Addressing overestimation of the prevalence of depression based on self-report screening questionnaires

Brett D. Thombs PhD, Linda Kwakkenbos PhD, Alexander W. Levis BSc, Andrea Benedetti PhD

Cite as: *CMAJ* 2018 January 15;190:E44-9. doi: 10.1503/cmaj.170691

Mental health disorders, including major depressive disorder, are classified in research using validated diagnostic interviews.[1,2] However, administering diagnostic interviews to large population samples to estimate prevalence is expensive because of the time and trained personnel that are required. This is likely why researchers increasingly use self-report screening questionnaires, which require fewer resources, to estimate prevalence. We searched PubMed from Jan. 1, 2017, to Mar. 14, 2017, for primary studies with titles that indicated that prevalence of depression or depressive disorders had been assessed. Prevalence was based on screening questionnaires in 17 of 19 studies (89%; Appendix 1, available at www.cmaj.ca/lookup/suppl/doi:10.1503/cmaj.170691/-/DC1). Many recent meta-analyses have also based estimates of prevalence of depression on screening questionnaires.[3–7] However, using screening questionnaires to estimate prevalence can overestimate prevalence and blur distinctions between low- and high-prevalence populations. We describe the problem and possible strategies for estimation of prevalence that are less resource intensive than conducting diagnostic interviews with all patients.

KEY POINTS

- The common practice of reporting the percentage of patients with scores above cut-off thresholds in screening questionnaires for depression as disorder prevalence substantially overestimates prevalence and misinforms users of epidemiological evidence.
- Exaggeration of the prevalence of depression is disproportionately high in low-prevalence populations and blurs distinctions between high- and low-prevalence populations.
- Researchers should use diagnostic interview methods that have been validated for estimating prevalence.
- A two-stage estimation method that combines screening questionnaires and diagnostic interviews can reduce resource requirements and generate valid prevalence estimates for depression.

How are patients classified with screening questionnaires for depression?

Typically, screening questionnaires for depression are completed independently by respondents. The questionnaires assess symptoms similar to those evaluated in diagnostic interviews, but they do not assess functional impairment or investigate non-psychiatric conditions that can produce similar symptoms. Patients are classified as likely or unlikely to have depression based on scores above or below a cut-off threshold. Researchers set cut-offs by comparing scores on a screening questionnaire to classifications based on validated diagnostic interviews and attempting to maximize correct classifications. Different approaches may be used,[8,9] but many researchers simply maximize combined sensitivity (probability that a person with depression is classified correctly) and specificity (probability that a person without depression is classified correctly).[10] Because screening is intended to identify previously unrecognized cases, cut-off thresholds for screening are set to cast a wide net and identify substantially more patients who may have depression than those who will meet diagnostic criteria based on a diagnostic interview.

How should percentage above cut-offs on screening questionnaires be interpreted?

Positive predictive value (PPV) is the percentage of patients with scores above a test cut-off who have the target condition. For screening questionnaires for depression, PPV is the percentage of patients with a positive screen who meet diagnostic criteria. Positive predictive value depends on test sensitivity, specificity and true prevalence, but because screening tests are designed to cast a wide net, PPV is often very low. In many medical settings, fewer than 3 of 10 patients with a positive screen have major depression.[11]

The percentage of patients above a cut-off threshold typically exceeds true prevalence substantially. This has been shown by several recent highly cited meta-analyses that combined results from primary studies that used validated diagnostic interviews and primary studies that reported percentages of patients above cut-off thresholds on screening questionnaires for depression. In

a meta-analysis involving patients who underwent bariatric surgery, 19% had depression in 34 studies based on evaluation by screening questionnaires, but the rate was 7% to 8% in six studies that used a validated diagnostic interview.[3] Another meta-analysis of 43 studies involving new fathers during the prenatal and postpartum periods reported an overall prevalence of depression of 10%; however, three included studies that used validated diagnostic interviews reported a prevalence less than 5%.[5,12] Yet another meta-analysis, on depression among medical students,[7] reported that 27% of participants from 183 studies had depression. However, the only included study that used a validated diagnostic interview reported 9% prevalence, which is comparable to the 9% prevalence among 18- to 25-year-olds and the 7% prevalence among 26- to 49-year-olds in the general population of the United States.[13]

Some researchers have attempted to address this problem by labelling the percentage of patients above cut-offs for screening questionnaires as the prevalence of "clinically significant" symptoms or "symptoms" of depression rather than depression.[14,15] However, these designations are not based on evidence that these cut-offs reflect a meaningful divide between impairment and nonimpairment. Furthermore, the percentage of patients above cut-off thresholds varies depending on the particular screening questionnaire and cut-off threshold used. For example, a systematic review of depression after myocardial infarction found that 31% of patients had a score at or above the standard cut-off of 10 on the Beck Depression Inventory, whereas only 16% had a score at or above the standard cut-off of 8 on the Hospital Anxiety and Depression Scale.[16]

Another concern is that screening tools for depression overestimate prevalence more in low true-prevalence populations than in high true-prevalence populations. Based on assumed values of sensitivity and specificity, the percentage of patients who would score above a cut-off threshold for a screening questionnaire can be calculated for different values of true prevalence. In Table 1, we used estimates of sensitivity and specificity for the standard cut-off of 10 or greater on the Patient Health Questionnaire-9 (PHQ-9) from a recent meta-analysis involving about 20 000 patients (12% had depression).[17] Sensitivity and specificity may vary by patient population symptom severity and, thus, by prevalence.[18,19] Therefore, Table 1 shows a basic scenario and scenarios where sensitivity and specificity are adjusted in calculations across prevalence. Estimated prevalence is substantially exaggerated when true prevalence is lowest. In all scenarios, the percentage of patients above the cut-off threshold is at least twice the true prevalence when true prevalence is 10% or less, but this ratio decreases as true prevalence increases. This is because the misclassification of noncases as cases of depression (false positives) is disproportionately high in low-prevalence populations and only minimally offset by false-negative screens, which occur

Table 1: Comparison of true prevalence and percentage of patients above a cut-off threshold for screening tests

True prevalence, %	Sensitivity	Specificity	Percentage of patients above cut-off	Percentage of patients with false-positive screens among those above cut-off	Percentage of patients with false-negative screens among those below cut-off	Percentage of patients above cut-off – true prevalence	Percentage of patients above cut-off/true prevalence
Basic scenario:* sensitivity and specificity were constant across levels of true prevalence							
0.0	78.0	87.0	13.0	100.0	0.0	13.0	–
5.0	78.0	87.0	16.3	76.0	1.3	11.3	3.3
10.0	78.0	87.0	19.5	60.0	2.7	9.5	2.0
15.0	78.0	87.0	22.8	48.6	4.3	7.8	1.5
20.0	78.0	87.0	26.0	40.0	5.9	6.0	1.3
25.0	78.0	87.0	29.3	33.3	7.8	4.3	1.2
30.0	78.0	87.0	32.5	28.0	9.8	2.5	1.1
Adjusted for varying sensitivity and specificity:† sensitivity changed by 1% and specificity by 2% per 10% change in prevalence							
0.0	77.0	89.0	11.0	100.0	0.0	11.0	–
5.0	77.5	88.0	15.3	74.6	1.3	10.3	3.1
10.0	78.0	87.0	19.5	60.0	2.7	9.5	2.0
15.0	78.5	86.0	23.7	50.3	4.2	8.7	1.6
20.0	79.0	85.0	27.8	43.2	5.8	7.8	1.4
25.0	79.5	84.0	31.9	37.6	7.5	6.9	1.3
30.0	80.0	83.0	35.9	33.1	9.4	5.9	1.2

*Based on sensitivity = 78% and specificity = 87%, which are estimates for the standard cut-off threshold of 10 or greater for the Patient Health Questionnaire-9 from a meta-analysis of published results from 21 292 patients (2573 cases, 12%).[17]
†Sensitivity and specificity may vary with disease prevalence.[18,19] Thus, estimates of sensitivity and specificity were adjusted upward or downward from a prevalence of 10% based on a meta-analysis[19] that found that sensitivity may decrease 1% and specificity may increase 2% per 10% reduction in prevalence.

when true cases are missed by the screening test. Consequently, even populations with very low prevalence appear to have high prevalence based on the percentage above a screening test cut-off; this is the case even if terms such as "clinically significant symptoms" are used to describe patients above the cut-off threshold. Calculations in Table 1 do not account for precision of sensitivity and specificity estimates or potential heterogeneity across samples, but these factors could potentially exacerbate this problem.

What are the alternatives for estimating prevalence of depression?

Three methods for generating prevalence estimates from screening questionnaires or from a combination of screening questionnaires and diagnostic interviews have been proposed, including back calculation based on sensitivity and specificity,[20] prevalence matching[8] and two-stage estimation.[21]

Back calculation

Back calculation involves adjusting the percentage above a cut-off threshold by existing estimates of sensitivity and specificity.[20] The percentage of patients above a cut-off is equal to the percentage with true positive results for screening plus the percentage with false-positive screens. Based on this, a simple formula can be derived to estimate disorder prevalence (derivation of the formula is presented in Appendix 2, available at www.cmaj.ca/lookup/suppl/doi:10.1503/cmaj.170691/-/DC1):

$$\text{Prevalence} = (\%\ \text{above cut-off} + \text{specificity} - 1)/(\text{sensitivity} + \text{specificity} - 1)$$

However, estimation based on this method assumes that the exact sensitivity and specificity are known for the population being studied, which rarely occurs in practice. A meta-analysis of screening and case finding for major depressive order using the PHQ-9, which included about 20 000 patients, had 95% confidence intervals (CIs) for the standard cut-off threshold of 10 or greater that ranged from 70% to 84% for sensitivity and 84% to 90% for specificity.[17] This lack of certainty about true sensitivity and specificity can lead to substantial swings in back-calculated prevalence.

Figure 1 shows the estimated prevalence generated across a range of percentages among patients with a score above the cut-off for a screening questionnaire. The red line shows the estimated prevalence based on point estimates of PHQ-9 sensitivity (78%) and specificity (87%).[17] The green line shows estimated prevalence based on the lower bound of the 95% CI for sensitivity and the upper bound for specificity, and the blue line for the opposite. As shown by the black lines, if 20% of patients have a score above the cut-off threshold, plausible estimates of true disorder prevalence

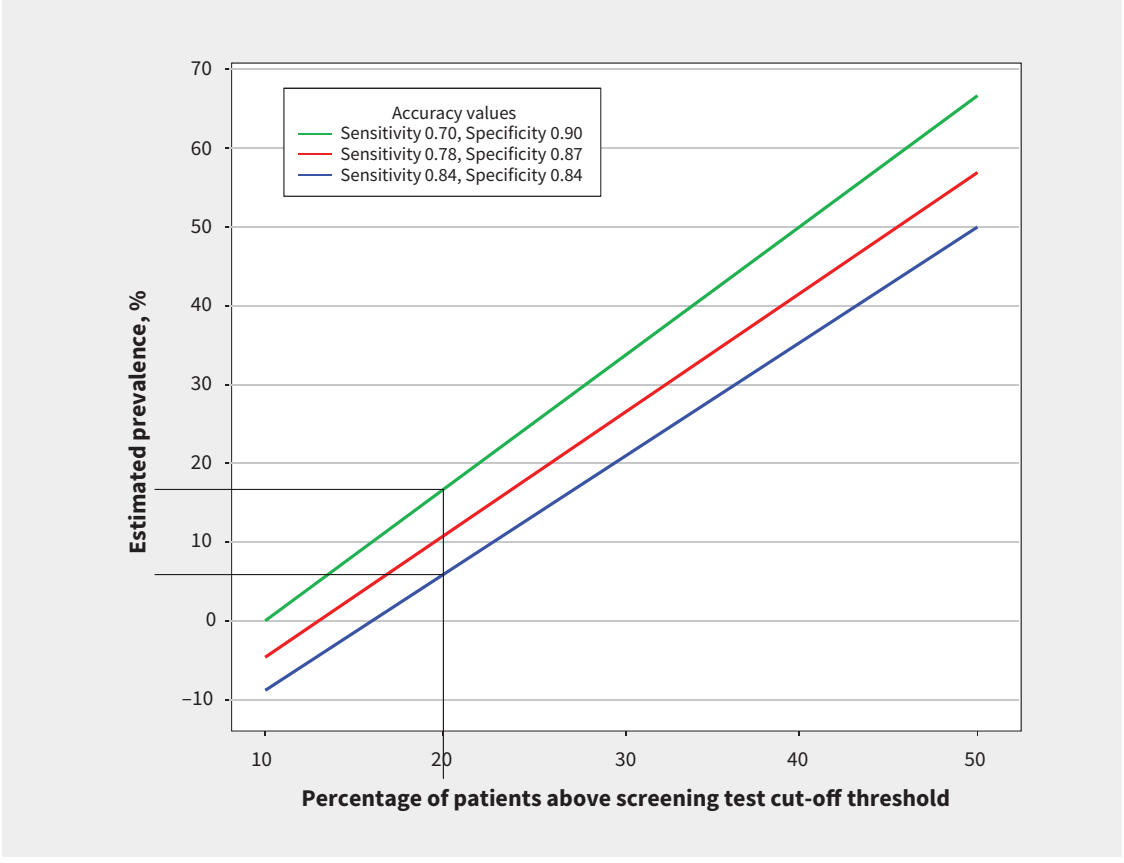

Figure 1: Estimated disorder prevalence based on the percentage of patients with scores above a cut-off threshold for a screening test, using estimates of sensitivity and specificity from a meta-analysis of the Patient Health Questionnaire-9 for detecting major depressive disorder.[17] Black lines highlight estimated prevalence for situations where 20% of patients have a score above the test cut-off threshold.

would range from 6% (blue line) to 17% (green line). However, this example likely underestimates the actual degree of imprecision that would be encountered in practice: we incorporated, for simplicity, ranges of estimates for sensitivity and specificity but ignored imprecision in the estimated percentage of patients with scores above the cut-off threshold for the screening questionnaire. We used CI estimates of sensitivity and specificity from a very large meta-analysis of the PHQ-9, but intervals for other screening questionnaires with less data would be even wider. Furthermore, we did not consider heterogeneity of estimates from different settings and the ramifications of this for implementation. An additional consideration is that estimated prevalence may actually be negative in

Table 2: Precision of two-stage prevalence estimation for true prevalence, sample size and percentage of patients with negative results for a screening test who were administered diagnostic interviews*

No. of patients	No. of patients with a positive screening result	No. of patients with a negative screening result	Percentage of patients with a negative screening result who were assessed with a diagnostic interview	No. of diagnostic interviews needed	95% CI	95% CI width	Raw difference between CI width and CI width when 100% of patients with negative screens were interviewed	Ratio of CI width to CI width when 100% of patients with negative screens were interviewed
True prevalence = 5%								
500	81	419	10.0	123	2.8–12.1	9.3	5.3	2.3
500	81	419	50.0	291	3.4–7.9	4.5	0.5	1.1
500	81	419	100.0	500	3.5–7.5	4.0	–	–
1000	163	838	10.0	246	3.0–7.9	4.9	2.2	1.8
1000	163	838	50.0	582	3.8–6.9	3.1	0.4	1.1
1000	163	838	100.0	1001	3.8–6.5	2.7	–	–
2000	325	1675	10.0	493	3.5–6.9	3.4	1.5	1.8
2000	325	1675	50.0	1163	4.0–6.2	2.2	0.2	1.1
2000	325	1675	100.0	2000	4.1–6.0	1.9	–	–
True prevalence = 10%								
500	98	403	10.0	138	5.9–15.9	10.0	4.7	1.9
500	98	403	50.0	299	7.5–13.6	6.1	0.8	1.2
500	98	403	100.0	501	7.6–12.9	5.3	–	–
1000	195	805	10.0	276	6.9–13.9	7.0	3.2	1.9
1000	195	805	50.0	598	8.1–12.3	4.2	0.5	1.1
1000	195	805	100.0	1000	8.3–12.0	3.7	–	–
2000	390	1610	10.0	551	7.6–12.5	4.9	2.2	1.8
2000	390	1610	50.0	1195	8.6–11.6	3.0	0.4	1.1
2000	390	1610	100.0	2000	8.8–11.4	2.6	–	–
True prevalence = 20%								
500	130	370	10.0	167	13.5–27.6	14.1	7.1	2.0
500	130	370	50.0	315	16.2–24.4	8.2	1.2	1.2
500	130	370	100.0	500	16.7–23.8	7.0	–	–
1000	260	740	10.0	334	15.1–25.0	9.9	4.9	2.0
1000	260	740	50.0	630	17.3–23.0	5.8	0.8	1.2
1000	260	740	100.0	1000	17.6–22.6	5.0	–	–
2000	520	1480	10.0	668	16.7–23.9	7.2	3.7	2.1
2000	520	1480	50.0	1260	18.0–22.1	4.1	0.6	1.2
2000	520	1480	100.0	2000	18.3–21.8	3.5	–	–

Note: CI = confidence interval. Numbers in the table are rounded to the nearest integer.
*Appendix 3 shows the estimation methods.

some scenarios where assumptions about sensitivity and specificity are inaccurate.

Prevalence matching

Prevalence matching[8] involves conducting very large research studies to set a cut-off for estimation of the prevalence of depression rather than screening for previously unidentified cases. This could be done by administering a screening tool and a validated diagnostic interview to all patients included in a study and setting a cut-off score that results in the percentage above the cut-off matching as closely as possible the number of patients with depression based on a validated diagnostic interview rather than to balance sensitivity and specificity. However, barriers to using this approach and generating accurate estimates of prevalence include the large number of patients who would need to be administered a diagnostic interview in the calibrating study and the high likelihood that results would not generalize well to other samples, given the substantial heterogeneity of results in existing studies of screening questionnaires.[17] Thus, estimates based on a cut-off score established in one study may be inaccurate when the cut-off is applied in other settings.

Two-stage prevalence estimation

In the two-stage approach,[21,22] first, all patients are administered a screening questionnaire. Then, all patients with positive screens, but only a randomly selected portion of patients with negative screens, are evaluated with a validated diagnostic interview. Prevalence is estimated by adding the number of patients with positive screens who meet diagnostic criteria and the number of patients with negative screens who also meet diagnostic criteria, weighting the latter to reflect their actual proportion of the total sample. This still requires diagnostic interviews but can reduce the number of interviews that need to be conducted substantially. Methods for implementing a two-stage approach have been described previously.[22]

Table 2 shows the precision of estimates that would likely be obtained using a two-stage approach. Precision, based on the width of estimated 95% CIs, is higher when true prevalence is lower, when the total number of patients is higher, and when a greater percentage of patients with negative results for screening are interviewed. In many scenarios, differences in precision are minimal, and this shows that investigators may be able to achieve sufficient precision to meet their needs by interviewing only a small proportion of patients with negative results for screening, which would have positive resource implications. The methods used to generate Table 2 can be found in Appendix 3, available at www.cmaj.ca/lookup/suppl/doi:10.1503/cmaj.170691/-/DC1.

What are the implications of these observations in depression research?

Screening tests for mental health and other types of screening questionnaires are not designed to make diagnostic classifications, and they are not calibrated to estimate prevalence. Using them in this way distorts prevalence estimates, often substantially, and does so disproportionately in low-prevalence populations. Estimating disorder prevalence with screening questionnaires misinforms evidence users, including health care decision-makers. It may also contribute to overdiagnosis, because practitioners may use the same methods to diagnose cases in clinical practice, and they may assume that they should be finding similar rates of disorders. Overdiagnosis can lead to inappropriate labelling and nocebo effects, as well as the unnecessary consumption of health care resources and potentially harmful treatment for patients who will not benefit.[23,24]

There are important implications for how research should be conducted and reported. First, prevalence estimates should be based on appropriate methods. Researchers should not report rates above cut-off thresholds in screening questionnaires as estimates of prevalence or clinical impairment. Second, systematic reviews and meta-analyses of the prevalence of depression should be based on results from validated diagnostic interviews. Third, comparisons between samples and descriptions of mental health symptoms based on depression screening tools should ideally use continuous scores rather than cut-off categories for screening questionnaires.[25] In some cases, categorical divisions may be helpful to illustrate data distributions and make comparisons, but there is no reason why the categories used should be dichotomous or based on cut-off thresholds of screening questionnaires. If categories are used, a clear rationale should be provided, including a justification for the category thresholds chosen. Finally, the knowledge needed to accurately implement back calculation and prevalence matching is not yet available. When efficient methods for estimating the prevalence of depression are needed, two-stage estimation of prevalence presents a viable option that can reduce resource use substantially and generate unbiased, reasonably precise prevalence estimates.

References

1. Wittchen H-U. Reliability and validity studies of the WHO-Composite International Diagnostic Interview (CIDI): a critical review. *J Psychiatr Res* 1994;28:57-84.
2. Spitzer RL, Williams JBW, Gibbon M, et al. The Structured Clinical Interview for DSM-III-R (SCID) – I: history, rationale, and description. *Arch Gen Psychiatry* 1992;49:624-9.
3. Dawes AJ, Maggard-Gibbons M, Maher AR, et al. Mental health conditions among patients seeking and undergoing bariatric surgery: a meta-analysis. *JAMA* 2016;315:150-63.
4. Edmondson D, Richardson S, Fausett JK, et al. Prevalence of PTSD in survivors of stroke and transient ischemic attack: a meta-analytic review. *PLoS One* 2013;8:e66435.
5. Paulson JF, Bazemore SD. Prenatal and postpartum depression in fathers and its association with maternal depression: a meta-analysis. *JAMA* 2010;303:1961-9.
6. Mata DA, Ramos MA, Bansal N, et al. Prevalence of depression and depressive symptoms among resident physicians: a systematic review and meta-analysis. *JAMA* 2015;314:2373-83.
7. Rotenstein LS, Ramos MA, Torre M, et al. Prevalence of depression, depressive symptoms, and suicidal ideation among medical students: a systematic review and meta-analysis. *JAMA* 2016;316:2214-36.
8. Kelly MJ, Dunstan FD, Lloyd K, et al. Evaluating cutpoints for the MHI-5 and MCS using the GHQ-12: a comparison of five different methods. *BMC Psychiatry* 2008;8:10.
9. Smits N, Smit F, Cuijpers P, et al. Using decision theory to derive optimal cut-off scores of screening instruments: an illustration explicating costs and benefits of mental health screening. *Int J Methods Psychiatr Res* 2007;16:219-29.
10. Youden WJ. Index for rating diagnostic tests. *Cancer* 1950;3:32-5.
11. Thombs BD, Arthurs A, El-Baalbaki G, et al. Risk of bias from inclusion of patients who already have diagnosis of or are undergoing treatment for depression in diagnostic accuracy studies of screening tools for depression: systematic review. *BMJ* 2011;343:d4825.

12. Thombs BD, Roseman M, Arthurs E. Prenatal and postpartum depression in fathers and mothers. *JAMA* 2010;304:961.

13. Behavioral health trends in the United States: results from the 2014 National Survey on Drug Use and Health. Rockville (MD): Center for Behavioral Health Statistics and Quality; 2015. Available: www.samhsa.gov/data/sites/default/files/NSDUH-FRR1-2014/NSDUH-FRR1-2014.htm (accessed 2017 Mar. 16).

14. Scott JE, Mathias JL, Kneebone AC. Depression and anxiety after total hip replacement among older adults; a meta-analysis. *Aging Ment Health* 2016;20:1243-54.

15. Buchberger B, Huppertz H, Krabbe L, et al. Symptoms of depression and anxiety in youth with type 1 diabetes: a systematic review and meta-analysis. *Psychoneuroendocrinology* 2016;70:70-84.

16. Thombs BD, Bass EB, Ford DE, et al. Prevalence of depression in survivors of acute myocardial infarction: review of the evidence. *J Gen Intern Med* 2006;21:30-8.

17. Moriarty AS, Gilbody S, McMillan D, et al. Screening and case finding for major depressive disorder using the Patient Health Questionnaire (PHQ-9): a meta-analysis. *Gen Hosp Psychiatry* 2015;37:567-76.

18. Usher-Smith JA, Sharp SJ, Griffin SJ. The spectrum effect in tests for risk prediction and diagnosis. *BMJ* 2016;353:i3139.

19. Leeflang MMG, Rutjes AWS, Reitsma JB, et al. Variation of a test's sensitivity and specificity with disease prevalence. *CMAJ* 2013;185:E537-44.

20. Rogan WJ, Gladen B. Estimating prevalence from the results of a screening test. *Am J Epidemiol* 1978;107:71-6.

21. Lehtinen V, Sohlman B, Nummelin T, et al. The estimated incidence of depressive disorder and determinants in the Finnish ODIN sample. *Soc Psychiatry Psychiatr Epidemiol* 2005;40:778-84.

22. Dunn G, Pickles A, Tansella M, et al. Two-phase epidemiological surveys in psychiatric research. *Br J Psychiatry* 1999;174:95-100.

23. Thombs BD, Coyne JC, Cuijpers P, et al. Rethinking recommendations for depression screening in primary care. *CMAJ* 2012;184:413-8.

24. Thombs BD, Ziegelstein RCRC. Does depression screening improve depression outcomes in primary care settings? *BMJ* 2014;348:g1253.

25. Altman DG, Royston P. The cost of dichotomising continuous variables. *BMJ* 2006;332:1080.

Competing Interests: None declared.

This article has been peer reviewed.

Affiliations: Lady Davis Institute of the Jewish General Hospital (Thombs, Kwakkenbos, Levis); Department of Psychiatry (Thombs, Kwakkenbos); Department of Epidemiology, Biostatistics, and Occupational Health (Thombs, Levis, Benedetti); Departments of Medicine, Psychology, and Educational and Counselling Psychology (Thombs), McGill University, Montréal, Que.; Behavioural Science Institute, Clinical Psychology (Kwakkenbos), Radboud University, Nijmegen, the Netherlands; Respiratory Epidemiology and Clinical Research Unit (Benedetti), McGill University Health Centre, Montréal, Que.

Contributors: Brett Thombs was responsible for the study concept and design. All of the authors participated in the conduct of analyses, contributed to interpretation of data, drafted sections of the manuscript, reviewed the manuscript critically for intellectual content, gave final approval of the version to be published and agreed to be accountable for all aspects of the work.

Acknowledgements: The authors thank Scott Patten, Kira Riehm, Ian Shrier and Roy Ziegelstein for their helpful feedback on earlier versions of this manuscript.

Funding: Brett Thombs and Andrea Benedetti were supported by researcher salary awards from the Fonds de recherche du Québec – Santé. Linda Kwakkenbos was supported by a Banting Postdoctoral Fellowship from the Canadian Institutes of Health Research. Alexander Levis was supported by a Masters Award from the Canadian Institutes of Health Research. There was no specific funding for this study, and no funding body had any input into any aspect of the study.

Correspondence to: Brett Thombs, brett.thombs@mcgill.ca