Reduced Accuracy of Intake Screening Questionnaires Tied to Quality Metrics

Jodi Simon, DrPH, MS[1]
Jeffrey Panzer, MD, MS[1,2]
Katherine M. Wright, MPH, PhD[3]
Abbey Ekong, MA[1]
Patrick Driscoll, RN, BSN, MPH[1]
Nivedita Mohanty, MD, MS[1,3]
Christine A. Sinsky, MD[4]

[1]AllianceChicago, Chicago, Illinois

[2]Tapestry 360 Health, Chicago, Illinois

[3]Northwestern University Feinberg School of Medicine, Chicago, Illinois

[4]American Medical Association, Chicago, Illinois

ABSTRACT

Clinical workflows that prioritize repetitive patient intake screening to meet performance metrics may have unintended consequences. This retrospective analysis of electronic health record data from 24 Federally Qualified Health Centers assessed effectiveness and accuracy of the 2-item Patient Health Questionnaire (PHQ-2) for depression screening and Generalized Anxiety Disorder 2 (GAD-2) for anxiety screening from 2019 to 2021. Scores of over 91% of PHQ-2 and GAD-2 tests indicated low likelihood of depression or anxiety, which diverged markedly from published literature on screening outcomes. Visit-based screenings linked to performance metrics may not be delivering the intended value in a real-world setting and risk distracting clinical effort from other high value activities.

Ann Fam Med 2023;21:444-447. https://doi.org/10.1370/afm.3019

INTRODUCTION

Primary care visits often start with a myriad of standardized intake screening questions that are tied to performance metrics and incorporated into electronic health records (EHRs). Prioritizing repetition of intake screening questionnaires at primary care visits may have unintended consequences such as administrative burden, provision of low-value care, and reduced clinical capacity to deliver other, high-value services.[1]

Prior work demonstrated high levels of repetition of 6 intake screening questionnaires tied to performance metrics (ie, Patient Health Questionnaire-2 [PHQ-2], tobacco use screening, etc) during visits to 25 Federally Qualified Health Centers (FQHCs) in 2019.[2] The current study extends this research by exploring the accuracy and utility of 2 of these validated questionnaires (PHQ-2, Generalized Anxiety Disorder 2 [GAD-2]) to better understand if they provide the expected value in real-world settings.

METHODS

We analyzed EHR data to (1) compare rates of positive PHQ-2 and GAD-2 tests administered within our study population to publicly available US Census data and published literature, and (2) to assess the accuracy of these instruments by comparing the PHQ-2 and GAD-2 scores to diagnoses for corresponding patients. The study population included patients aged 18 years and older with at least 1 visit between 2019 and 2021 to 1 of 24 FQHCs (spanning 11 states). The 2 questionnaires were selected because they are widely implemented at the FQHCs and are linked to performance metrics for the National Committee for Quality Assurance Patient-Centered Medical Home recognition[3] and/or the Health Resources and Services Administration's Uniform Data System[4] and they are embedded into the intake form of the EHR. Questionnaires are predominately administered verbally during the intake process by medical assistants.

To make our results comparable to the US Census Bureau's 2021 Household Pulse Survey (HPS), we applied HPS sample weights to generate nationally representative estimates of adults experiencing symptoms of depression and anxiety as measured by the PHQ-2 and GAD-2.[5]

To assess accuracy, we examined score distributions for PHQ-2 and GAD-2 screenings completed by patients with subsequent new evidence of depression and anxiety (delineated as a new diagnosis in the EHR). We compared the ability of

Conflicts of interest: authors report none. Dr Sinsky is employed by the American Medical Association.

CORRESPONDING AUTHOR

Jodi Simon
AllianceChicago
225 W. Illinois Street, 5th Floor
Chicago, IL 60654
jsimon@alliancechicago.org

the screeners to detect disease to sensitivity rates in published literature.

This study was granted an exemption from review by the Chicago Department of Public Health Institutional Review Board.

RESULTS

Screenings, including 1,883,317 PHQ-2s and 1,573,107 GAD-2s, were performed on 380,057 patients. Of these, 92.3% (1,738,534/1,883,317) of PHQ-2 tests and 91.4% (1,437,234/1,573,107) of GAD-2 tests resulted in a cumulative score of 0 or 1, indicating low likelihood of depression (for PHQ-2) and anxiety (for GAD-2) (Figure 1). The mean (SD) PHQ-2 score was 0.29 (1.024). The mean (SD) GAD-2 score was 0.35 (1.193). The median (interquartile range [IQR]) was 0.00 (0.00-0.00) for both instruments. Score distributions show 11% of patients had a positive PHQ-2 score (≥2) on their first screen, compared with 26% to 43% of first screens in the literature[6-9] and census data sets[5] (Figure 2). Similarly, score distributions show 11% of patients had a positive GAD-2 score (≥2) on their first screen, compared with 47% to 53% in census data sets[5] and previous literature.[10]

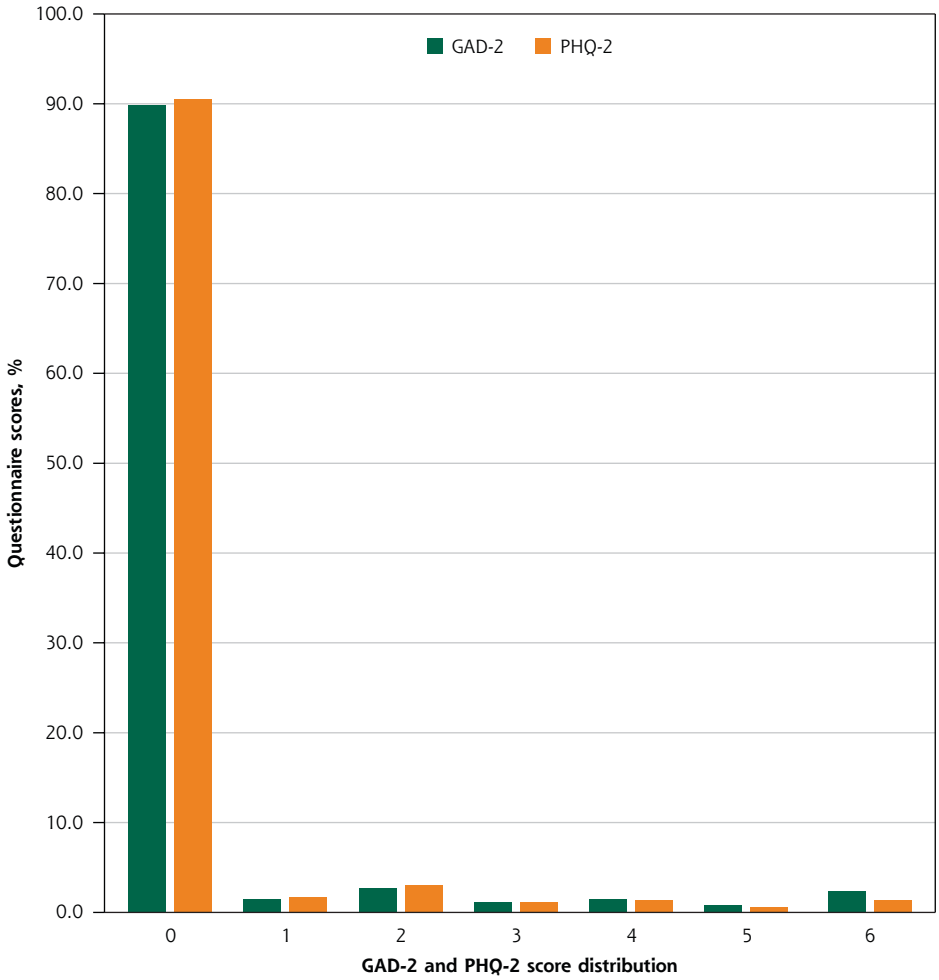
Narrowing the analysis to patients with new diagnoses (excluding patients without a diagnosis or with a prior diagnosis), we found 42.3% (10,624/25,116) of patients with a new depression diagnosis scored 0 or 1 on the PHQ-2 within the previous 30 days. Of patients with a new anxiety diagnosis, 42.7% (16,272/38,127) scored 0 or 1 on the GAD-2. Said another way, screening only detected risk in 57.7% of patients subsequently diagnosed with depression and 57.3% of patients subsequently diagnosed with anxiety.

DISCUSSION

Our prior study demonstrated that intake screening questionnaires during primary care visits in FQHCs are often administered repetitively in order to meet performance metrics.[2] The current results suggest that existing workflows for screening are also less effective in detecting depression and anxiety than expected. In this real-world setting, PHQ-2 and GAD-2 results were more frequently negative (normal) when compared with settings described in published literature and census data. Although FQHC patients may differ from those in the literature and census data, these differences are unlikely to account for this disparity. In fact, the patients we studied are likely to have a relatively high prevalence of depression and anxiety because FQHC patients are predominantly low income[11,12] and because the study period overlapped with the COVID-19 pandemic.[13,14]

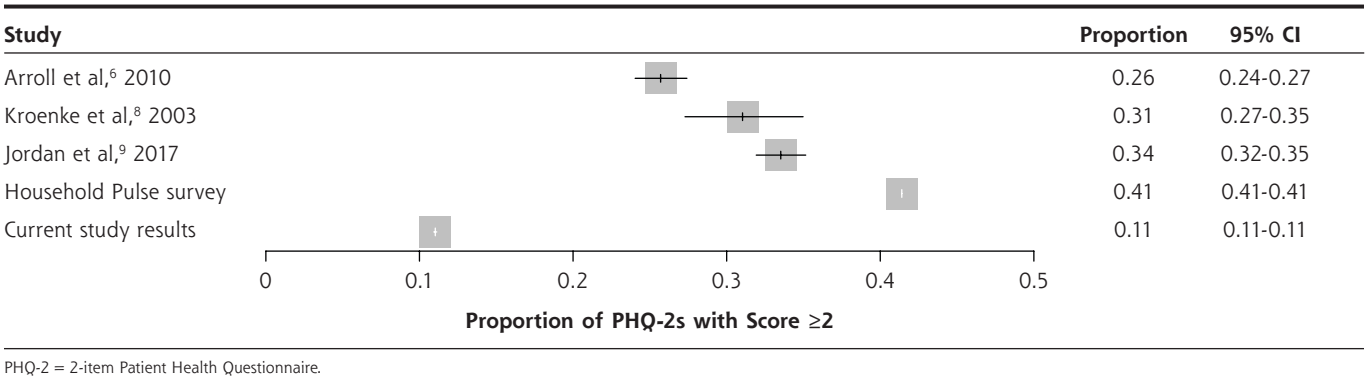
We also evaluated PHQ-2 and GAD-2 results in patients who develop new diagnoses of depression or anxiety. In these patients, the PHQ-2 and GAD-2 had disease detection rates of less than 60%, compared with 90+% sensitivity in published literature.[6-8] We acknowledge that documentation on a diagnosis list in an EHR is not gold standard proof that the patient has depression or anxiety. Nonetheless, low positivity (<60%) in a screening test among patients diagnosed within 30 days of screening warrants further exploration.

Figure 1. GAD-2 and PHQ-2 score distributions.



GAD-2 = Generalized Anxiety Disorder 2 questionnaire; PHQ-2 = 2-item Patient Health Questionnaire.

Figure 2. Comparison of positive PHQ-2 rates.



PHQ-2 = 2-item Patient Health Questionnaire.

These results raise the possibility that when done frequently to meet performance thresholds, such screenings may be performed in a perfunctory or inconsistent manner that reduces sensitivity. Preliminary qualitative findings based on structured interviews with clinicians, staff, and patients demonstrate variation in questionnaire administration and time constraints as underlying factors leading to inaccuracies, but future, more comprehensive work in this area is needed.

The US Preventive Services Task Force (USPSTF) recently issued draft recommendations that primary care clinicians screen all adults aged <65 years for anxiety. The recommendations state that "more studies are needed on the diagnostic accuracy of screening tools that are feasible for use in primary care."[15] Our findings indicate potentially compromised accuracy of anxiety and depression screeners when their implementation is driven by a need to meet performance measures and they are embedded into EHRs and visit workflows. Some improvement suggestions are to screen at predetermined intervals rather than at every clinical encounter and to rely on self-administration methods, either electronic or paper, which may have higher fidelity and reliability[16] and cause less burden to staff and patients.

Our study has broad relevance for policy makers, regulators, measure developers, and clinician organizations that extends beyond depression and anxiety screening. Focusing on incentivized process measures like intake screening questionnaires leads to repetitive[2] and, we hypothesize, inaccurate completion. The impact on outcomes that matter (ie, reducing mortality and morbidity from depression and anxiety) may not be as favorable as previously perceived, and ineffective screening may unintentionally detract from clinical care because care teams and patients have less time and cognitive energy to focus on other priorities during busy clinical encounters. The importance of not confusing metrics with objectives ("surrogation") is described in the Harvard Business Review article "Don't Let Metrics Undermine Your Business."[17] Our findings suggest similar wisdom could be useful in health care, given the implementation of care processes like depression and anxiety screening to meet a performance metric may inadvertently lead to reduced accuracy and low-value care.

Read or post commentaries in response to this article.

Key words: performance measures; health care quality; administrative burden; practice-based research; PHQ-9; quality improvement; physician burnout

Submitted October 25, 2022; submitted, revised, March 16, 2023; accepted March 29, 2023.

Funding support: This work was funded by the American Medical Association Practice Transformation Initiative.

Disclaimer: The opinions expressed in this article are those of the author(s) and should not be interpreted as American Medical Association policy.

Previous presentations: Illinois Primary Health Care Association Annual Leadership Conference; Chicago, Illinois; October 6, 2022.

Acknowledgments: Elizabeth Adetoro assisted with study design, data collection, validation, and interpretation. Ryan Jaeger, AllianceChicago, created the data set for analysis.

REFERENCES

1. Sinsky CA, Panzer J. The solution shop and the production line — the case for a frameshift for physician practices. *N Engl J Med*. 2022;386(26):2452-2453. 10.1056/nejmp2202511
2. Simon J, Panzer J, Adetoro E, et al. Frequency of administration of standardized screening questions in Federally Qualified Health Centers. *JAMA Intern Med*. 2021;181(9):1253-1255. 10.1001/jamainternmed.2021.2503
3. National Committee for Quality Assurance. Patient-centered medical home: developing the business case from a practice perspective. Accessed Dec 20, 2020. https://www.ncqa.org/programs/health-care-providers-practices/patient-centered-medical-home-pcmh/
4. Health Resources and Services Administration. Uniform Data System: reporting instructions for calendar year 2020 health center data. Updated Aug 21, 2020. Accessed May 7, 2021. https://bphc.hrsa.gov/sites/default/files/bphc/datareporting/pdf/2020-uds-manual.pdf
5. US Census Bureau. Week 40 household pulse survey: December 1 – December 13. Census.gov. Published Jan 19, 2022. Accessed Oct 24, 2022. https://www.census.gov/data/tables/2021/demo/hhp/hhp40.html
6. Arroll B, Goodyear-Smith F, Crengle S, et al. Validation of PHQ-2 and PHQ-9 to screen for major depression in the primary care population. *Ann Fam Med*. 2010;8(4):348-353. 10.1370/afm.1139
7. Levis B, Sun Y, He C, et al; Depression Screening Data (DEPRESSD) PHQ Collaboration. Accuracy of the PHQ-2 alone and in combination with the PHQ-9 for screening to detect major depression: systematic review and meta-analysis. *JAMA*. 2020;323(22):2290-2300. 10.1001/jama.2020.6504
8. Kroenke K, Spitzer RL, Williams JB. The Patient Health Questionnaire-2: validity of a two-item depression screener. *Med Care*. 2003;41(11):1284-1292. 10.1097/01.mlr.0000093487.78664.3c

9. Jordan P, Shedden-Mora MC, Löwe B. Psychometric analysis of the Generalized Anxiety Disorder scale (GAD-7) in primary care using modern item response theory. *PLoS One*. 2017;12(8):e0182162. 10.1371/journal.pone.0182162
10. Plummer F, Manea L, Trepel D, McMillan D. Screening for anxiety disorders with the GAD-7 and GAD-2: a systematic review and diagnostic metaanalysis. *Gen Hosp Psychiatry*. 2016;39:24-31. 10.1016/j.genhosppsych.2015.11.005
11. Belle D. Poverty and women's mental health. *Am Psychol*. 1990;45(3):385-389. 10.1037/0003-066X.45.3.385
12. Santiago CD, Kaltman S, Miranda J. Poverty and mental health: how do low-income adults and children fare in psychotherapy? *J Clin Psychol*. 2013;69(2):115-126. 10.1002/jclp.21951
13. Jia H, Guerin RJ, Barile JP, et al. National and state trends in anxiety and depression severity scores among adults during the COVID-19 pandemic — United States, 2020–2021. *MMWR Morb Mortal Wkly Rep*. 2021;70:1427–1432. 10.15585/mmwr.mm7040e3external icon
14. Ettman CK, Cohen GH, Abdalla SM, et al. Persistent depressive symptoms during COVID-19: a national, population-representative, longitudinal study of U.S. adults. *Lancet Reg Health Am*. 2022;5:100091. 10.1016/j.lana.2021.100091
15. Draft recommendation statement: screening for anxiety in adults. United States Preventive Services Taskforce. Published Sep 20, 2022. Accessed Oct 24, 2022. https://www.uspreventiveservicestaskforce.org/uspstf/draft-recommendation/anxiety-adults-screening#bootstrap-panel–7
16. Carolan R, Marshall D, Kulkarni P, et al. Comparing the rate of positive PHQ-2 in self-administered paper versus provider-administered verbal screening tools. Poster presented Medical Education Research Forum 2019, Henry Ford Hospital. https://scholarlycommons.henryford.com/merf2019qi/4
17. Harris M, Tayler B. Don't let metrics undermine your business. Harvard Business Review. Published Aug 27, 2019. Accessed Oct 24, 2022. https://hbr.org/2019/09/dont-let-metrics-undermine-your-business