The Use of Computational Phenotypes within Electronic Healthcare Data to Identify Transgender People in the United States: A Narrative Review

Theo G. Beltran[1], Elle Lett[2], Tonia Poteat[3], and Juan Hincapie-Castillo[1]

[1]The University of North Carolina at Chapel Hill Gillings School of Global Public Health
[2]Center for Applied Transgender Studies
[3]The University of North Carolina at Chapel Hill School of Medicine

March 15, 2023

Abstract

Purpose: With the expansion of research utilizing electronic healthcare data to identify transgender (TG) population health trends, the validity of computational phenotype algorithms to identify TG patients is not well understood. We aim to identify the current state of the literature that has utilized CPs to identify TG people within electronic healthcare data and their validity, potential gaps, and a synthesis of future recommendations based on past studies. **Methods:** Authors searched the National Library of Medicine's PubMed, Scopus, and the American Psychological Association Psyc Info's databases to identify studies published in the United States that applied CPs to identify TG people within electronic health care data. **Results:** Twelve studies were able to validate or enhance the positive predictive value (PPV) of their CP through manual chart reviews (n=5), hierarchy of code mechanisms (n=4), key text-strings (n=2), or self-surveys (n=1). CPs with the highest PPV to identify TG patients within their study population contained diagnosis codes and other components such as key text-strings. However, if key text-strings were not available, researchers have been able to find most TG patients within their electronic healthcare databases through diagnosis codes alone. **Conclusion:** CPs with the highest accuracy to identify TG patients contained diagnosis codes along with components such as procedural codes or key text-strings. CPs with high validity are essential to identifying TG patients when self-reported gender identity is not available. Still, self-reported gender identity information should be collected within electronic healthcare data as it is the gold standard method to better understand TG population health patterns.

INTRODUCTION

There is an estimated 2 million adults who identify as transgender (TG) living in the United States.[1] Unjust discrimination and violence have led to disproportionate health burdens among TG populations that have been consistently reported, such as higher prevalence of mental health distress, substance misuse, and HIV when compared to cisgender people (i.e. those whose sex assigned at birth aligns with their current gender identity).[2,3] While health literature on TG individuals is growing, this population is largely overlooked in epidemiologic studies due to small sample size limitations and inconsistent gender identity data collection measures.[4] Recruiting a large sample size of TG people is labor intensive and costly, leading researchers to resort to real-world data (RWD) sources like electronic healthcare databases to create efficient methods for identifying these patients.

Computational phenotypes (CPs) have become emerging tools to distinguish groups of patients with shared characteristics within electronic healthcare databases, and they have an important role in TG health-related research.[5] In brief, CPs are algorithms that use a combination of diagnostic and procedure codes, medication records, and demographic characteristics to identify patient populations within healthcare utilization data.[6] Given the varying data models from RWD sources, there is not a single standardized method to identify TG

patients. A systematic review in 2016 assessing the variation in prevalence estimates of TG people using self-reported gender identity information from surveys and TG-related diagnosis codes from electronic healthcare data across the world highlighted the lack of standardization and significant heterogeneity of ascertainment of TG status across studies as an important barrier for research.[4]

To date, there has not been a review of published literature on CPs to better understand their ability to identify TG people and their health utilization patterns within electronic health care data. Similarly, there has not been a comprehensive assessment of validity of such CP algorithms in this setting. In this narrative review, we aim to discuss the existing literature that has utilized CPs to identify TG people within electronic healthcare data and their validity, potential gaps, and a synthesis of future recommendations based on current knowledge.

METHODS

Authors searched the National Library of Medicine's PubMed, Scopus, and the American Psychological Association Psyc Info's databases to identify studies published in the United States that applied CPs to identify TG people within electronic health care data. Multiple combinations of search terms included: "transgender" "electronic health records" "computational phenotype" and "electronic medical records" (full search strategy in **Supplemental Table 1**). The electronic search included all papers published through September 2022. Our narrative review focused on research articles applying algorithms to electronic health care databases to identify TG patients. We excluded studies that used surveys, did not use data from the United States, used qualitative methodologies, or lesbian, gay, bisexual, transgender, and queer (LGBTQ) research that did not include TG people or separate their results. We excluded these studies as we wanted to focus on current measures within the United States healthcare system, where gender identity information is not often available. We also wanted to ensure United States health insurance codes were utilized, as this is an emerging area of identifying TG patients in large databases. Two reviewers (T.G.B. and J.H.C.) independently reviewed the citation index for possible inclusion, while discrepancies were resolved through consensus. Papers were reviewed and analyzed through Covidence.[7] While not a comprehensive systematic review, we follow the PRISMA statement to report our results in the spirit of transparency and reproducibility.

RESULTS

Of the 718 articles initially identified within the comprehensive search, 17 studies utilized original CPs to identify TG patients within electronic healthcare databases (**Figure 1**).[6,8–23] Twelve of these studies were able to validate or enhance the positive predictive value of their CP through manual chart reviews (n=5), hierarchy of code mechanisms (n=4), key text-strings (n=2), or self-surveys (n=1). Of these twelve studies, three used administrative claims data and nine used electronic health record (EHR) data. Claims data contains claims information on patient utilization of prescription fills and medical services for the purposes of documenting administrative and healthcare billing reimbursement, while EHR data are medical charts in a digitized format and recorded by providers for the purposes of patient clinical care.[24–26]

Administrative Claims Data and Augmented Data

Proctor and colleagues used the Centers for Medicare & Medicaid Services Chronic Conditions Data Warehouse to identify TG Medicare patients in 2013 (n=4098).[17] This study utilized a CP of ICD-9 diagnosis codes that were relevant to TG status, then subsequently addressed concerns of coding errors by validating their method through a specific logic. This logic required that patients with an initial ICD-9 diagnosis code relevant to TG status must have at least one or more of the following: more than 1 claim of an ICD-9 diagnosis code relevant to TG status in 2012, 2013, or 2014, an unspecified endocrine disorder code used by providers for TG patients to avoid TG stigma through health insurance, a sex hormone prescription in 2013, a principle diagnosis code from the ICD-9 diagnosis codes relevant to TG status, or a billing claim condition code 45 modifier or KX modifier. Proctor and colleague's CP found a sensitivity of 89.26% though their internal hierarchal method in identifying TG patients using Medicare insurance.[17] Their study was able to find 66.03% of TG Medicare patients using ICD-9 diagnosis codes relevant to TG status in 2013 alone. Approximately forty percent were identified with similar claims in 2012 and 2014, as well as through sex

hormone prescriptions or a principle diagnosis code specific to TG status.[17]

Jasuja and colleagues uses a similar approach to Proctor and colleagues by using a hierarchal method of claims data in order to validate their CP.[17,19] Jasuja and colleague's retrospective analysis of administrative claims from OptumLabs Data Warehouse from 2006 to 2017 identified 27,227 potential TG patients. They initially use gender identity disorder diagnosis codes, then incorporates endocrine non-specific codes, procedure codes relevant to TG status, then sex hormone receipts discordant with sex recorded in the claims data as a validation method to improve their accuracy of identifying their TG patients. To enhance their positive predictive value (PPV) even further, they required non-specific endocrine disorder codes to not be followed by non-diabetes codes such as thyroid or adrenal diseases and utilized a technical panel of experts to categorize a list of procedural codes that could be used for TG patients undergoing gender affirmation surgery. This study specified a minimum dosage for hormone replacement therapy to exclude non-TG patients who may receive these prescriptions at a lower dose. Using ICD-9 and ICD-10 gender identity disorder diagnosis codes alone, they were able to find 69% of TG patients. The added internal validation method of non-specific endocrine disorder codes with TG-related procedure codes added 16% TG people, and non-specific endocrine disorder codes along with sex-discordant hormone prescriptions added another 15%. They were also able to remove 1.2% of patients from the overall TG cohort after validation methods revealed their procedure or hormone prescription codes did not align with gender identity status, such as an estrogen prescription along with a transmasculine-identified procedure code (e.g., bilateral mastectomy).[19]

Yee and colleagues also used a hierarchal method approach within Oregon Medicaid claims data from 2010 to 2019 to identify TG patients with at least one ICD-9 or 10 gender identity-related diagnosis code.[18] In their approach for confirming additional details once a patient was identified as TG, they differentiated sex assigned at birth (SAB) information from self-reported gender in the enrollment file. They operationalized this by determining whether a patient's procedural or medication codes differed from their recorded SAB. As this method relied on gender identity-related diagnosis codes as entry into their study cohort, 100% of their patients had a diagnosis code. Of the 2,940 beneficiaries identified as TG, they were able to confirm 92.1% as TG using the hierarchy method described by Proctor et. al.[17] They also used a sensitivity analyses that included all changes in recorded gender and found an additional 16% transmasculine and 21% transfeminine patients.

These reviewed studies provide evidence that CPs applied in claims data with the strongest sensitivity and PPV contained ICD-9 or ICD-10 gender identity disorder codes along with additional diagnosis codes and procedural codes, although diagnosis codes were able to identify most TG patients.

Electronic Health Records and Augmented Data

Roblin and colleague's retrospective cohort study uses Kaiser Permanente Georgia's EHR system from 2006 to 2015 to potentially identify TG patients (n=271).[13] The authors describe a 3-step algorithm, which included an initial EHR search for International Classification of Disease (ICD)-9 diagnosis codes (**Supplemental Table 1**) and key text-strings relevant to TG status from supplemental digitized provider notes, validation of TG status through having at least two diagnosis codes or validation by manual review of text-strings, then determination of patient sex assigned at birth after their inclusion in the cohort. After internal validation of patients through a committee manually reviewing the key-text strings, they found that the application of key text-strings only, diagnosis codes only, and both diagnosis and key text-strings led to positive predictive values (PPV) of 45%, 56%, and 100%, respectively. A similar study by Quinn and colleagues used Kaiser Permanente's Georgia and California EHR system to identify potential TG individuals (n=6456) to build the Study of Transition, Outcomes and Gender (STRONG) cohort.[14]The study uses the same 3-step algorithm, and published their full extensive list of key text-strings. In this study, only 10% of patients were found from diagnosis codes alone, while 61% were found from both diagnosis codes and keywords. The PPV for key text-strings, diagnosis codes, and both were 26%, 54%, and 98% respectively.

Gerth and colleague's study utilizes the STRONG cohort to assess agreement between medical records and a self-reported survey.[15] The survey contained the recommended self-reported gender identity method that

asks for sex assigned at birth and current gender identity.[5,27] They distributed the survey to a subset of cohort members in order to confirm TG status (transmasculine or transfeminine) based on gender affirming treatment (e.g. testosterone, estrogen hormone therapy) and surgery (e.g. chest or genital reconstruction surgery) through Kaiser Permanente. They found high agreement between self-reported gender identity and gender affirming treatment records with a sensitivity of 99% and specificity of 99%.[15]

Guo et. al built upon Quinn and colleagues' work to apply a CP within the University of Florida Health integrated data repository which included the Epic EHR system from 2012 to 2019.[6,14]They used gender identity information, ICD-9 and ICD-10 diagnosis codes, Current Procedural Terminology (CPT) codes, and key text-strings relevant to TG status in clinician text notes as potential mechanisms to find the best performing CP for their data. Authors validated their CPs through a manual chart review of selected samples and then identified subgroups and used natal sex assignment for confirmation of transmasculine or transfeminine gender identity. Guo and colleagues found 19,600 potential TG patients and their best performing CP for both structured and unstructured data was when a TG patient had a recorded TG gender identity or had at least one relevant diagnosis code and at least one relevant key text-string relevant to TG status, which led to an F1-score of 0.954.[6]

Foer and colleagues' retrospective chart review used Epic data from two primary academic teaching institutions in Boston, MA from 2015-2019 to identify 13,424 potential TG patients.[20] They were able to utilize key text-strings within clinician notes with TG-related text, as well as F64 ICD-10 diagnosis codes, and gender identity field entries. Manual chart reviews were performed on a subset to validate the classification of patients as gold standard. They were able to find all patients through a legal sex field (100%), while sex assigned at birth was available for 48.7% of patients, and 48% had a completed gender identity field. They found 15.7% of TG patients through diagnosis and key text-strings, 89% from key text-strings alone, 14% from a gender identity field (14%), 1.2% from ICD diagnosis codes, and 5.1% from TG status listing. After validation via chart review of a subset of 324 patients, they confirmed 8% of patients as TG. 24 patients with gender fields alone were misclassified as TG when they were cisgender based on chart reviews. However, they had a high specificity after applying their algorithm to a random set of patients and found none to be TG. In this study, key text-strings and diagnosis codes were more sensitive to identify TG patients than gender related fields.[20]

Blosnich and colleagues applied a CP of ICD-9 and ICD-10 diagnosis codes relevant to TG status to identify 7560 TG patients through the US Department of Veterans Affairs Corporate Data Warehouse from 2000 to 2016.[16] Their validation method used a search algorithm of clinical text notes to find key text-strings related to TG status. Their search algorithm reached a sensitivity of 89.30%, with a specificity of 99.95%. False positives were similar to Roblin and colleagues of key text-strings that were discussions about TG relatives or friends of the patient.[13] They were also able to find false negatives through key text-strings for 1.1% of patients.[16]

Wolfe and colleagues used EHR from the Veterans Health Administration from 2006 to 2018 to create their cohort of TG veterans (n=10,769).[21] Their CP included: 1) 1 or more gender identity disorder diagnosis code in outpatient or inpatient data during the study period, 2) a diagnosis code of non-specified endocrine disorder, 3) change in sex marker field lasting at least 1 year to reflect stability, 4) sex hormone prescription discordant with sex, and 5) excluded those with specific non-diabetes endocrine code, such as adrenal or thyroid disease, and prostate cancer, as well as had minimum dosage levels for hormones. They used a hierarchal strategy that prioritized diagnosis codes or hormones, then non-specific endocrine disorder with hormone prescription, then endocrine disorders with change in sex markers, then hormone therapy with change in sex marker, to finally hormone prescriptions only, which is very similar to Jasuja et al.[19] They validated the algorithm through performing a chart review of a random sample of veterans from each of the 5 groups. Wolfe and colleagues found that TG veterans with a gender identity disorder diagnosis code had the highest positive predictive value (83%) compared to non-gender identity disorder coded veterans (2%), and concluded that gender identity disorder diagnosis codes were the most reliable approach for identification of TG patients in the VHA.[21]

Alpert and colleague's cross-sectional study utilized CancerLinQ data by the American Society of Clinical Oncology (ASCO) Learning HealthCare System to identify TG cancer patients (n=557).[22] Their CP had three categories: category 1) diagnosis related to gender identity (transsexualism or gender identity disorder); (category 2) recorded gender male with at least one diagnosis code indicating cancer of the ovaries, cervix, vulva, vagina, uterus, placenta, or other related organs; and/or (category 3) recorded gender female with at least one diagnosis code indicating cancer of the prostate, testes, penis, or other related organs. 557 individuals matched their inclusion criteria within CancerLinQ data: 42 in category 1, 316 in category 2, and 199 in category 3. 76% of those with an ICD-9 or ICD-10 diagnosis code relevant to TG status were confirmed to be TG, while only 2% and 3% were identified through categories 2 and 3, respectively.[22] There was very low specificity for categories 2 and 3, as many patients identified ended up being false positives (i.e. cisgender).

Chyten-Brennan and colleagues created a CP to identify TG patients (n=213) among people living with HIV through the Montefiore Health System in New York City from 1997 to 2017.[23] Their CP contained: 1) ICD-9 or ICD-10 diagnosis codes; 2) gender-affirming medications; 3) key text-strings, and 4) gender identity variables (e.g., yes/no field for TG). After manual chart review to validate TG status, they were able to confirm 84% of patients (PPV). Only 13.5% were identified through ICD-9 or ICD-10 diagnosis codes alone, while 60% were found from multiple categories. They were not able to confirm the TG status of 22% of those found only through ICD-9 or ICD-10 diagnosis codes. However, they were able to accurately identify 15% of TG patients through HIV-funding related gender identity data, which is not found in other EHR-based algorithms. Without this data, they would have differentially misclassified a large portion of TG people, which would lead to biased estimates.

EHR data was able to overcome the key limitation of validation for claims data by having access to conduct manual chart reviews, as well as self-reported gender identity when the data was collected and available. Similar to claim-based CPs, the strongest CPs in EHR data contained diagnosis codes accompanied by other information, which for EHR data was key text-strings relevant to TG status. If key text-strings were available, the PPV of the CP has the potential to be 100%.[13] In terms of algorithm components to identify TG patients, Wolfe et. al and Alpert et. al were able to find the highest proportion of TG patients through diagnosis codes alone.[21,22] However, Chyten-Brennan and colleagues were only able to identify 13.5% of TG patients through diagnosis codes, and Foer and colleagues found that key text-strings were able to identify almost 90% of patients.[20,23] Additionally, Chyten-Brennan and colleagues access to self-reported gender identity data added a large amount of TG patients that would have otherwise been classified as cisgender through their medical records alone.

DISCUSSION

Within a variety of validation methods, several studies have found that the CPs with the highest PPV to identify TG patients within their study population contained diagnosis codes accompanied by other information, such as procedural or prescription codes in claims data, and key text-strings in EHR data. If key text-strings were not available, most researchers have been able to find most TG patients within their electronic healthcare databases through diagnosis codes alone.

The articles reviewed contained several strengths. The use of CPs provides a low-cost, rapid approach to identify TG people who are missed by traditional structures. Roblin and colleagues provide a replicable SAS program to be applied in other healthcare systems with similar structures.[13] Proctor and colleagues offer a hierarchal framework for enhancing a CP by using additional criteria to confirm TG status, and Blosnich and colleagues validated their CP of diagnosis codes through clinician text notes within the VHA health system, which is particularly useful when there is no gold standard of self-reported gender identity.[16,17] Yee and colleagues developed a way to using differential SAB and self-reported gender to determine TG status in claims data.[18] To help improve their PPV, Jasuja and colleagues had a technical panel of experts in clinical management of TG patients decide on which procedural codes to include, and also conduct chart reviews.[19] These additional approaches added 31% of patients outside of diagnosis codes, and they found systematic differences between those found through diagnosis codes and those without. This is important for health

disparities research, as those who were older were more likely to be found without gender-identity specific diagnosis codes, and also can be used to improve the diversity and generalizability of study samples as it did for Jasuja and colleagues.[19]

Quinn and colleagues were able to use key text-strings within data structures that had access to provider notes to create one of the largest cohorts (N=6500) to date for TG people, who can be hard to recruit into cohort studies based on stigmatization or marginalization in society.[14] These key text-strings were carefully created through study stakeholders built within study design, who were also part of the TG community. Stakeholders also created a comprehensive list of hormone medications and procedures for gender affirmation, which may have helped increase the sensitivity of their CP. Guo and colleagues extended Quinn and colleague's CP by using both structured and unstructured data, which gave them the opportunity to use self-reported gender identity data when available, added ICD-10 diagnosis codes, expanded the list of key text-strings to improve sensitivity, and created an automated algorithm that does not require extensive manual chart reviews.[6] Their final reported CP was simple (at least one diagnosis code and one key text-string) and generalizable to other health systems with similar structures.

In addition, the data systems used were rich and highly detailed, such as Wolfe and colleagues paper which used the nation's largest integrated health care system.[21] Chyten-Brennan and colleagues were also able to use HIV-funded data that uniquely collected information on gender identity, further strengthening their ability to identify TG patients.[23]

However, the included studies also have limitations. Refinements of CPs are required to stay up to date on current terminology as TG terms will change over time. The most common false positive in CPs was the misuse of key text-strings by providers that were meant to discuss a relative or loved one of the patient, but not the patient themselves. Therefore, CPs must be careful to assess their validity. Further, it is difficult to validate CPs due to limited access to self-reported gender identity data, and although alternative methods were used in these studies, self-reported gender identity should be utilized as the gold standard.[5,27,28] Self-reported gender identity data can lead to bias, particularly if there are no options for those who are non-binary to accurately report their gender identity. The findings of these studies are also not generalizable to the entire TG population since not all TG patients have a TG-related diagnosis, especially those who do not seek gender affirming care. Many studies are limited to the current health care system they are using (ex. Kaiser data, Medicaid data), which will also limit generalizability of the data. In addition to this, there are lack of protections to access gender affirming health care, and not all TG people disclose their TG gender identity to their providers or surveys. This means the true prevalence of TG patients within samples may be higher due to underreporting of transgender gender identities.

Findings from the included studies have provided avenues for future research. The use of more natural language processing methods to identify nuance in CP performance are needed, especially within studies that apply key text-string methods. Papers also call for the standardization of CPs for collection at the population level, and the utilization of accessible software to apply the CP to other healthcare systems with similar data structures. All studies advocated for the incorporation of the recommended two-step method of self-reported gender identification in both EHR and claims data sources, which is still lacking in many data structures and advocated for by other reviews of TG health research.[5,29] This would also allow researchers to be able to identify transmasculine (TM) and transfeminine (TF) patients, since there are many health differences, such as differences in the need for recommended preventive screenings such as for cervical cancer.[6,19] Future work also calls for more analysis on community level differences in nomenclature and terminology related to TG people of color, and there is a need for larger ongoing longitudinal studies where data is aggregated over time and across place to assess differences between TM, TF, and TG people of color. Authors also emphasize that it is imperative for the medical community to advocate on behalf of TG patients to ensure gender-affirming medical and surgical care is protected by federal law. To ensure more holistic stories of the data are depicted, additional mixed-methods studies are needed as evidence gaps remain for contextual factors specific to the TG experience.

Quantitative evidence of CPs used to identify TG patients can have high validity when self-reported gender

identity is not available. While diagnosis codes relevant to TG status are primarily used, other forms of identification such as key text-strings and hormone prescriptions, non-specific endocrine disorder codes are useful additions to consider for researchers planning to use CPs in their TG health research.

CONCLUSION

The results of several reviewed studies show that the CPs with the highest accuracy to identify TG patients contained diagnosis codes and another component, such as and procedural and medication codes for claims data and key text-strings for EHR data. However, researchers have been able to find most TG patients within their electronic healthcare databases through diagnosis codes alone, although differences may occur depending on the data structure of the health system utilized. These findings support the conclusion that CPs are essential to identifying TG patients when self-reported gender identity is not available to understand TG population health patterns.

REFERENCES

1. We Are Here: Understanding The Size Of The LGBTQ+ Community. Published online 2021. Accessed January 10, 2022. https://hrc-prod-requests.s3-us-west-2.amazonaws.com/We-Are-Here-120821.pdf

2. Reisner SL, Poteat T, Keatley J, et al. Global health burden and needs of transgender populations: a review. *Lancet Lond Engl* . 2016;388(10042):412-436. doi:10.1016/S0140-6736(16)00684-X

3. Cisgender Definition & Meaning - Merriam-Webster. Accessed March 1, 2023. https://www.merriam-webster.com/dictionary/cisgender

4. Collin L, Reisner SL, Tangpricha V, Goodman M. Prevalence of Transgender Depends on the "Case" Definition: A Systematic Review.*J Sex Med* . 2016;13(4):613-626. doi:10.1016/j.jsxm.2016.02.001

5. Reisner SL, Deutsch MB, Bhasin S, et al. Advancing Methods for U.S. Transgender Health Research. *Curr Opin Endocrinol Diabetes Obes* . 2016;23(2):198-207. doi:10.1097/MED.0000000000000229

6. 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. *medRxiv* . Published online August 6, 2020:2020.08.04.20168161. doi:10.1101/2020.08.04.20168161

7. Covidence systematic review software. Accessed August 19, 2022. https://www.covidence.org/

8. Workman TE, Goulet JL, Brandt C, et al. A Prototype Application to Identify LGBT Patients in Clinical Notes. In: ; 2020:4270-4275. doi:10.1109/BigData50022.2020.9378109

9. McDowell A, Progovac AM, Cook BL, Rose S. Estimating the Health Status of Privately Insured Gender Minority Children and Adults.*LGBT Health* . 2019;6(6):289-296. doi:10.1089/lgbt.2018.0238

10. Obedin-Maliver J, Light A, de Haan G, Jackson RA. Feasibility of Vaginal Hysterectomy for Female-to-Male Transgender Men. *Obstet Gynecol* . 2017;129(3):457-463. doi:10.1097/AOG.0000000000001866

11. Dragon CN, Laffan AM, Erdem E, et al. Health Indicators for Older Sexual Minorities: National Health Interview Survey, 2013-2014.*LGBT Health* . 2017;4(6):398-403. doi:10.1089/lgbt.2016.0203

12. Progovac AM, Cook BL, Mullin BO, et al. Identifying Gender Minority Patients' Health And Health Care Needs In Administrative Claims Data.*Health Aff Proj Hope* . 2018;37(3):413-420. doi:10.1377/hlthaff.2017.1295

13. Roblin D, Barzilay J, Tolsma D, et al. A novel method for estimating transgender status using electronic medical records. *Ann Epidemiol* . 2016;26(3):198-203. doi:10.1016/j.annepidem.2016.01.004

14. Quinn VP, Nash R, Hunkeler E, et al. Cohort profile: Study of Transition, Outcomes and Gender (STRONG) to assess health status of transgender people. *BMJ Open* . 2017;7(12). doi:10.1136/bmjopen-2017-018121

15. Gerth J, Becerra-Culqui T, Bradlyn A, et al. Agreement between Medical Records and Self-Reports: Implications for Transgender Health Research. *Rev Endocr Metab Disord* . 2018;19(3):263-269. doi:10.1007/s11154-018-9461-4

16. Blosnich JR, Cashy J, Gordon AJ, et al. Using clinician text notes in electronic medical record data to validate transgender-related diagnosis codes. *J Am Med Inform Assoc JAMIA* . 2018;25(7):905-908. doi:10.1093/jamia/ocy022

17. Proctor K, Haffer SC, Ewald E, Hodge C, James CV. Identifying the Transgender Population in the Medicare Program. *Transgender Health* . 2016;1(1):250-265. doi:10.1089/trgh.2016.0031

18. Yee K, Lind BK, Downing J. Change in Gender on Record and Transgender Adults' Mental or Behavioral Health. *Am J Prev Med* . Published online December 14, 2021. doi:10.1016/j.amepre.2021.10.016

19. Jasuja GK, de Groot A, Quinn EK, et al. Beyond Gender Identity Disorder Diagnosis Codes: An Examination of Additional Methods to Identify Transgender Individuals in Administrative Databases. *Med Care* . 2020;58(10):903-911. doi:10.1097/MLR.0000000000001362

20. Foer D, Rubins DM, Almazan A, Chan K, Bates DW, Hamnvik OPR. Challenges with Accuracy of Gender Fields in Identifying Transgender Patients in Electronic Health Records. *J Gen Intern Med* . 2020;35(12):3724-3725. doi:10.1007/s11606-019-05567-6

21. Wolfe HL, Reisman JI, Yoon SS, et al. Validating Data-Driven Methods for Identifying Transgender Individuals in the Veterans Health Administration of the US Department of Veterans Affairs. *Am J Epidemiol* . 2021;190(9):1928-1934. doi:10.1093/aje/kwab102

22. Alpert AB, Komatsoulis GA, Meersman SC, et al. Identification of Transgender People With Cancer in Electronic Health Records: Recommendations Based on CancerLinQ Observations. *JCO Oncol Pract* . 2021;17(3):e336-e342. doi:10.1200/OP.20.00634

23. Chyten-Brennan J, Patel VV, Ginsberg MS, Hanna DB. Algorithm to identify transgender and gender nonbinary individuals among people living with HIV performs differently by age and ethnicity. *Ann Epidemiol* . 2021;54:73-78. doi:10.1016/j.annepidem.2020.09.013

24. Strom BL. Overview of Electronic Databases in Pharmacoepidemiology. In: *Pharmacoepidemiology* . ; 2019:203-210. doi:10.1002/9781119413431.ch11

25. Gerhard T, Moride Y, Pottegård A, Pratt N. Encounter Databases. In:*Pharmacoepidemiology* . ; 2019:211-240. doi:10.1002/9781119413431.ch12

26. Horton DB, Bhullar H, Carty L, et al. Electronic Health Record Databases. In: *Pharmacoepidemiology* . ; 2019:241-289. doi:10.1002/9781119413431.ch13

27. The GenIUSS Group. Best practices for asking questions to identify transgender and other gender minority Respondents on population-based surveys. Published online 2014. https://escholarship.org/content/qt3qk7s1g6/qt3qk7s1g6.pdf#page=27

28. Deutsch MB, Green J, Keatley J, Mayer G, Hastings J, Hall AM. Electronic medical records and the transgender patient: recommendations from the World Professional Association for Transgender Health EMR Working Group. *J Am Med Inform Assoc JAMIA* . 2013;20(4):700-703. doi:10.1136/amiajnl-2012-001472

29. Tate CC, Ledbetter JN, Youssef CP. A two-question method for assessing gender categories in the social and medical sciences. *J Sex Res* . 2013;50(8):767-776. doi:10.1080/00224499.2012.690110

Figure 1. PRISMA Flow Diagram of Narrative Review of Transgender Computational Phenotype Literature