BMJ Mental Health

Prevalence estimates of major depressive disorder in 27 European countries from the European Health Interview Survey: accounting for imperfect diagnostic accuracy of the PHQ-8

Felix Fischer ,[1] Dario Zocholl,[2] Geraldine Rauch,[2] Brooke Levis,[3,4] Andrea Benedetti,[4,5,6] Brett Thombs ,[3,4,6,7,8,9] Matthias Rose,[1,10] Polychronis Kostoulas[11]

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/bmjment-2023-300675).

For numbered affiliations see end of article.

Correspondence to
Dr Felix Fischer, Department of Psychosomatic Medicine, Charité Universitätsmedizin Berlin, Berlin, Germany; felix.fischer@charite.de

Received 10 February 2023
Accepted 6 March 2023
Published Online First 5 April 2023

ABSTRACT

Background Cut-offs on self-report depression screening tools are designed to identify many more people than those who meet criteria for major depressive disorder. In a recent analysis of the European Health Interview Survey (EHIS), the percentage of participants with Patient Health Questionnaire-8 (PHQ-8) scores ≥10 was reported as major depression prevalence.

Objective We used a Bayesian framework to re-analyse EHIS PHQ-8 data, accounting for the imperfect diagnostic accuracy of the PHQ-8.

Methods The EHIS is a cross-sectional, population-based survey in 27 countries across Europe with 258 888 participants from the general population. We incorporated evidence from a comprehensive individual participant data meta-analysis on the accuracy of the PHQ-8 cut-off of ≥10. We evaluated the joint posterior distribution to estimate the major depression prevalence, prevalence differences between countries and compared with previous EHIS results.

Findings Overall, major depression prevalence was 2.1% (95% credible interval (CrI) 1.0% to 3.8%). Mean posterior prevalence estimates ranged from 0.6% (0.0% to 1.9%) in the Czech Republic to 4.2% (0.2% to 11.3%) in Iceland. Accounting for the imperfect diagnostic accuracy resulted in insufficient power to establish prevalence differences. 76.4% (38.0% to 96.0%) of observed positive tests were estimated to be false positives. Prevalence was lower than the 6.4% (95% CI 6.2% to 6.5%) estimated previously.

Conclusions Prevalence estimation needs to account for imperfect diagnostic accuracy.

Clinical implications Major depression prevalence in European countries is likely lower than previously reported on the basis of the EHIS survey.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Based on the Patient Health Questionnaire-8 (PHQ-8), the prevalence of current depressive disorder among participants in the European Health Interview Survey (EHIS; 27 countries, n=258 888) was recently reported to be 6.4%.

⇒ A 2021 individual participant data meta-analysis (44 studies, 9242 participants) found that the PHQ-9, which performs equivalently to the PHQ-8, identifies 2.5 times as many major depression cases as a validated semi-structured diagnostic interview.

WHAT THIS STUDY ADDS

⇒ In this study, we accounted for the imperfect diagnostic accuracy of the PHQ-8 using a Bayesian framework and estimated a considerably smaller overall prevalence of 2.1% across Europe.

⇒ Despite large differences in the proportion of observed positive tests between countries, accounting for diagnostic accuracy of the PHQ-8 resulted in insufficient power to establish prevalence differences.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ This study highlights a method to account for imperfect diagnostic accuracy in prevalence estimation and suggests that prevalence estimates from population-based surveys need to be interpreted carefully.

BACKGROUND

Depression is a leading cause of burden from disease worldwide,[1] responsible for approximately 2.2 million excess deaths in 2010[2] with numbers rising over recent decades.[3] Effective public health interventions could reduce the burden of depression.[4] Such strategies, however, must be informed by accurate data collected in large-scale population-based studies.

The largest study of depression prevalence across European countries in recent years was based on data from the European Health Interview Survey (EHIS), a large-scale population survey intended to inform health policy in Europe.[5] Participants with scores ≥10 on the Patient Health Questionnaire-8 (PHQ-8) screening tool were classified as having major depression. Authors reported an overall prevalence of 6.4% (95% CI 6.2% to 6.5%) and estimates for 27 European countries that ranged from 2.6% (2.1% to 3.0%) in the Czech Republic to 10.3% (9.3% to 11.3%) in Iceland. There were

To cite: Fischer F, Zocholl D, Rauch G, *et al*. *BMJ Ment Health* 2023;**26**:1–7.

large and significant differences between European countries.[5] This paper has been broadly cited.

Depression screening tools, including the PHQ-8 used in the EHIS, are designed to identify many more people than those who will be eventually determined to have a disorder after more comprehensive clinical assessment. These questionnaires are not designed to make definitive clinical diagnoses,[6 7] as positive screening results can be explained, for example, by accompanying depressive symptoms of other mental disorders. Also, persons who actually have a major depression, but are responding to treatment, might screen negative.

Ideally, major depression prevalence would hence be estimated on the basis of validated diagnostic interviews which are designed to replicate the diagnostic process, including assessment of symptom severity and impairment and ruling out alternative origins. These methods have been used in large population-based studies,[8] and the resources required can be reduced using strategies such as two-step implementation in conjunction with self-report screening tools.[9]

Nonetheless, brief self-report assessments are attractive and commonly used in studies on the prevalence of mental health conditions, as they are easy to administer to a large number of participants. Ignoring their imperfect diagnostic accuracy, however, leads to exaggerated prevalence estimates. An individual participant data meta-analysis of 44 primary studies (9242 participants) found that scores of ≥10 on the PHQ-9, which performs equivalently to the PHQ-8,[6] overestimated major depressive prevalence by 2.5 times compared with a validated semi-structured diagnostic interview.[10] This finding is consistent with earlier research[7 11] and research on other depression screening tools.[12 13]

Reporting of the positive test rate of the PHQ-8 as an index of population prevalence leads to an overestimation of true prevalence and can misinform public health policy making. It is therefore essential to account for imperfect diagnostic accuracy properly to estimate prevalence. One strategy is to incorporate prior information about sensitivity and specificity[9] in a Bayesian framework[14–16] in order to estimate depression prevalence.

OBJECTIVE

The objectives of this study were to (1) estimate prevalence of major depression in Europe, taking into account the imperfect diagnostic accuracy of the PHQ-8 screening tool, (2) assess differences in prevalence between countries and (3) compare results with those from a previous EHIS study, which assumed perfect PHQ-8 diagnostic accuracy for identifying major depression.

METHODS

This was a cross-sectional study that analysed data from the second wave of the EHIS. We included data from 27 countries where the PHQ-8 was administered. It is reported in accordance with the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) guideline.[17]

Data

The EHIS is a population-based study, which provides harmonised data on health status, healthcare use and health determinants as well as socioeconomic background variables across European countries. The EHIS is nationally organised and conducted every 5 years. The second wave took place between 2013 and 2015 in all European Union member states, Iceland and Norway.[18] It included 30 countries, of which 27 administered the PHQ-8 and were included in the present study. Study planning, sampling, data collection, procession and submission were conducted following guidelines from Eurostat.[19]

All EHIS study material was translated following a standardised translation protocol, including at least two translators with the target language as their mother tongue. In 21 countries, the study material was pretested. Data were collected in different countries by face-to-face interviews, telephone interviews, postal or online questionnaires or a combination of these methods. Data were mainly collected in 2014 and took 8 months on average for each country.[18]

The EHIS targets the European population aged 15 years and older living in private households. Depending on the country, one of three types of sampling frame was used: population registers, dwelling registers and population censes. The most common sampling design was a two-stage or three-stage stratified or systematic (cluster) sampling design with the individual being the ultimate sampling unit. The overall sample size in each country was determined to achieve a precision requirement of <1 percentage point error for the most critical variable in the survey. Out of 27 countries included in this study, for 9 the minimum effective sample size was not available, for 5 the target was not achieved and for 13 the minimum effective sample size was obtained. The unit non-response rate by country ranged from 16% to 70% with the highest rate of non-response when only self-administered questionnaires were collected.

Weighting factors for each individual were calculated to model the eligible population of each country, to account for design features of the survey and to reduce bias caused by non-response. Details can be found in the EHIS Methodological manual[19] (pp 156–160). Eurostat reported that study guidelines and implementation regulations have been closely followed, resulting in an overall sufficient or even good comparability across countries of the data.[18]

Primary measure

The PHQ-9 consists of nine items representing the nine symptoms included in diagnostic criteria for major depression according to the Diagnostic and Statistical Manual of Mental Disorders, fourth edition.[20] Each item is scored using a 4-category Likert scale that reflects frequency during the past 2 weeks (0–3 points; *not at all* to *nearly every day*). The PHQ-9 has been widely adopted in clinical research and practice.[21] The PHQ-8, which is commonly used in research and was used in the EHIS, omits an item of the PHQ-9 regarding suicidal ideation and self-harm. Total scores for the PHQ-8 can range from 0 to 24. Based on 27 studies included in an individual participant data meta-analysis of the PHQ-8 (6362 participants, 790 major depression cases), the cut-off of ≥10 maximised combined sensitivity and specificity compared with major depression classification based on a semi-structured interview. Sensitivity and specificity were estimated to be 0.86 (95% CI 0.80 to 0.90) and 0.86 (95% CI 0.83 to 0.89), although with considerable heterogeneity across studies.[6]

Statistical analysis

We used a Bayesian Latent Class Model to estimate major depression prevalence based on the PHQ-8.[15 22] The two latent (unobserved) classes represent disease status (major depressive disorder (MDD), no MDD). Based on observed PHQ-8 test status and prior information on PHQ-8 test characteristics, we estimate the probability of class membership, which is depression prevalence.

Replicating Arias-de la Torre *et al*,[5] PHQ-8 scores ≥10 were considered positive. Prior information on sensitivity and

specificity from an comprehensive meta-analysis[6] was employed probabilistically as prior distributions. The model was fitted using Markov Chain Monte Carlo methods.[23] Complete statistical analysis methods are described in online supplemental appendix 1.

Model

Given the differences in sampling and data collection, we considered the EHIS as 27 independent studies (one per country). In any country i, the observed number of the test positives y_i out of the n_i tested individuals was assumed to follow a binomial distribution with country-specific probability for a positive test p_i:

$$y_i \sim Binomial\left(n_i, p_i\right)$$

where p_i can be expressed as the sum of the probabilities for true positive (TP_i) and false positive tests (FP_i). These can be expressed in terms of prevalence ($Prev_i$), sensitivity (Se_i) and specificity (Sp_i):

$$p_i = TP_i + FP_i = Se_i \times Prev_i + \left(1 - Sp_i\right) \times \left(1 - Prev_i\right)$$

We used a joint multivariate normal prior on the logit of Se_i and Sp_i [24]:

$$\text{logit}\begin{pmatrix} Se_i \\ Sp_i \end{pmatrix} \sim N\left[\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \Sigma\right]$$

and a beta prior for $Prev_i$:

$$Prev_i \sim Beta\left(a, b\right)$$

Prior specification and model estimation

Priors represent existing knowledge about model parameters. Prior information about sensitivity (Se_i) and specificity (Sp_i) of the PHQ-8 was derived from a recent comprehensive individual participant data meta-analysis.[6] We modelled the prior distribution from estimates of mean logit-sensitivity β_0, mean logit-specificity β_1, between-study variances τ_0^2, τ_1^2 and between-study correlation ρ.[25] For the cut-off of ≥10, this yielded the following prior distribution:

$$\text{logit}\begin{pmatrix} Se_i \\ Sp_i \end{pmatrix} \sim N\left[\begin{pmatrix} 1.84 \\ 1.81 \end{pmatrix}, \begin{pmatrix} 0.37 & -0.07 \\ -0.07 & 0.77 \end{pmatrix}\right]$$

In our main analysis, we did not include informative prior information on depression prevalence to maintain comparability to the analysis reported by Arias-de la Torre *et al*[5] and used a uniform prior:

$$Prev_i \sim Beta\left(1, 1\right)$$

To investigate the appropriateness of the priors derived above and to inform prior sensitivity analysis, we performed prior predictive checks (see online supplemental appendix 2).

Model fitting

All models were fitted in Stan[26] using Markov Chain Monte Carlo sampling (4 chains, 5000 iterations, 2500 warm-up iterations). We examined trace plots, $\hat{R}$ values, effective sample size and autocorrelation plots to assess model convergence.[23] We performed posterior predictive checks to investigate whether the model was adequate to describe the observed data. Code to fit the model can be accessed at Open Science Framework (https://osf.io/w7fj2).

Interpretation

We assessed the joint posterior distribution of the model parameters $Prev_i, Se_i, Sp_i$. We reported posterior means and 95% CrIs of Se_i, Sp_i and $Prev_i$ for each country and compared the marginal and joint posterior distributions of Se_i and Sp_i to the respective prior distributions.

Based on the joint posterior distribution, we estimated the expected numbers of true positives ($TP_i = Prev_i \times Se_i$), false positives ($FP_i = \left(1 - Prev_i\right) \times \left(1 - Sp_i\right)$), true negatives ($TN_i = \left(1 - Prev_i\right) \times Sp_i$) and false negatives ($FN_i = Prev_i \times \left(1 - Se_i\right)$) for each country. We estimated the ratio between true prevalence and positive test frequency $\left(\frac{TP_i + FN_i}{TP_i + FP_i}\right)$ as well as the posterior probability $Pr\left(\left(TP_i + FP_i\right) > \left(TP_i + FN_i\right)\right)$ that the positive test frequency overestimates the true prevalence.

Finally, we assessed major depression prevalence differences between countries by inspecting the respective posterior distributions and reported the mean and 95% CrI of these differences for all pairwise comparisons. We also assessed, for each pair of countries, the posterior probability that the prevalence difference between both was greater/smaller than 0. A posterior probability >95% was considered as strong evidence for an actual prevalence difference between countries.

To investigate the robustness of our analysis against different priors, we investigated the impact of prior adjustments derived from prior predictive checks on the posterior distributions. See online supplemental appendix 1 for more details.

FINDINGS

The EHIS microdata contained 316 333 observations from 30 countries. We excluded 39 608 observations from Belgium, Spain and the Netherlands, where the PHQ-8 was not administered, 10 139 observations, where a proxy answered the survey instead of the selected person and 7694 observations, where the PHQ-8 sum score could not be calculated due to missing items. A participant flow chart and detailed information of PHQ-8 item non-response are provided in online supplemental appendices 3 and 4.

Overall, 258 888 participants were included in the study, of which 15 757 (6.1%) had a positive depression screening test with a PHQ-8 score of ≥10. Table 1 shows demographic characteristics of the sample (weighted, crude numbers are reported in online supplemental appendix 5). Table 2 shows the sample size as well as the absolute and relative number of positive tests (PHQ-8 score ≥10) for each country (weighted, crude numbers are reported in online supplemental appendix 6). The weighted mean PHQ-8 score was 2.8 (SD=3.8).

Model sampling performed well without any indication of problems. Empirical indicators such as trace plots, autocorrelation plots, $\hat{R}$ (all parameters <1.002) and effective sample size (2225–22 148) indicated appropriate exploration and convergence of the posterior distribution. Posterior predictive checks indicated that the fitted model could generate the observed data well (see online supplemental appendix 7).

Prevalence analysis

Figure 1 shows the posterior prevalence of major depression for each of the included countries. We estimated an overall prevalence of 2.1% (95% CrI 1.0% to 3.8%), which is considerably lower and less precise than the prevalence estimate of 6.4% (95% CI 6.2% to 6.5%) reported by Arias-de la Torre *et al*.[5]

We observed this pattern in every country. Although 10.3% of the participants in Iceland had a PHQ-8 score ≥10, the mean posterior prevalence estimate was only 4.2%. The 95% CrI

Table 1 Sociodemographic data of included participants (weighted)

	Overall (n=334 852 5250.1)
Age (years)	
15–29	66 601 524.7 (19.9)
30–44	83 083 006.0 (24.8)
45–59	86 348 395.6 (25.8)
60–74	65 622 957.7 (19.6)
75+	33 196 641.2 (9.9)
Sex	
Female	174 641 045.7 (52.2)
Male	160 211 479.4 (47.8)
Urbanisation	
Densely populated	120 420 878.1 (36.0)
Intermediate populated	110 379 374.2 (33.0)
Thinly populated	10 3064 481.0 (30.8)
Missing	987 791.8 (0.3)
Job status	
Carries out a job or profession, including unpaid work for a family business or holding, an apprenticeship or paid traineeship, etc	174 862 298.4 (52.2)
Unemployed	20 169 819.1 (6.0)
Pupil, student, further training, unpaid work experience	26 104 910.7 (7.8)
In retirement or early retirement or has given up business	79 511 329.6 (23.7)
Permanently disabled	6 656 137.5 (2.0)
In compulsory military or community service	192 853.4 (0.1)
Fulfilling domestic tasks	18 453 105.6 (5.5)
Other inactive person	7 104 640.4 (2.1)
Missing	1 797 430.5 (0.5)
Marital status	
Never married and never been in a registered partnership	101 791 196.0 (30.4)
Married or in a registered partnership	181 540 296.0 (54.2)
Widowed or in registered partnership that ended with death of partner	26 114 781.4 (7.8)
Divorced or in registered partnership that was legally dissolved	24 315 934.2 (7.3)
Missing	1 090 317.6 (0.3)

Crude numbers are reported in online supplemental appendix 3.

Table 2 Weighted sample size (n), absolute and relative number of positive tests (PHQ-8 ≥10) per country

Country	N	PHQ-8 ≥10	%
Austria	7 172 701.5	308 059.4	4.29
Bulgaria	5 134 235.7	335 436.5	6.53
Croatia	3 294 195.5	106 647.3	3.24
Cyprus	662 487.8	21 934.2	3.31
Czech Republic	8 724 200.5	224 999.2	2.58
Denmark	4 437 100.7	318 182.0	7.17
Estonia	1 093 388.3	72 635.7	6.64
Finland	3 752 576.1	196 393.2	5.23
France	46 273 994.9	3 251 141.1	7.03
Germany	68 578 374.3	6 337 385.6	9.24
Greece	8 526 355.1	288 774.9	3.39
Hungary	8 197 507.0	653 793.0	7.98
Iceland	238 579.1	24 636.4	10.33
Ireland	3 259 505.8	249 867.0	7.67
Italy	45 251 959.1	1 723 576.8	3.81
Latvia	1 563 899.3	72 034.3	4.61
Lithuania	2 406 381.6	72 479.6	3.01
Luxembourg	368 501.2	35 889.6	9.74
Malta	343 875.2	11 236.2	3.27
Norway	4 228 740.5	220 131.2	5.21
Poland	29 231 241.2	1 261 011.0	4.31
Portugal	8 765 340.3	802 390.8	9.15
Romania	16 633 668.4	728 839.9	4.38
Slovakia	4 594 060.7	117 703.3	2.56
Slovenia	1 660 406.5	91 404.0	5.50
Sweden	7 294 332.9	638 141.6	8.75
UK	43 164 916.0	3 193 101.0	7.40
Overall	334 852 525.2	21 357 825.1	6.38

Weights were calculated by the EHIS to account for design features of the survey and to reduce bias caused by non-response. Crude numbers are reported in online supplemental appendix 5.
EHIS, European Health Interview Survey; PHQ-8, Patient Health Questionnaire-8.

indicates that a wide range of prevalence values from 0.2% to 11.9% would be in line with prior information on PHQ-8 test characteristics and the observed data.

Table 3 reports the posterior means and 95% CrIs for the percentage of TP, FP, TN and FN PHQ-8 results. For example, we expect in Austria 95.3% of all tests to be TN, 0.8% TP, 0.4% FN and 3.5% FP. Hence, the ratio of prevalence (TP+FN) to positive tests (TP+FP) is 0.27 (05% CrI 0.01 to 0.90). The posterior probability that the true prevalence was smaller than the positive test frequency was 98.3% for Austria, which was similar across all countries.

We estimated the largest prevalence difference between Iceland and the Czech Republic with an estimated difference of 3.6% (95% CrI −0.6% to 11.3%, Pr=0.079). For no pairwise comparison, we can determine (Pr < 0.05) which country has the lower depression prevalence (see online supplemental appendix 8).

Analysis of sensitivity and specificity

Prior predictive checks indicated that the observed numbers of positive tests in the EHIS were unlikely given the prior information on specificity (see online supplemental appendix 2). The posterior distribution of Sp_i indicated this as well (see online supplemental appendix 9). Sp_i was estimated to be at least 0.90, whereas the prior distribution suggested a Sp_i between 0.70 and 0.90. The posterior Sp_i distribution had a greater mean and smaller variance compared with the prior distribution, suggesting that specificity in the EHIS was greater than in the available diagnostic studies that we used to establish our prior distribution. Online supplemental appendix 8 provides a more detailed explanation, how our model holds information about Sp_i despite that true depression status is not available, and shows the joint prior and the posterior distributions of Se_i and Sp_i, as well as their 95% CrIs.

Prior sensitivity analysis

We report results of a prior sensitivity analysis in online supplemental appendix 10. Compared with our main analysis, imposing a liberal assumption that depression prevalence is likely >0.5% and <25.8% resulted in minimally higher prevalence estimates with comparable uncertainty. We found a more pronounced effect when we additionally adjusted the prior on

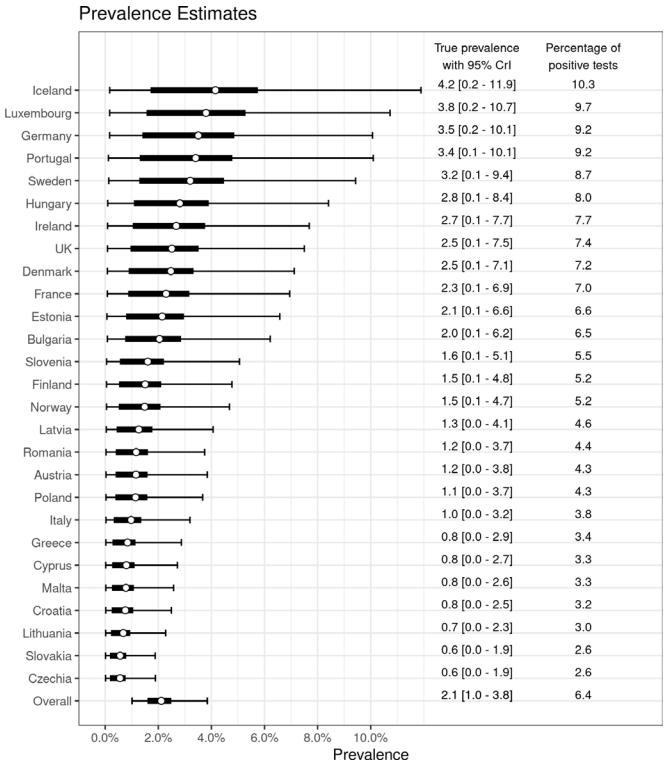

Figure 1 Mean posterior prevalence estimates and credible intervals (CrI), bold line indicates 50% CrI, thin line indicates 95% CrI. Note that width of credible intervals is reflecting sample size, and the functional relation between prevalence, sensitivity and specificity. For higher positive test rates, a wider range of combinations of these parameters are in line with the data. See online supplemental appendices 11 and 12 for a more detailed explanation.

specificity, leading to higher prevalence estimates compared with our initial analysis. Nonetheless, the posterior distribution suggested that actual prevalence was lower than the observed positive test frequency. Even in a purely hypothetical scenario in which we assumed unrealistically high a priori sensitivity and specificity of 95% and 97%, respectively, with high precision of the prior distributions, the uncertainty of the estimates remained rather large and observed positive test frequencies considerably overestimated the prevalence.

DISCUSSION

We incorporated the best available evidence on the diagnostic accuracy of the PHQ-8, using a cut-off of ≥10, to estimate major depression prevalence in Europe using data from the EHIS survey. Our main findings were that major depression prevalence across Europe was 2.1% (95% CrI 1.0% to 3.8%), that accounting for the imperfect diagnostic accuracy resulted in insufficient power to establish prevalence differences between countries and that previous prevalence estimates from the EHIS are likely overestimates.

Our depression prevalence estimates are in line with studies reporting depression prevalence on the basis of fully structured interviews, for example, 3.3% in the UK[27] and 2.0% in the Netherlands.[28] On the contrary, a survey in the Czech Republic using the Mini-International Neuropsychiatric Interview (MINI) reported depression prevalence of 4.0%, well above our estimate of 0.6%.[29] However, comparisons of these estimates need to consider that semi-structured and fully structured interviews are as well susceptible to measurement error.

The imprecision of the PHQ-8 as a diagnostic tool resulted in large uncertainty about depression prevalence. Hence, power to detect prevalence differences between countries was small. Despite the number of positive tests in Iceland (10.3%) being almost 4 times larger than in the Czech Republic (2.6%), there was still a 7.9% posterior probability that actual prevalence of major depression in Iceland is smaller than in the Czech Republic. This result is in contrast to the conclusion from Arias-de la Torre *et al*[5] that prevalence varies substantially and statistically significantly across European countries. Furthermore, we show that the relative frequency of positive tests as reported by Arias-de la Torre *et al*[5] almost certainly overestimates depression prevalence, even if we assume sensitivity and specificity to be substantially higher than previously reported.

Our analysis should not be mistaken for evidence that depression prevalence does not vary across Europe. For example, the posterior probability for depression prevalence being twice as high in Iceland compared with Czech Republic is 84.1%, indicating that the observed data are in line with a broad range of true prevalence differences. Therefore, relying on the PHQ-8 in a population-based survey alone seems insufficient to assess depression prevalence.

We incorporated sensitivity and specificity estimates reported from a large individual participant data meta-analysis[6] as the best currently available evidence. Nonetheless, we found a mismatch between prior and posterior information on specificity, suggesting that the PHQ-8 diagnostic accuracy in population-based study is different.[6] None of the primary studies using semi-structured diagnostic interviews were conducted in the general population; the mean PHQ-8 score of all included primary studies was 5.9 (SD=5.6),[6] whereas in the EHIS it was 2.8 (SD=3.8). This suggests that risk for depression in the negatively screened is lower than in diagnostic studies. Furthermore, diagnostic studies had on average only 236 participants and 29 major depression cases, resulting in imprecise and heterogenous estimates of sensitivity and specificity.[6] Direct sampling of sensitivity and specificity from the predictive posterior distribution of the individual participant data meta-analysis[6] would enable us to account for uncertainty on estimated heterogeneity, relax distributional assumptions or model associations between diagnostic accuracy and prevalence. Our approach of constructing priors from the published parameter estimates is less accurate, but can be replicated more easily in independent studies.

More precise information on the diagnostic accuracy of the PHQ-8 in the general population would be needed to obtain more meaningful prevalence estimates. This could be potentially achieved by tailoring diagnostic accuracy priors to the specific populations; however, existing evidence did not even allow assessment of country specific diagnostic accuracy. Further possibilities include using different cut-off values or the diagnostic algorithm of the PHQ-9. A promising approach is use a risk model based on the PHQ-8 sum score. To date, participants with a PHQ-8 score of 0 and 9 both constitute negative screens and are treated as equivalent using the standard cut-off of ≥10, although both apparently have a very different probability of major depression. Such a risk model is not available yet for the PHQ-8.

A limitation of the present analysis is that we used a uniform prior for prevalence in the main analysis to maintain comparability with the results of Arias-de la Torre *et al*.[5] Furthermore, one should be aware that our prevalence estimates did not reflect any previous knowledge on major depression prevalence across Europe.

Table 3 Estimated number of true positives (TP), false positives (FP), true negatives (TN), false negatives (FN) (in thousands), ratio of true prevalence and positive rate with the respective 95% CrIs and posterior probability (Pr) that positive rate overestimates true prevalence

	Percentage of				Ratio of true prevalence and positive test rate	Pr (true prevalence<positive test rate)
Country	TP	FP	TN	FN		
Austria	0.8 (0.0 to 2.3)	3.5 (2.0 to 4.3)	95.3 (93.7 to 95.7)	0.4 (0.0 to 2.0)	0.27 (0.01 to 0.90)	0.983
Bulgaria	1.5 (0.1 to 3.9)	5.1 (2.6 to 6.5)	92.9 (90.4 to 93.5)	0.6 (0.0 to 3.1)	0.31 (0.01 to 0.95)	0.979
Croatia	0.5 (0.0 to 1.6)	2.7 (1.7 to 3.2)	96.5 (95.5 to 96.8)	0.2 (0.0 to 1.2)	0.23 (0.01 to 0.77)	0.990
Cyprus	0.6 (0.0 to 1.7)	2.8 (1.7 to 3.3)	96.4 (95.4 to 96.7)	0.2 (0.0 to 1.3)	0.24 (0.01 to 0.82)	0.989
Czech Republic	0.4 (0.0 to 1.2)	2.2 (1.4 to 2.6)	97.2 (96.5 to 97.4)	0.2 (0.0 to 0.9)	0.22 (0.01 to 0.73)	0.992
Denmark	1.7 (0.1 to 4.5)	5.5 (2.7 to 7.1)	92.1 (89.2 to 92.8)	0.8 (0.0 to 3.6)	0.35 (0.01 to 0.99)	0.976
Estonia	1.5 (0.1 to 4.1)	5.1 (2.6 to 6.6)	92.7 (90.0 to 93.4)	0.6 (0.0 to 3.3)	0.32 (0.01 to 0.99)	0.976
Finland	1.1 (0.0 to 3.0)	4.2 (2.3 to 5.2)	94.3 (92.5 to 94.8)	0.4 (0.0 to 2.3)	0.29 (0.01 to 0.91)	0.984
France	1.6 (0.1 to 4.3)	5.4 (2.7 to 7.0)	92.3 (89.5 to 93.0)	0.7 (0.0 to 3.5)	0.33 (0.01 to 0.99)	0.976
Germany	2.5 (0.1 to 6.2)	6.7 (3.0 to 9.1)	89.8 (85.7 to 90.7)	1.0 (0.0 to 5.1)	0.38 (0.02 to 1.09)	0.966
Greece	0.6 (0.0 to 1.7)	2.8 (1.7 to 3.4)	96.3 (95.1 to 96.6)	0.3 (0.0 to 1.5)	0.25 (0.01 to 0.85)	0.985
Hungary	2.0 (0.1 to 5.1)	6.0 (2.9 to 7.9)	91.2 (87.8 to 92.0)	0.8 (0.0 to 4.2)	0.35 (0.01 to 1.06)	0.970
Iceland	3.0 (0.1 to 7.2)	7.3 (3.2 to 10.2)	88.5 (83.7 to 89.7)	1.1 (0.0 to 6.0)	0.40 (0.02 to 1.15)	0.958
Ireland	1.9 (0.1 to 4.9)	5.7 (2.8 to 7.6)	91.6 (88.6 to 92.3)	0.8 (0.0 to 3.7)	0.35 (0.01 to 1.00)	0.975
Italy	0.7 (0.0 to 2.0)	3.1 (1.8 to 3.8)	95.9 (94.6 to 96.2)	0.3 (0.0 to 1.6)	0.26 (0.01 to 0.84)	0.987
Latvia	0.9 (0.0 to 2.5)	3.7 (2.1 to 4.6)	95.0 (93.4 to 95.4)	0.4 (0.0 to 2.1)	0.28 (0.01 to 0.88)	0.983
Lithuania	0.5 (0.0 to 1.4)	2.5 (1.6 to 3.0)	96.8 (95.8 to 97.0)	0.2 (0.0 to 1.2)	0.23 (0.01 to 0.76)	0.990
Luxembourg	2.7 (0.1 to 6.6)	7.0 (3.1 to 9.6)	89.2 (84.9 to 90.3)	1.1 (0.0 to 5.3)	0.39 (0.02 to 1.10)	0.963
Malta	0.5 (0.0 to 1.6)	2.7 (1.7 to 3.3)	96.5 (95.5 to 96.8)	0.2 (0.0 to 1.3)	0.24 (0.01 to 0.79)	0.990
Norway	1.0 (0.0 to 2.9)	4.2 (2.3 to 5.2)	94.4 (92.5 to 94.8)	0.4 (0.0 to 2.3)	0.29 (0.01 to 0.90)	0.982
Poland	0.8 (0.0 to 2.3)	3.5 (2.0 to 4.3)	95.3 (93.9 to 95.7)	0.3 (0.0 to 1.8)	0.27 (0.01 to 0.85)	0.986
Portugal	2.4 (0.1 to 6.1)	6.7 (3.0 to 9.1)	89.9 (85.9 to 90.8)	1.0 (0.0 to 5.0)	0.37 (0.01 to 1.10)	0.966
Romania	0.8 (0.0 to 2.3)	3.6 (2.0 to 4.4)	95.3 (93.8 to 95.6)	0.4 (0.0 to 1.9)	0.27 (0.01 to 0.85)	0.985
Slovakia	0.4 (0.0 to 1.1)	2.2 (1.4 to 2.6)	97.3 (96.5 to 97.4)	0.2 (0.0 to 1.0)	0.22 (0.01 to 0.74)	0.992
Slovenia	1.1 (0.0 to 3.1)	4.4 (2.4 to 5.5)	94.0 (91.9 to 94.5)	0.5 (0.0 to 2.6)	0.29 (0.01 to 0.92)	0.980
Sweden	2.3 (0.1 to 5.7)	6.4 (3.1 to 8.6)	90.4 (86.6 to 91.2)	0.9 (0.0 to 4.7)	0.37 (0.02 to 1.08)	0.966
UK	1.8 (0.1 to 4.7)	5.6 (2.7 to 7.3)	91.9 (88.9 to 92.6)	0.7 (0.0 to 3.7)	0.34 (0.01 to 1.01)	0.974

More precise prevalence estimates could be achieved with different strategies. One would be to update prior information on diagnostic accuracy with study specific information on diagnostic accuracy. Such could be obtained using a two stage approach, where a semi-structured clinical interview is conducted in a subsample of the survey participants.[7 9] However, an appropriate sampling strategy and a sufficient number of interviews is vital to obtain precise and valid diagnostic accuracy estimates. Further strategies include incorporating auxiliary data, for example, from health insurance records,[30] or systematic evidence synthesis of all prevalence studies using different assessment tools like depression screeners, fully and semi-structured interviews.

CLINICAL IMPLICATIONS

Our analysis indicates that depression screening tools like the PHQ-8 should not be analysed or interpreted as if they were equivalent to diagnostic interviews in population-based studies. One should therefore not mistake the positive test frequency as a measure of prevalence. Rather, if a screening tool is to be used to attempt to estimate prevalence, appropriate statistical methods must be used to account for the less-than-perfect diagnostic accuracy of depression screening tools, even when sensitivity and specificity appear to be high. Our results call for the development of methods to estimate the probability of major depression more precisely from the PHQ-8 and other depression screening tools, for example, by risk models that are based on the sum score.

Policy makers should be aware that our analysis suggests that less people in Europe suffer from major depressive disorder and that differences between countries are likely smaller than previously reported.[5] Public health policy making must not necessarily rely on depression prevalence, but previous prevalence estimates are misleading. Evidence from different sources must be weighted carefully.

Author affiliations
[1]Department of Psychosomatic Medicine, Center for Internal Medicine and Dermatology, Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health, Berlin, Germany
[2]Institute of Biometry and Clinical Epidemiology, Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health, Berlin, Germany
[3]Lady Davis Institute for Medical Research, Jewish General Hospital, Montréal, Québec, Canada
[4]Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montréal, Québec, Canada
[5]Respiratory Epidemiology and Clinical Research Unit, McGill University Health Centre, Montréal, Québec, Canada
[6]Department of Medicine, McGill University, Montréal, Québec, Canada
[7]Department of Psychiatry, McGill University, Montréal, Québec, Canada
[8]Department of Psychology, McGill University, Montréal, Québec, Canada
[9]Biomedical Ethics Unit, McGill University, Montréal, Québec, Canada
[10]Department of Quantitative Health Sciences, University of Massachusetts Medical School, Worcester, Massachusetts, USA
[11]Faculty of Public Health, University of Thessaly, Karditsa, Greece

Contributors FF and PK conceived the study. FF and DZ performed data analysis. FF, DZ, GR, BL, AB, BT, MR and PK contributed to study design and interpretation.

FF drafted the manuscript. DZ, GR, BL, AB, BT, MR and PK provided critical reviews and approved the final manuscript. FF as the guarantor of the study accepts full responsibility for the work and the conduct of the study, had access to the data, and controlled the decision to publish.

Funding BL was supported by a Fonds de recherche du Québec—Santé (FRQS) Postdoctoral Training Fellowship, AB by a FRQS researcher salary award and BT by a Tier 1 Canada Research Chair, all outside of the present work.

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval This specific study is a secondary analysis of a public and anonymised dataset, which had obtained ethics approval and therefore required no additional ethics approval. The EHIS microdata is available at institutional level from Eurostat. All protocols for conducting the survey for data collection are available on the official Eurostat website at: EUR-Lex-02008R1338–20210101-EN-EUR-Lex. Participants gave informed consent to participate in the study before taking part.

Provenance and peer review Not commissioned; externally peer reviewed.

EHIS microdata are available from Eurostat (https://ec.europa.eu/eurostat). Our statistical analysis can be replicated using the aggregated data presented in table 2, using code published on the Open Science Framework (https://osf.io/w7fj2).

ORCID iDs
Felix Fischer http://orcid.org/0000-0002-9693-6676
Brett Thombs http://orcid.org/0000-0002-5644-8432

REFERENCES

1 Ferrari AJ, Charlson FJ, Norman RE, *et al*. Burden of depressive disorders by country, sex, age, and year: findings from the global burden of disease study 2010. *PLoS Med* 2013;10.
2 Patel V, Chisholm D, Parikh R, *et al*. Addressing the burden of mental, neurological, and substance use disorders: key messages from disease control priorities, 3rd edition. *Lancet* 2016;387:1672–85.
3 Liu Q, He H, Yang J, *et al*. Changes in the global burden of depression from 1990 to 2017: findings from the global burden of disease study. *J Psychiatr Res* 2020;126:134–40.
4 Herrman H, Kieling C, McGorry P, *et al*. Reducing the global burden of depression: a lancet-world psychiatric association commission. *Lancet* 2019;393:e42–3.
5 Arias-de la Torre J, Vilagut G, Ronaldson A, *et al*. Prevalence and variability of current depressive disorder in 27 european countries: a population-based study. *Lancet Public Health* 2021;6:e729–38.
6 Wu Y, Levis B, Riehm KE, *et al*. Equivalency of the diagnostic accuracy of the PHQ-8 and PHQ-9: a systematic review and individual participant data meta-analysis. *Psychol Med* 2020;50:1368–80.
7 Thombs BD, Kwakkenbos L, Levis AW, *et al*. Addressing overestimation of the prevalence of depression based on self-report screening questionnaires. *CMAJ* 2018;190:E44–9.
8 Alonso J, Angermeyer MC, Bernert S, *et al*. Prevalence of mental disorders in Europe: results from the European study of the epidemiology of mental disorders (esemed) project. *Acta Psychiatr Scand Suppl* 2004;109:21–7.
9 Taub NA, Morgan Z, Brugha TS, *et al*. Recalibration methods to enhance information on prevalence rates from large mental health surveys. *Int J Methods Psychiatr Res* 2005;14:3–13.
10 Levis B, Benedetti A, Ioannidis JPA, *et al*. Patient health questionnaire-9 scores do not accurately estimate depression prevalence: individual participant data meta-analysis. *J Clin Epidemiol* 2020;122:S0895-4356(19)30735-8:115–128..
11 Levis B, Yan XW, He C, *et al*. Comparison of depression prevalence estimates in meta-analyses based on screening tools and rating scales versus diagnostic interviews: a meta-research review. *BMC Med* 2019;17:65:65.:.
12 Lyubenova A, Neupane D, Levis B, *et al*. Depression prevalence based on the Edinburgh postnatal depression scale compared to structured clinical interview for DSM disorders classification: systematic review and individual participant data meta-analysis. *Int J Methods Psychiatr Res* 2021;30:e1860.
13 Brehaut E, Neupane D, Levis B, *et al*. Depression prevalence using the HADS-D compared to scid major depression classification: an individual participant data meta-analysis. *J Psychosom Res* 2020;139:110256.
14 Lewis FI, Torgerson PR. A tutorial in estimating the prevalence of disease in humans and animals in the absence of a gold standard diagnostic. *Emerg Themes Epidemiol* 2012;9:9:1–8.:.
15 Joseph L, Gyorkos TW, Coupal L. Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. *Am J Epidemiol* 1995;141:263–72.
16 McInturff P, Johnson WO, Cowling D, *et al*. Modelling risk when binary outcomes are subject to error. *Stat Med* 2004;23:1095–109.
17 von Elm E, Altman DG, Egger M, *et al*. The strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. *Ann Intern Med* 2007;147:573–7.
18 Eurostat. *Quality report of the second wave of the european health interview survey: 2018 edition*. Publications Office of the European Union, 2018.
19 Eurostat. *European health interview survey (EHIS wave 2) - methodological manual*. Luxembourg: Publications Office of the European Union, 2013. Available: https://ec.europa.eu/eurostat/product?code=KS-RA-13-018
20 Spitzer RL, Kroenke K, Williams JBW. Validation and utility of a self-report version of PRIME-MD: the PHQ primary care study. *JAMA* 1999;282:1737–44.
21 Kroenke K. PHQ-9: global uptake of a depression scale. *World Psychiatry* 2021;20:135–6.
22 Speybroeck N, Devleesschauwer B, Joseph L, *et al*. Misclassification errors in prevalence estimation: Bayesian handling with care. *Int J Public Health* 2013;58:791–5.
23 Gelman A, Carlin JB, Stern HS, *et al*. *Bayesian data analysis*. Boca Raton: CRC Press, 2013.
24 Riley RD, Ahmed I, Debray TPA, *et al*. Summarising and validating test accuracy results across multiple studies for use in clinical practice. *Stat Med* 2015;34:2081–103.
25 Riley RD, Dodd SR, Craig JV, *et al*. Meta-Analysis of diagnostic test studies using individual patient data and aggregate data. *Stat Med* 2008;27:6111–36.
26 Carpenter B, Gelman A, Hoffman MD, *et al*. Stan: a probabilistic programming language. *J Stat Softw* 2017;76:1.
27 McManus S, Bebbington P, Jenkins R, *et al*. Mental health and wellbeing in england: adult psychiatric morbidity survey 2014. leeds: NHS digital. n.d. Available: http://discover.ukdataservice.ac.uk/catalogue?sn=6379
28 van Loo HM, Beijers L, Wieling M, *et al*. Prevalence of internalizing disorders, symptoms, and traits across age using advanced nonlinear models. *Psychol Med* 2021;53:1–10.
29 Formánek T, Kagström A, Cermakova P, *et al*. Prevalence of mental disorders and associated disability: results from the cross-sectional czech mental health study (CZEMS). *Eur Psychiatry* 2019;60:1–6.
30 Edwards J, Pananos AD, Thind A, *et al*. A bayesian approach to estimating the population prevalence of mood and anxiety disorders using multiple measures. *Epidemiol Psychiatr Sci* 2021;30:e4.