Journal of the American Medical Informatics Association, 25(7), 2018, 905–908
doi: 10.1093/jamia/ocy022
Advance Access Publication Date: 4 April 2018
Brief Communication

Brief Communication

Using clinician text notes in electronic medical record data to validate transgender-related diagnosis codes

John R Blosnich,[1] John Cashy,[1] Adam J Gordon,[1,2,3] Jillian C Shipherd,[4,5,6] Michael R Kauth,[4,7,8,9] George R Brown,[10,11] and Michael J Fine[1,2]

[1]Center for Health Equity Research and Promotion, VA Pittsburgh Healthcare System, Pittsburgh, PA, 15240, USA, [2]Division of General Internal Medicine, School of Medicine, University of Pittsburgh, Pittsburgh, PA, 15213, USA, [3]MIRECC, VA Pittsburgh Healthcare System, Pittsburgh, PA, 15213, USA, [4]LGBT Health Program, Office of Patient Care Services, Department of Veterans Affairs, Washington, DC, 20420, USA, [5]VA Boston Healthcare System, National Center for PTSD, Women's Health Sciences Division, Boston, MA, 02130, USA, [6]Department of Psychiatry, Boston University, Boston, MA, 02118. USA, [7]South Central MIRECC, Michael E. DeBakey VA Medical Center, Houston, TX, 77030, USA, [8]Department of Psychiatry and Behavioral Sciences, Baylor College of Medicine, Houston, TX, 77030, USA, [9]Houston VA HSR&D Center for Innovations in Quality, Effectiveness and Safety, Houston, TX, 77021, USA, [10]Department of Psychiatry and Behavioral Sciences, East Tennessee State University, Johnson, TN, 37604, USA and [11]Mountain Home VA Medical Center, Mountain Home, TN, 37684, USA

Corresponding Author: John R Blosnich, Department of Veterans Affairs, VA Pittsburgh Healthcare System, Center for Health Equity Research and Promotion, University Drive C (151C-U), Building 30, Pittsburgh, PA, 15240-1001, USA; john.blosnich@va.gov

Received 13 July 2017; Revised 8 January 2018; Editorial Decision 23 February 2018; Accepted 26 February 2018

ABSTRACT

Objective: Transgender individuals are vulnerable to negative health risks and outcomes, but research remains limited because data sources, such as electronic medical records (EMRs), lack standardized collection of gender identity information. Most EMR do not include the gold standard of self-identified gender identity, but International Classification of Diseases (ICDs) includes diagnostic codes indicating transgender-related clinical services. However, it is unclear if these codes can indicate transgender status. The objective of this study was to determine the extent to which patients' clinician notes in EMR contained transgender-related terms that could corroborate ICD-coded transgender identity.

Methods: Data are from the US Department of Veterans Affairs Corporate Data Warehouse. Transgender patients were defined by the presence of ICD9 and ICD10 codes associated with transgender-related clinical services, and a 3:1 comparison group of nontransgender patients was drawn. Patients' clinician text notes were extracted and searched for transgender-related words and phrases.

Results: Among 7560 patients defined as transgender based on ICD codes, the search algorithm identified 6753 (89.3%) with transgender-related terms. Among 22 072 patients defined as nontransgender without ICD codes, 246 (1.1%) had transgender-related terms; after review, 11 patients were identified as transgender, suggesting a 0.05% false negative rate.

Conclusions: Using ICD-defined transgender status can facilitate health services research when self-identified gender identity data are not available in EMR.

INTRODUCTION

Transgender individuals—people for whom sex assigned at birth is inconsistent with gender identity—are vulnerable to health and heath care inequities,[1] including HIV infection,[2] self-directed violence,[3,4] and trauma[5] and discrimination.[6] Approximately 1.4 million adults in the US identify as transgender,[7] but research remains limited due to scant data sources that include gender identity.

Electronic heath records (EHRs) data provide opportunities to gather health information about transgender individuals.[8–10] Because of its focus on the ecology of information across the medical setting—from provider psychology, to documenting processes, to infrastructural resources[11]—medical informatics provides critical lenses through which to view transgender health research in EHR. More specifically, from a public health informatics level,[12,13] identification of specific patient populations of interest (e.g., minority patient populations) is necessary to initiate any research endeavor, such as health services research or disease surveillance. Health services research about transgender populations has been considerably limited because of the first conundrum of how best to elucidate the research population of interest. Currently, there are two primary ways transgender populations can be identified in EHR.

First, there is a "gold standard" of self-identified gender identity, i.e., having patients report their own gender identity, which is ideal because it would provide the clearest indication of defining the population of interest. Currently Fenway Health, a large community health center in Boston, is one of the few healthcare centers that collects patients' self-reported gender identity, which has facilitated health services research studies of transgender patients.[14–16] These studies, though using the gold standard, focus on a limited sample of predominantly treatment-seeking urban transgender individuals. Unfortunately, few healthcare systems collect self-identified gender identity. Despite initial interest by the Institute of Medicine to recommend collection of gender identity in EHRs, it did not matriculate to the final list of recommended elements.[17] Thus, researchers using EHR data for transgender health have had to think creatively to find transgender patients.

In the absence of the gold standard, a second method is using International Classification of Diseases (ICD) diagnostic codes related to transgender status (e.g., gender dysphoria [GD]). This method is particularly prone to questions germane to the medical informatics field because ICD data are distilled through many filters (e.g., a provider's clinical judgment and the processes of translating and documenting that judgment into a diagnostic code). Consequently, it is unclear to what extent ICD documentation can indicate a patient's transgender status.

ICD 9 documentation of transgender status has been used in EHR studies within large healthcare systems (e.g., Department of Veterans Affairs [VA][3] and the Centers for Medicaid and Medicare Services),[18] which have large patient populations but lack self-identified gender identity data. To operationalize transgender identity, VA studies used ICD diagnosis codes related to Gender Identity Disorder (GID) or GD.[3,19,20] In addition to unknown validity, ICD-defined transgender identity likely underestimates transgender populations because not all transgender individuals have ICD codes for GID or GD.[6] Because insurers use ICD codes, confirming accuracy of ICD codes to identify transgender populations would assist policy makers and researchers in accurately identifying health care needs of transgender populations.

From a medical informatics perspective, clinician chart notes represent a data source that may indicate a provider's cognitive rationale for diagnostic judgment or may offer narrative evidence that would support inference of a patient's gender identity. Roblin et al studied EHR-based documentation of transgender status using ICD and clinical text note data from Kaiser Permanente Georgia, finding only 40% of their sample had both ICD codes and 1 of 6 transgender-related terms. As a foundational study into *how* transgender patients could be identified (i.e., text, ICD codes, or both), because the analyses focused solely on transgender patients, it remains unclear the extent to which clinical text notes validate transgender-related ICD codes when trying to discern transgender and nontransgender patients. Moreover, it is unclear how findings may be replicated and may differ between a private vs federally subsidized healthcare system.

In the absence of self-identified gender identity in VA data, the aim of this study was to determine the extent to which patients' clinician notes contained transgender-related terms that could corroborate a patient's ICD-coded transgender identity. We hypothesized that: (1) >50% of transgender VA patients (i.e., patients with GID or GD ICD codes) would have transgender-related terms in their clinical progress notes, and (2) <1% of a sample of nontransgender VA patients (i.e., patients without GID or GD ICD codes) would have transgender-related terms in their clinician notes.

METHODS

The VA maintains a consolidated architecture for its EHR through its Corporate Data Warehouse (CDW). The current investigation focused on data from all inpatient and outpatient visits within the VA between fiscal years (FY) 2000–2016. In October 2015, the VA began to transition its coding from the ICD, 9th Revision (ICD9) to the ICD10 catalogue of diagnostic codes, so this project incorporated both ICD catalogues. More detailed information about VA's CDW has been published previously.[21] Although the VA collects biological sex of all patients as either male or female, there currently is not standardized data collection about patients' gender identity.

Consistent with prior research about transgender veterans,[4,19] transgender identity was defined by the presence of any one of several ICD9 and ICD10 codes associated with transgender-related clinical services (see Table 1). For each transgender patient, a comparison group of three patients without transgender-related codes (i.e., nontransgender veterans) was randomly drawn from the same VA Medical Center and FY of the transgender patient's index diagnosis with a transgender-related code. To be included in the study, the patients had to have at least one inpatient or outpatient visit that contained clinician notes.

Full clinician notes from inpatient and outpatient visits for the analytic sample were extracted from the CDW. Among the analytic sample, nearly all (95%) patients' clinician notes contained less than 7000 characters total; the maximum was 450 000 characters. Analyses were performed with R statistical software[22] and the tm text mining package.[23] The search algorithm was conducted in 2 phases. The first phase included only patients with transgender-related ICD codes because they constituted the group in which terms or phrases were most likely to appear. Punctuation, numbers, and common stop words (e.g., "and") were removed from the text of the progress notes. Next, a list was generated containing the frequencies of all words in the progress notes. The first author reviewed the list of terms and identified any variants and misspellings of terms such as "transgender" and "trans-gender," or "transsexual," "transexual," and "trans-sexual," which were incorporated into building the search algorithm (See supplementary Table S2 for syntax).

In the second phase, the search algorithm was conducted on the groups of patients with and without transgender-related ICD codes.

Table 1. ICD-9 and ICD-10 Diagnoses Related to Transgender Status

Diagnosis name	ICD-9	ICD-10
Transvestic fetishism	302.3	F65.1
Gender identity disorder in adolescence and adulthood	302.85	F64.1
Gender identity disorder of childhood	302.6	F64.2
Other gender identity disorder	N/A	F64.8
Gender identity disorder, unspecified	N/A	F64.9
Transsexualism	302.5	F64.0
Personal history of sex reassignment	N/A	Z87.890

N/A denotes that the code was listed in ICD-10 but did not exist in ICD-9.

A regular expression matching was performed to identify text patterns that would indicate transgender identity, such as "trans" closely followed (e.g., within three words) by "sex" or "gender" or "vest," or "gender" closely followed by "assign" or "confus" (e.g., confusion, confused) or "disorder." Any term or phrase matches among the group of patients without transgender-related ICD codes were examined in the full context of the note by the first author to determine whether they indicated transgender identity.

Among the transgender group, term or phrase matches were deemed "true positives." Among the nontransgender group, term or phrase matches that were reviewed and considered valid were deemed "false negatives." This study was approved by the institutional review board of (institution name masked for peer-review).

RESULTS

We identified 7643 unique patients with one or more transgender-related diagnosis code(s), of which 83 (1.1%) had no text notes and could not be included in the study. Among these 7560 transgender patients, the search algorithm identified transgender-related terms in the progress notes of 6753 patients (89.3%), considered true positives. Words and phrases identified were fairly clear, (e.g., "she is transgender," "transgender male," "going to last phase of sex reassignment").

We identified 22 929 patients without transgender-related codes, of which 857 (3.7%) had no text notes and were excluded from the comparison group. Among these 22 072 patients, the algorithm identified 246 (1.1%) patients with one or more transgender-related terms in their notes (i.e., false negatives). Of these 246 patients, 113 (45.9%) had the word "transgender" appear because of wording in a demographic template created by select VA Medical Centers (e.g., "Is the Veteran lesbian, gay, bisexual, or transgender"). For 100 (40.7%) patients, the terms or phrases were flagged in relation to either medical shorthand notations (e.g., using "trans exam" for transillumination exam) or misspellings (e.g., "gasexchange" instead of "gas exchange," which was flagged as matching the search phrase "sex change"). Another 22 cases (8.9%) were identified due to the patient discussing another person's transgender identity (e.g., spouse, child). The remaining 11 cases (4.5%) appeared to be language clearly indicating that transgender-related terms were used in regard to the transgender identity of the patient. Consequently, the false negative rate was 0.05% (11 of 22 072).

DISCUSSION

The results supported our initial hypotheses; in fact, resulting in a higher percentage (89.3%) of transgender patients with transgender related-terms. Our findings also support Roblin et al,[8] suggesting the validity of ICD codes to identify samples of transgender patients.

The results suggest high sensitivity, however there are logistical and cultural caveats to using ICD codes to define transgender identity in EHR studies. Logistically, there are many reasons that a transgender patient may not have an ICD code for GID or GD. For instance, a transgender patient may elect to receive their transgender-related care (e.g., cross-sex hormone prescriptions) outside of the VA but come to the VA for nontransgender related care. Correspondingly, if a transgender patient is not experiencing "clinically significant distress or impairment in social, occupational, or other important areas of functioning," as outlined in the definition of GD, a diagnosis may not appear.[24] Culturally, the medicalization of transgender identities requires a diagnosis to access treatments, which can be stigmatizing because it pathologizes gender identity.[25] Consequently, further research exploring clinicians' beliefs and cognition on decision making processes involved in constructing notes for transgender patients, language choice, or how patient interactions affect their note-writing would illuminate data input processes for future informatics-related research. Further research is also needed to explore whether type and prevalence of ICD codes may be associated with how notes are constructed.

Importantly, the issue of false positives (e.g., 10.7% of persons with transgender-related ICD codes did not have transgender-related terms) remains complex. Some clinicians may have purposefully kept transgender-related terms out of the notes at the behest of their patients or to protect their patients' privacy. Thus, it is difficult to interpret absence of a term as evidence of a false positive. Further research is needed to develop more sophisticated methods to make these determinations.

Several limitations should be noted. First, text notes were composed by clinicians and cannot be interpreted as a patient's self-identified transgender identity or disclosure on behalf of the patient. Relatedly, although we refined our search algorithm, it may have missed cases due to variations (e.g., idiosyncratic provider shorthand) for which we could not account. Second, ICD codes are used to characterize the services delivered to the patient, which are not necessarily the products of a diagnostic assessment by a mental health professional (i.e., ICD codes can be used without a diagnostic assessment). Additionally, the transgender-related ICD codes do not match the Diagnostic and Statistical Manual from which US mental health professionals make psychiatric diagnoses. Thus, it is important to note that ICD codes do not necessarily imply diagnoses from psychological assessments. Third, the VA has recently undertaken several systemic initiatives to improve transgender health care,[26] which may reduce the generalizability of this study to other US healthcare systems. Fourth, 4.8% of the original sample were excluded because they lacked clinician notes, which could have been due to several scenarios, such as the patient enrolled for care and listed as a patient but did not show up for their appointment. Lastly, this study was not a true sensitivity/specificity study because it lacked the gold standard of self-identity; however, it incorporated an innovative, reproducible method of using clinical text notes as an expeditious proxy verification of ICD-defined transgender identity.

CONCLUSION

This study offers evidence of validity in using ICD-defined transgender identity to facilitate health services research. Until health systems include standardized self-identified gender identity information in their data collection systems,[27] researchers using ICD codes to

identify transgender patients may consider using clinician text notes as an additional step when defining patient samples.

FUNDING

This work was supported by a Career Development Award (CDA 14-408) from the Department of Veterans Affairs, Health Service Research & Development to JRB.

COMPETING INTERESTS

The authors have no financial interests to disclose.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

ACKNOWLEDGMENTS

This work was supported using resources and facilities at the VA Informatics and Computing Infrastructure (VINCI), VA HSR RES 13-457. The authors thank Allen Faler with VINCI for his assistance with data extraction and management. The opinions expressed in this work are those of the authors and do not necessarily reflect those of the funders, institutions, the Department of Veterans Affairs, or the United States Government.

REFERENCES

1. Institute of Medicine. *The Health of Lesbian, Gay, Bisexual, and Transgeder People: Building a Foundation for Better Understanding*. Washington DC: Institute of Medicine; 2011.
2. Baral SD, Poteat T, Strömdahl S, Wirtz AL, Guadamuz TE, Beyrer C. Worldwide burden of HIV in transgender women: a systematic review and meta-analysis. *Lancet Infect Dis* 2013; 13 (3): 214–222.
3. Blosnich JR, Brown GR, Shipherd JC, Kauth M, Piegari RI, Bossarte RM. Prevalence of gender identity disorder and suicide risk among transgender veterans utilizing veterans health administration care. *Am J Public Health*. 2013; 103 (10): e27–32.
4. Blosnich JR, Brown GR, Wojcio S, Jones KT, Bossarte RM. Mortality among veterans with transgender-related diagnoses in the Veterans Health Administration, FY2000-2009. *LGBT Health* 2014; 1 (4): 269–276.
5. Shipherd JC, Maguen S, Skidmore WC, Abramovitz SM. Potentially traumatic events in a transgender sample: frequency and associated symptoms. *Traumatology* 2011; 17 (2): 56–67.
6. Grant JM, Mottet LA, Tanis J, Herman JL, Harrison J, Keisling M. *National Transgender Discrimination Survey Report on Health and Health Care*. Washington, DC: National Center for Transgender Equality and the National Gay and Lesbian Task Force; 2010.
7. Flores AR, Herman JL, Gates GJ, Brown TNT. *How Many Adults Identify as Transgender in the United States?* Los Angeles, CA: The Williams Institute; 2016.
8. 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.
9. Deutsch MB, Green J, Keatley J, *et al*. Electronic medical records and the transgender patient: recommendations from the World Professional Association for Transgender Health EMR Working Group. *J Am Med Inform Assoc* 2013; 20 (4): 700–703.
10. Deutsch MB, Buchholz D. Electronic health records and transgender patients—practical recommendations for the collection of gender identity data. *J Gen Int Med* 2015; 30 (6): 843.
11. Greenes RA, Shortliffe EH. Medical informatics: an emerging academic discipline and institutional priority. *JAMA*. 1990; 263 (8): 1114–1120.
12. Hersh WR. Medical informatics: improving health care through information. *JAMA*. 2002; 288 (16): 1955–1958.
13. Kukafka R. Public health informatics: the nature of the field and its relevance to health promotion practice. *Health Promotion Pract* 2005; 6 (1): 23–28.
14. Reisner SL, White JM, Bradford JB, Mimiaga MJ. Transgender health disparities: comparing full cohort and nested matched-pair study designs in a community health center. *LGBT Health* 2014; 1 (3): 177–184.
15. Reisner SL, White JM, Mayer KH, Mimiaga MJ. Sexual risk behaviors and psychosocial health concerns of female-to-male transgender men screening for STDs at an urban community health center. *AIDS Care* 2014; 26 (7): 857–864.
16. Reisner SL, Vetters R, Leclerc M, *et al*. Mental health of transgender youth in care at an adolescent urban community health center: a matched retrospective cohort study. *J Adolesc Health* 2015; 56 (3): 274–279.
17. Institute of Medicine. *Capturing Social and Behavioral Domains in Electronic Health Records: Phase 1*. Washington, DC: The National Academies Press; 2014.
18. 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.
19. Kauth MR, Shipherd JC, Lindsay J, Blosnich JR, Brown GR, Jones KT. Access to care for transgender veterans in the Veterans Health Administration: 2006-2013. *Am J Public Health* 2014; 104 (Suppl 4): S532–534.
20. Brown GR, Jones KT. Mental health and medical health disparities in 5135 transgender veterans receiving healthcare in the veterans health administration: a case-control study. *LGBT Health* 2016; 3 (2): 122–131.
21. Fihn SD, Francis J, Clancy C, *et al*. Insights from advanced analytics at the Veterans Health Administration. *Health Affairs* 2014; 33 (7): 1203–1211.
22. R Core Team. R: A language and environment for statistical computing. *[computer program]*. Vienna, Austria: R Foundation for Statistical Computing; 2013.
23. Feinerer I. *Introduction to the TM Package Text Mining in R*. 2013. ftp://videolan.cs.pu.edu.tw/network/CRAN/web/packages/tm/vignettes/tm.pdf. Accessed June 8, 2017.
24. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders: DSM-5*. Washington, DC: American Psychiatric Association; 2013.
25. Hughto JMW, Reisner SL, Pachankis JE. Transgender stigma and health: a critical review of stigma determinants, mechanisms, and interventions. *Soc Sci Med* 2015; 147: 222–231.
26. Kauth MR, Shipherd JC. Transforming a system: improving patient-centered care for sexual and gender minority veterans. *LGBT Health* 2016; 3 (3): 177–179.
27. Collin L, Reisner SL, Tangpricha V, Goodman M. Prevalence of transgender depends on the "case" definition: a systematic review. *J Sexual Med* 2016; 13 (4): 613–626.