGD status can be inferred from their SOGI data, i.e., if a patient's sex assigned at birth does not match their gender identity. Figure 1 describes the algorithm for classifying patients as TGGD or non-TGGD based on diagnostic codes and SOGI data. Non-TGGD patients were also gathered from patients who do not have SOGI data associated with their health record.

A one-to-one matching of transgender to cisgender patients on age, sex assigned at birth, and followup time was performed using the *ccoptimalmatch* R library.[6] The resulting matched-patient corpus was split into training, and test sets. Two CP algorithms were constructed, a baseline CP that mirrors typical rule-based approaches described in the literature by looking for the presence of diagnostic codes (Table 1), and a CP that uses a random forest (RF) classifier. The RF classifier was trained with 5-fold cross validation on features including conditions, drug exposures, measurements, observations, and procedures using an approach based on the APHRODITE project[7]. Both CPs were evaluated on a held-out set of patients from the GS.

Results

The compiled OMOP CDM instance was created in July 2021 and consists of 144,228 patients who were selected based on having at least one visit since January 2010 and having any of a number of potentially TGGD-related diagnoses, medications, and procedures. Manual chart review has been conducted on a small subset of patients (330 at the time of this writing) in order to validate the extracted SOGI data. Cohen's kappa statistic (0.93) pertaining to the TGGD status of the patient demonstrated "almost perfect" agreement among reviewers[8].

The GS corpus contains 4513 TGGD patients, one-to-one matched with a corresponding cisgender patient. Ten percent (451 patients) of the extended GS was held out to use for final testing of CPs. Performance of the baseline CP on the GS test set is in line with performances reported in the literature[4] as it suffers from relatively poor sensitivity (Table 2). The RF CP demonstrated improved sensitivity over the baseline CP (0.905 vs 0.529) with a modest drop in specificity (Table 2).

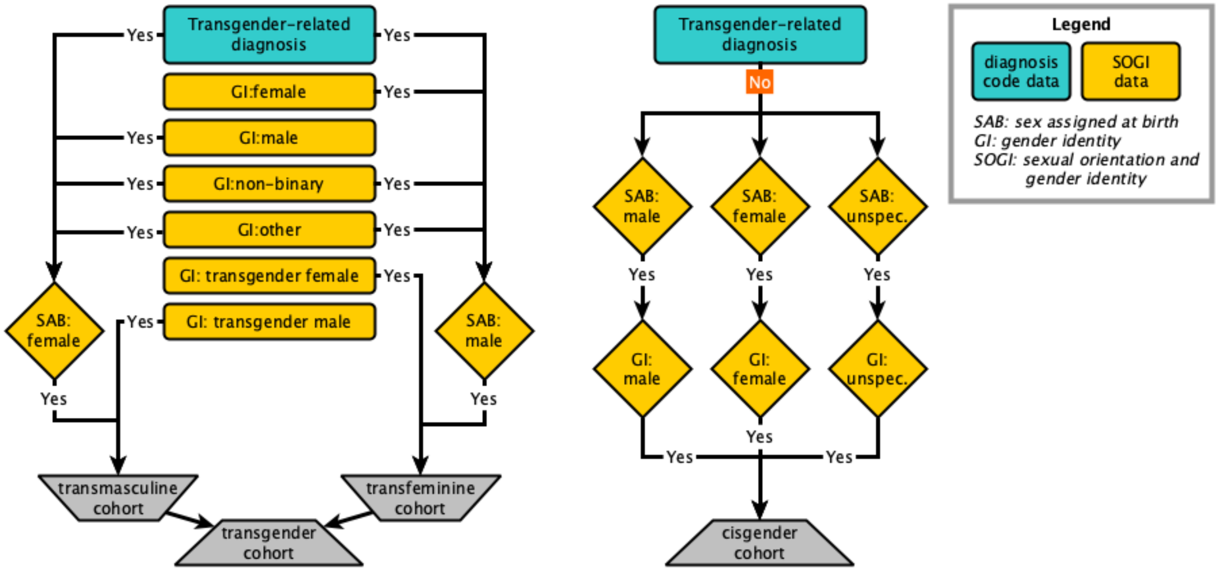

Figure 1: Flow chart depicting the algorithm for defining the gold standard corpus of patients based on diagnostic codes and SOGI data.

Model	Test Corpus	Accuracy (95% CI)	Sensitivity	Specificity	Positive predictive value	Negative predictive value
Baseline CP	GS Full	0.644 (0.550, 0.738)	0.499	0.822	0.776	0.572
Baseline CP	GS Test	0.765 (0.682,0.848)	0.529	1.000	1.000	0.680
RF CP	GS Test	0.8825 (0.860,0.903)	0.905	0.860	0.866	0.900

Table 2: Performance for the baseline and random forest (RF) computational phenotypes (CP) evaluated on the gold standard corpus. GS Full = full GS (4513 patients); GS Test = held out portion of the GS (10%, 451 patients); CI = confidence interval

Limitations to this work include the exclusion of persons under the age of 18, the reliance on data from a single academic institution, the lack of inclusion of procedure and medication codes in the baseline CP, among others, all of which speak to the potential for our methodology to not generalize to other data.

Conclusion

While further error analysis and experimentation with random forest and other ML techniques are warranted, the work presented here demonstrates the potential for machine learning CP approaches to help address the challenge of identifying TGGD persons for observational research.

Acknowledgements

This work was supported by the Health Data Compass Data Warehouse project (healthdatacompass.org). Thank you to the anonymous reviewers for their insightful comments.

References

1. McDowell A, Dowshen NL. Measuring and Addressing Inequities in Health Care Access for Transgender and Gender Diverse Youth in the U.S. J Adolesc Health. 2021 Mar 1;68(3):431–2.
2. Cathcart-Rake EJ, Kling JM, Carroll EF, Davidge-Pitts C, Le-Rademacher J, Ridgeway JL, et al. Understanding Disparities: A Case Illustrative of the Struggles Facing Transgender and Gender Diverse Patients With Cancer. J Natl Compr Canc Netw. 2023 Feb 1;21(2):227–30.
3. Bragazzi NL, Khamisy-Farah R, Converti M. Ensuring equitable, inclusive and meaningful gender identity- and sexual orientation-related data collection in the healthcare sector: insights from a critical, pragmatic systematic review of the literature. Int Rev Psychiatry. 2022 May 19;34(3–4):282–91.
4. Beltran TG, Lett E, Poteat T, Hincapie-Castillo J. The Use of Computational Phenotypes within Electronic Healthcare Data to Identify Transgender People in the United States: A Narrative Review [Internet]. Preprints; 2023 Mar [cited 2023 Jun 1]. Available from: https://www.authorea.com/users/596098/articles/629754-the-use-of-computational-phenotypes-within-electronic-healthcare-data-to-identify-transgender-people-in-the-united-states-a-narrative-review?commit=3c242a0d74f89883e3620029d11a709e709e9700
5. De Vries H, Elliott MN, Kanouse DE, Teleki SS. Using pooled kappa to summarize interrater agreement across many items. Field Methods. 2008;20(3):272–82.
6. Mamouris P, Nassiri V, Molenberghs G, van den Akker M, van der Meer J, Vaes B. Fast and optimal algorithm for case-control matching using registry data: application on the antibiotics use of colorectal cancer patients. BMC Med Res Methodol. 2021 Apr 2;21(1):62.
7. Banda JM, Halpern Y, Sontag D, Shah NH. Electronic phenotyping with APHRODITE and the Observational Health Sciences and Informatics (OHDSI) data network. AMIA Jt Summits Transl Sci Proc AMIA Jt Summits Transl Sci. 2017;2017:48–57.
8. McHugh ML. Interrater reliability: the kappa statistic. Biochem Medica. 2012 Oct 15;22(3):276–82.