Appendix to Fischer et al. 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

Appendix 1: Detailed statistical analysis plan ................................................................. 2
Appendix 2: Prior predictive checks ................................................................................. 5
Appendix 3: Participant Flowchart .................................................................................. 7
Appendix 4: Missing PHQ-8 item-level data .................................................................. 8
Appendix 4: Crude sociodemographic descriptives ......................................................... 10
Appendix 5: Crude sample size and number of positive tests per country ...................... 11
Appendix 6: Posterior predictive check .......................................................................... 12
Appendix 7: Estimated prevalence differences between countries ............................... 14
Appendix 9: Posterior sensitivity and specificity distributions ......................................... 15
Appendix 10: Prior sensitivity analysis .......................................................................... 17
Appendix 11: A note on the width of the credible intervals ............................................. 21
Appendix 12: A note on the rank order of prevalence estimates and prevalence differences between countries ................................................................. 22
References ..................................................................................................................... 23
Appendix 1: Detailed statistical analysis plan

We used a Bayesian Latent Class Model (BLCM) to estimate major depression prevalence based on the PHQ-8.[1] Major depression status was modelled as two latent (unobserved) classes, taking into account PHQ-8 test characteristics and observed PHQ-8 test status. Replicating Arias-de la Torre et al.[2], PHQ-8 scores ≥ 10 were considered positive. Necessary parameter constraints on sensitivity and specificity are employed probabilistically as prior distributions. A general approach to fit such Bayesian models is Markov Chain Monte Carlo (MCMC) methods. These allow sampling from the joint posterior distribution of all model parameters, even when likelihood functions are analytically intractable.[3] Inference on model parameters such as prevalence can then be directly deducted from the posterior distribution.

Model

The EHIS has been conducted in each participating country by national statistical offices or similar entities using slightly different sampling schemes and data collection procedures.[4] To account for this heterogeneity, we considered the EHIS as multiple studies (one per country), in particular since sensitivity and specificity of the PHQ-8 are known to considerably vary across studies. In any country $i$, the observed number of the test positives $y_i$ out of the $n_i$ tested individuals is assumed to follow a binomial distribution:

$$y_i \sim \text{Binomial}(n_i, p_i)$$

where $p_i$ is the probability of observing a positive test (the positive test rate) in the country $i$. Since positive tests are either true or false positives (TP and FP, respectively), the positive test rate is modelled as a function of the country-specific true prevalence of major depressive disorder ($\text{Prev}_i$) and the sensitivity ($\text{Se}_i$) and specificity ($\text{Sp}_i$) of the diagnostic test:

$$p_i = TP_i + FP_i = \text{Se}_i \times \text{Prev}_i + (1 - \text{Sp}_i) \times (1 - \text{Prev}_i)$$

As it is impossible to estimate three unknown quantities ($\text{Prev}_i, \text{Se}_i, \text{Sp}_i$) from a single quantity ($y_i$), we need to include external information on the diagnostic test characteristics to be able to estimate the true prevalence.[5] Within a Bayesian framework, we express such a priori information about model parameters in terms of prior distribution.

In this case, we impose a joint multivariate normal prior on the logit of $\text{Se}_i$ ($\beta_0$) and $\text{Sp}_i$ ($\beta_1$). This prior allows to account for the typically negative correlation of sensitivity and specificity between studies:[6]

$$\text{logit} \left( \frac{\text{Se}_i}{\text{Sp}_i} \right) \sim N \left( \begin{pmatrix} \beta_1 \\ \beta_0 \end{pmatrix}, \Sigma \right)$$
Since prevalence is a probability (i.e., values range from 0 to 1), the beta distribution is a natural choice for a prior distribution:

\[ \text{Prev}_i \sim \text{Beta}(a, b) \]

For each country, we use the same set of priors since we have no a priori expectations about specific differences.

**Prior specification and model estimation**

Prior knowledge about sensitivity \((Se_i)\) and specificity \((Sp_i)\) of the PHQ-8 can be drawn from diagnostic studies. In a recent comprehensive individual participant data meta-analysis on the diagnostic accuracy PHQ-8 [7] sensitivity \((Se)\) and specificity \((Sp)\) were modeled using a bivariate generalized logistic linear model.[8] A multivariate normal prior distribution of sensitivity and specificity can thus be specified based on the estimated mean logit-sensitivity \(\beta_0\), mean logit-specificity \(\beta_1\), between-study variances \(\tau_0^2\), \(\tau_1^2\), \(\tau_0^2\) and between-study correlation \(\rho\) from this bivariate generalized logistic linear model.[8]

\[
\begin{pmatrix}
\text{logit } (Se_i) \\
\text{logit } (Sp_i)
\end{pmatrix}
\sim N \left( \begin{pmatrix}
\beta_1 \\
\beta_0
\end{pmatrix}, \begin{pmatrix}
\tau_1^2 & \tau_1 \tau_0 \rho \\
\tau_1 \tau_0 \rho & \tau_0^2
\end{pmatrix} \right)
\]

For the cutoff of 10 (included), the respective model parameters were \(\beta_0 = 1.84\), \(\beta_1 = 1.81\), \(\tau_1 = 0.88\), \(\tau_0 = 0.61\), \(\rho = 0.13\). This yields the following multivariate distribution:

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

Further, we used a uniform prior for prevalence in each country.

\[ \text{Prev}_i \sim \text{Beta}(1,1) \]

We choose this prior on prevalence in our main analysis to maintain comparability to the analysis by Arias-de la Torre et al.[2] Although uniform, this prior might be informative given the functional relationships between prevalence, sensitivity and specificity in this case. This means that although we impose no information on the prevalence through the specific prior distribution, the information we provide about sensitivity and specificity might make some prevalence values less likely than others and the prior on prevalence is therefore not really uninformative.
Prior predictive checks
In order to investigate the appropriateness of the priors derived above, we performed prior predictive checks by randomly generating data given these priors under the model. Results are described in the Appendix 4.

Model fitting
All models were fitted in Stan [9] using Markov Chain Monte Carlo (MCMC) sampling (4 chains, 5,000 iterations, 2,500 warm-up iterations). MCMC sampling is a method to obtain samples from a distribution, which can be applied in situations where it is not mathematically feasible to determine the target distribution analytically. It is important to ensure that the Markov Chain has successfully explored the parameter space and has reached a stable state, i.e., has converged to an equilibrium, so that the samples are actually generated from the target distribution. We examined trace plots, values, effective sample size (ESS), autocorrelation plots as indicators of posterior exploration and model convergence.[3] We performed posterior predictive checks to investigate whether the model is adequate to describe the observed data.

Interpretation
In order to draw inferences from the model, we assessed the joint posterior distribution of the model parameters $\text{Prev}_i, \text{Se}_i, \text{Sp}_i$. We compared the marginal and joint posterior distributions of $\text{Se}_i$ and $\text{Sp}_i$ to the respective prior distributions. We report means and 95% credible intervals of $\text{Se}_i$, $\text{Sp}_i$ and $\text{Prev}_i$ for each country.

Based on the joint posterior distribution, we furthermore estimated the expected numbers of true positives ($\text{TP}_i = \text{Prev}_i \times \text{Se}_i$), false positives ($\text{FP}_i = (1 - \text{Prev}_i) \times (1 - \text{Sp}_i)$), true negatives ($\text{TN}_i = (1 - \text{Prev}_i) \times \text{Sp}_i$), and false negatives ($\text{FN}_i = \text{Prev}_i \times (1 - \text{Se}_i)$) for each country. In order to investigate whether the number of positive tests overestimates the prevalence of major depressive disorder we estimated the ratio of positive tests and the prevalence ($\frac{\text{TP}}{\text{TP} + \text{FP}}$) as well as the posterior probability $\text{Pr}((\text{TP}_i + \text{FN}_i) < (\text{TP}_i + \text{FP}_i))$ that the positive test rate 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% credible interval 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 different from 0. A posterior probability greater 95% was considered as strong evidence for an actual prevalence difference between countries.

In order to investigate the robustness of our analysis against different priors, we also fitted the models with prior adjustments deducted from prior predictive checks and compared the posterior prevalence distributions to the main analysis.
Appendix 2: Prior predictive checks

Prior predictive checking helps to assess whether the prior distributions are actually in line with previous knowledge. The idea is to generate data only based on the prior distributions. This generated data should be in a reasonable range to conclude that the prior distributions are appropriately defined. For example, if the prior predictive checks show that there is a large probability of depression prevalence values above 50%, the priors could be considered as inappropriate because they do not represent current knowledge well. Similarly, if the prior predictive distribution of the prevalence would be very narrow, say all resulting prevalence estimates between 5% and 6%, the priors could be discarded for being too informative.

For the initial priors we found that the prior predictive distribution of the positive test frequencies is close to uniform, i.e., the a priori probability that the majority of all tests are positive is approximately 50%. As this seems unrealistic to assume, we imposed a weakly informative prior distribution on prevalence which puts 90% of the prior weight on MDD prevalence between 0.51% and 25.89% (median: 6.70%):

\[ \text{Pre}v_1 \sim \text{Beta}(1,10) \]

Still, prior predictive checks showed that there was more than 99% prior probability to observe positive test rates above 5%. This can be explained by the specificity. Given a specificity of 86%, we would expect a positive test rate of at least 14% by false positive tests alone. Hence, a lower positive test rate can only be observed if specificity is higher than in the initial prior. We therefore additionally adjusted the multivariate prior on \( Se_1 \) and \( Sp_1 \):

\[
\begin{pmatrix}
\text{logit} (Se_1) \\
\text{logit} (Sp_1)
\end{pmatrix}
\sim N \left( \begin{pmatrix} 1.84 \\ 3.00 \end{pmatrix}, \begin{pmatrix} 0.37 & -0.07 \\ -0.07 & 0.77 \end{pmatrix} \right)
\]

This multivariate prior now puts 95% of the probability of \( Sp_1 \) between 0.78 and 0.99 with a median of 0.95. The prior probability to observe positive test rates below 5% was now approximately 5%, which seems to be a reasonable prior assumption.
Figure 1: Comparison of prior predictive distribution of the positive test rates under different prior specifications. A prior predictive distribution represents the data that is expected given the prior, i.e., without taking any data into account. The prior specification of the main analysis yielded prior predictive positive rates that are higher than typically expected. Under a more informative prevalence prior, a Beta(1,10) distribution, the predictive distribution came closer to a realistic expectation, however, small prevalence estimates were still too unlikely. When additionally adjusting the prior on specificity, so that a prior specificity of 95% was assumed, prior predicted positive test rates were more reasonably distributed.
Appendix 3: Participant Flowchart

Figure 2: Flowchart of study participants
Appendix 4: Missing PHQ-8 item-level data

<table>
<thead>
<tr>
<th>Country</th>
<th>PHQ Item 1</th>
<th>PHQ Item 2</th>
<th>PHQ Item 3</th>
<th>PHQ Item 4</th>
<th>PHQ Item 5</th>
<th>PHQ Item 6</th>
<th>PHQ Item 7</th>
<th>PHQ Item 8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Austria</td>
<td>&lt;0.1</td>
<td>&lt;0.1</td>
<td>&lt;0.1</td>
<td>&lt;0.1</td>
<td>&lt;0.1</td>
<td>&lt;0.1</td>
<td>&lt;0.1</td>
<td>&lt;0.1</td>
</tr>
<tr>
<td>Bulgaria</td>
<td>3.4</td>
<td>2.9</td>
<td>1.9</td>
<td>1.6</td>
<td>2.6</td>
<td>5.1</td>
<td>3.8</td>
<td>4.4</td>
</tr>
<tr>
<td>Croatia</td>
<td>0.5</td>
<td>0.4</td>
<td>0.4</td>
<td>0.4</td>
<td>0.4</td>
<td>0.7</td>
<td>0.5</td>
<td>0.8</td>
</tr>
<tr>
<td>Cyprus</td>
<td>&lt;0.1</td>
<td>&lt;0.1</td>
<td>&lt;0.1</td>
<td>&lt;0.1</td>
<td>&lt;0.1</td>
<td>&lt;0.1</td>
<td>&lt;0.1</td>
<td>&lt;0.1</td>
</tr>
<tr>
<td>Czechia</td>
<td>&lt;0.1</td>
<td>&lt;0.1</td>
<td>&lt;0.1</td>
<td>&lt;0.1</td>
<td>&lt;0.1</td>
<td>&lt;0.1</td>
<td>&lt;0.1</td>
<td>&lt;0.1</td>
</tr>
<tr>
<td>Denmark</td>
<td>2.7</td>
<td>2.8</td>
<td>2.0</td>
<td>2.3</td>
<td>3.5</td>
<td>2.7</td>
<td>2.2</td>
<td>2.5</td>
</tr>
<tr>
<td>Estonia</td>
<td>&lt;0.1</td>
<td>&lt;0.1</td>
<td>&lt;0.1</td>
<td>&lt;0.1</td>
<td>&lt;0.1</td>
<td>&lt;0.1</td>
<td>&lt;0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>Finland</td>
<td>3.9</td>
<td>3.6</td>
<td>2.9</td>
<td>3.1</td>
<td>2.9</td>
<td>5.6</td>
<td>4.4</td>
<td>3.0</td>
</tr>
<tr>
<td>France</td>
<td>3.4</td>
<td>3.4</td>
<td>2.6</td>
<td>2.7</td>
<td>4.0</td>
<td>3.1</td>
<td>2.6</td>
<td>2.7</td>
</tr>
<tr>
<td>Germany</td>
<td>0.6</td>
<td>0.7</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>0.4</td>
<td>0.4</td>
<td>0.4</td>
</tr>
<tr>
<td>Greece</td>
<td>0.4</td>
<td>0.4</td>
<td>0.3</td>
<td>0.3</td>
<td>0.2</td>
<td>0.6</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>Hungary</td>
<td>0.4</td>
<td>0.2</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>0.3</td>
<td>0.2</td>
<td>0.1</td>
</tr>
<tr>
<td>Iceland</td>
<td>0.9</td>
<td>0.4</td>
<td>0.6</td>
<td>0.8</td>
<td>0.8</td>
<td>1.1</td>
<td>0.5</td>
<td>2.1</td>
</tr>
<tr>
<td>Ireland</td>
<td>7.0</td>
<td>8.1</td>
<td>6.4</td>
<td>6.2</td>
<td>7.6</td>
<td>8.0</td>
<td>7.5</td>
<td>8.6</td>
</tr>
<tr>
<td>Italy</td>
<td>1.1</td>
<td>1.1</td>
<td>1.1</td>
<td>1.2</td>
<td>1.2</td>
<td>1.1</td>
<td>1.1</td>
<td>1.2</td>
</tr>
<tr>
<td>Latvia</td>
<td>0.3</td>
<td>0.3</td>
<td>0.3</td>
<td>0.2</td>
<td>0.2</td>
<td>0.8</td>
<td>0.3</td>
<td>2.1</td>
</tr>
<tr>
<td>Lithuania</td>
<td>1.1</td>
<td>0.9</td>
<td>0.7</td>
<td>0.7</td>
<td>0.9</td>
<td>1.8</td>
<td>1.5</td>
<td>2.1</td>
</tr>
<tr>
<td>Luxembourg</td>
<td>4.2</td>
<td>4.8</td>
<td>3.3</td>
<td>3.0</td>
<td>4.2</td>
<td>4.2</td>
<td>3.5</td>
<td>4.2</td>
</tr>
<tr>
<td>Malta</td>
<td>0.8</td>
<td>0.6</td>
<td>0.6</td>
<td>0.6</td>
<td>0.6</td>
<td>0.7</td>
<td>0.6</td>
<td>0.7</td>
</tr>
<tr>
<td>Norway</td>
<td>0.6</td>
<td>0.3</td>
<td>0.3</td>
<td>0.3</td>
<td>0.3</td>
<td>0.5</td>
<td>0.3</td>
<td>0.4</td>
</tr>
<tr>
<td>Poland</td>
<td>0.5</td>
<td>0.4</td>
<td>0.3</td>
<td>0.3</td>
<td>0.3</td>
<td>0.6</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>Portugal</td>
<td>0.4</td>
<td>0.3</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
<td>0.5</td>
<td>0.3</td>
<td>0.4</td>
</tr>
<tr>
<td>Romania</td>
<td>&lt;0.1</td>
<td>&lt;0.1</td>
<td>&lt;0.1</td>
<td>&lt;0.1</td>
<td>&lt;0.1</td>
<td>&lt;0.1</td>
<td>&lt;0.1</td>
<td>&lt;0.1</td>
</tr>
<tr>
<td>Slovakia</td>
<td>&lt;0.1</td>
<td>&lt;0.1</td>
<td>&lt;0.1</td>
<td>&lt;0.1</td>
<td>&lt;0.1</td>
<td>&lt;0.1</td>
<td>&lt;0.1</td>
<td>&lt;0.1</td>
</tr>
<tr>
<td>Slovenia</td>
<td>1.6</td>
<td>1.9</td>
<td>1.5</td>
<td>1.5</td>
<td>2.1</td>
<td>1.8</td>
<td>1.7</td>
<td>1.6</td>
</tr>
<tr>
<td>Sweden</td>
<td>2.8</td>
<td>2.7</td>
<td>2.2</td>
<td>3.3</td>
<td>2.7</td>
<td>2.4</td>
<td>1.8</td>
<td>2.0</td>
</tr>
<tr>
<td>UK</td>
<td>0.5</td>
<td>0.4</td>
<td>0.4</td>
<td>0.4</td>
<td>0.4</td>
<td>0.5</td>
<td>0.4</td>
<td>0.6</td>
</tr>
</tbody>
</table>

Frequencies of missing item responses by country (in percent)
<table>
<thead>
<tr>
<th>Country</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Austria</td>
<td>100.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Bulgaria</td>
<td>92.3</td>
<td>2.6</td>
<td>1.4</td>
<td>1.0</td>
<td>0.7</td>
<td>0.4</td>
<td>0.4</td>
<td>0.2</td>
<td>1.2</td>
</tr>
<tr>
<td>Croatia</td>
<td>98.5</td>
<td>0.8</td>
<td>0.2</td>
<td>0.1</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.3</td>
</tr>
<tr>
<td>Cyprus</td>
<td>100.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Czechia</td>
<td>99.9</td>
<td>0.1</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Denmark</td>
<td>93.8</td>
<td>3.2</td>
<td>0.6</td>
<td>0.3</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
<td>0.3</td>
<td>1.3</td>
</tr>
<tr>
<td>Estonia</td>
<td>99.8</td>
<td>0.2</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Finland</td>
<td>89.8</td>
<td>6.1</td>
<td>0.8</td>
<td>0.4</td>
<td>0.2</td>
<td>0.3</td>
<td>0.4</td>
<td>0.7</td>
<td>1.5</td>
</tr>
<tr>
<td>France</td>
<td>91.3</td>
<td>5.0</td>
<td>1.0</td>
<td>0.4</td>
<td>0.3</td>
<td>0.2</td>
<td>0.2</td>
<td>0.5</td>
<td>1.2</td>
</tr>
<tr>
<td>Germany</td>
<td>98.3</td>
<td>1.1</td>
<td>0.2</td>
<td>0.1</td>
<td>0.0</td>
<td>0.0</td>
<td>0.1</td>
<td>0.1</td>
<td>0.2</td>
</tr>
<tr>
<td>Greece</td>
<td>98.6</td>
<td>0.8</td>
<td>0.2</td>
<td>0.1</td>
<td>0.1</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.1</td>
</tr>
<tr>
<td>Hungary</td>
<td>99.2</td>
<td>0.7</td>
<td>0.1</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.1</td>
</tr>
<tr>
<td>Ireland</td>
<td>95.3</td>
<td>3.5</td>
<td>0.8</td>
<td>0.2</td>
<td>0.1</td>
<td>0.1</td>
<td>0.0</td>
<td>0.0</td>
<td>0.1</td>
</tr>
<tr>
<td>Italy</td>
<td>98.3</td>
<td>0.6</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
</tr>
<tr>
<td>Latvia</td>
<td>97.4</td>
<td>1.9</td>
<td>0.3</td>
<td>0.1</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.1</td>
</tr>
<tr>
<td>Lithuania</td>
<td>96.9</td>
<td>1.2</td>
<td>0.6</td>
<td>0.4</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>0.5</td>
</tr>
<tr>
<td>Luxembourg</td>
<td>90.6</td>
<td>4.4</td>
<td>1.0</td>
<td>0.6</td>
<td>0.3</td>
<td>0.4</td>
<td>0.3</td>
<td>0.7</td>
<td>1.7</td>
</tr>
<tr>
<td>Malta</td>
<td>98.6</td>
<td>0.6</td>
<td>0.2</td>
<td>0.1</td>
<td>0.0</td>
<td>0.1</td>
<td>0.1</td>
<td>0.0</td>
<td>0.4</td>
</tr>
<tr>
<td>Norway</td>
<td>98.8</td>
<td>0.7</td>
<td>0.2</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.2</td>
</tr>
<tr>
<td>Poland</td>
<td>98.7</td>
<td>0.7</td>
<td>0.2</td>
<td>0.1</td>
<td>0.1</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.2</td>
</tr>
<tr>
<td>Portugal</td>
<td>98.7</td>
<td>0.7</td>
<td>0.3</td>
<td>0.1</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Romania</td>
<td>100.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Slovakia</td>
<td>100.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Slovenia</td>
<td>95.7</td>
<td>2.5</td>
<td>0.2</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>0.2</td>
<td>0.2</td>
<td>0.9</td>
</tr>
<tr>
<td>Sweden</td>
<td>91.2</td>
<td>6.1</td>
<td>0.8</td>
<td>0.3</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>0.5</td>
<td>0.8</td>
</tr>
<tr>
<td>UK</td>
<td>98.8</td>
<td>0.8</td>
<td>0.1</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.3</td>
</tr>
</tbody>
</table>

Frequencies of missing item responses by questionnaire (in percent)
Appendix 5: Crude sociodemographic descriptives

<table>
<thead>
<tr>
<th>Age</th>
<th>Overall (N=258,888)</th>
</tr>
</thead>
<tbody>
<tr>
<td>15–29</td>
<td>42,311 (16.3%)</td>
</tr>
<tr>
<td>30–44</td>
<td>58,862 (22.7%)</td>
</tr>
<tr>
<td>45–59</td>
<td>68,919 (26.6%)</td>
</tr>
<tr>
<td>60–74</td>
<td>61,030 (23.6%)</td>
</tr>
<tr>
<td>75+</td>
<td>27,766 (10.7%)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sex</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>141,578 (54.7%)</td>
</tr>
<tr>
<td>Male</td>
<td>117,310 (45.3%)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Urbanization</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Densely populated</td>
<td>86,340 (33.4%)</td>
</tr>
<tr>
<td>Intermediate-populated</td>
<td>77,832 (30.1%)</td>
</tr>
<tr>
<td>Thinly populated</td>
<td>94,422 (36.5%)</td>
</tr>
<tr>
<td>Missing</td>
<td>294 (0.1%)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Job status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carries out a job or profession, including unpaid work for a family business or holding, an apprenticeship or paid traineeship, etc.</td>
</tr>
<tr>
<td>Unemployed</td>
</tr>
<tr>
<td>Pupil, student, further training, unpaid work experience</td>
</tr>
<tr>
<td>In retirement or early retirement or has given up business</td>
</tr>
<tr>
<td>Permanently disabled</td>
</tr>
<tr>
<td>In compulsory military or community service</td>
</tr>
<tr>
<td>Fulfilling domestic tasks</td>
</tr>
<tr>
<td>Other inactive person</td>
</tr>
<tr>
<td>Missing</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Marital Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Never married and never been in a registered partnership</td>
</tr>
<tr>
<td>Married or in a registered partnership</td>
</tr>
<tr>
<td>Widowed or in registered partnership that ended with death of partner</td>
</tr>
<tr>
<td>Divorced or in registered partnership that was legally dissolved</td>
</tr>
<tr>
<td>Missing</td>
</tr>
</tbody>
</table>
Appendix 6: Crude sample size and number of positive tests per country

<table>
<thead>
<tr>
<th>Country</th>
<th>n</th>
<th>PHQ_8 ≥ 10</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Austria</td>
<td>15,701</td>
<td>585</td>
<td>3.73</td>
</tr>
<tr>
<td>Bulgaria</td>
<td>5,258</td>
<td>390</td>
<td>7.42</td>
</tr>
<tr>
<td>Croatia</td>
<td>5,016</td>
<td>198</td>
<td>3.95</td>
</tr>
<tr>
<td>Cyprus</td>
<td>4,695</td>
<td>180</td>
<td>3.83</td>
</tr>
<tr>
<td>Czech Republic</td>
<td>6,607</td>
<td>209</td>
<td>3.16</td>
</tr>
<tr>
<td>Denmark</td>
<td>5,449</td>
<td>379</td>
<td>6.96</td>
</tr>
<tr>
<td>Estonia</td>
<td>5,439</td>
<td>357</td>
<td>6.56</td>
</tr>
<tr>
<td>Finland</td>
<td>5,146</td>
<td>248</td>
<td>4.82</td>
</tr>
<tr>
<td>France</td>
<td>14,191</td>
<td>949</td>
<td>6.69</td>
</tr>
<tr>
<td>Germany</td>
<td>24,404</td>
<td>2,269</td>
<td>9.30</td>
</tr>
<tr>
<td>Greece</td>
<td>7,834</td>
<td>310</td>
<td>3.96</td>
</tr>
<tr>
<td>Hungary</td>
<td>5,777</td>
<td>477</td>
<td>8.26</td>
</tr>
<tr>
<td>Iceland</td>
<td>3,812</td>
<td>386</td>
<td>10.13</td>
</tr>
<tr>
<td>Ireland</td>
<td>9,046</td>
<td>627</td>
<td>6.93</td>
</tr>
<tr>
<td>Italy</td>
<td>21,934</td>
<td>877</td>
<td>4.00</td>
</tr>
<tr>
<td>Latvia</td>
<td>6,607</td>
<td>310</td>
<td>4.69</td>
</tr>
<tr>
<td>Lithuania</td>
<td>4,982</td>
<td>176</td>
<td>3.53</td>
</tr>
<tr>
<td>Luxembourg</td>
<td>3,629</td>
<td>358</td>
<td>9.86</td>
</tr>
<tr>
<td>Malta</td>
<td>3,974</td>
<td>133</td>
<td>3.35</td>
</tr>
<tr>
<td>Norway</td>
<td>8,069</td>
<td>374</td>
<td>4.64</td>
</tr>
<tr>
<td>Poland</td>
<td>22,076</td>
<td>1,023</td>
<td>4.63</td>
</tr>
<tr>
<td>Portugal</td>
<td>17,974</td>
<td>1,774</td>
<td>9.87</td>
</tr>
<tr>
<td>Romania</td>
<td>16,422</td>
<td>850</td>
<td>5.18</td>
</tr>
<tr>
<td>Slovakia</td>
<td>5,489</td>
<td>160</td>
<td>2.91</td>
</tr>
<tr>
<td>Slovenia</td>
<td>5,914</td>
<td>326</td>
<td>5.51</td>
</tr>
<tr>
<td>Sweden</td>
<td>5,737</td>
<td>489</td>
<td>8.52</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>17,706</td>
<td>1,343</td>
<td>7.58</td>
</tr>
<tr>
<td>Overall</td>
<td>258,888</td>
<td>15,757</td>
<td>6.09</td>
</tr>
</tbody>
</table>
Appendix 7: Posterior predictive check

Posterior predictive checks can be used to investigate whether the final model generates data that is similar to the observed data. Doing so helps to assess the goodness-of-fit of the model to the data. In other words, it is a check whether the model is useful to explain the data. We found that our final model did very well reproduce the observed proportion of positive tests in each country, indicating an appropriate fit of the model to the observed data. We report the posterior predictive distributions of positive test rate of main analysis. The histograms shows the expected positive test rate on basis of the posterior distributions of prevalence, sensitivity and specificity, dashed line indicates observed positive test rate. In all countries, expected positive test rates align closely around the observed positive test rate, indicating appropriate model fit.
Figure 3: Expected frequencies of positive screening tests from posterior distributions
Appendix 8: Estimated prevalence differences between countries

<table>
<thead>
<tr>
<th>Country</th>
<th>Estimate</th>
<th>95% CI</th>
<th>p(Prevalence Difference &lt; 0)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Austria</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bulgaria</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Croatia</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cyprus</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Czechia</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Denmark</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Estonia</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Finland</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>France</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Germany</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Greece</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hungary</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Iceland</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ireland</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Italy</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Latvia</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lithuania</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Luxembour</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Malta</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Norway</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Poland</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Portugal</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Romania</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Slovakia</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Slovenia</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sweden</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Posterior mean with 95% credible interval of prevalence differences between countries and the respective posterior probability that prevalence difference < 0.
Appendix 9: Posterior sensitivity and specificity distributions

Apparently, there is no information on true depression status in the data. We still found considerable difference between prior and posterior distributions of specificity, indicating that there is some information about specificity in the data. This can be explained by the number of positive tests being an upper bound to the number of false positive tests.

Consider 10 out of 100 tests being positive. A hypothetical specificity of 80% implies that 20% of the healthy participants will have a false positive test. Our observation of 10 positive tests would then only be plausible, if prevalence is at least 50%, otherwise we would expect more than 10 false-positives. To observe not more than 10 positive tests, sensitivity would then be necessarily close to 0, otherwise we would observe more true positives. Hence, our observed test rate is only possible if either the sensitivity is very low under a high prevalence or if the specificity is very high under a low prevalence. The first case is very unlikely given our prior information on sensitivity. Based on our prior knowledge, it therefore is more plausible that specificity is higher than the 80% assumed at first than that sensitivity is close to 0%, which leads to an updated information about the specificity.
Figure 4: Joint prior 80%, 90%, 95% density regions (grey, dashed) and 1000 draws from the posterior distributions (solid) for sensitivity and specificity in each country. Error bars indicate marginal 95% credible interval for country specific posterior sensitivity and specificity.
Appendix 10: Prior sensitivity analysis

Priors are a central element of Bayesian data analysis as we combine the prior information with the likelihood of the observed data to generate posterior distributions of the model parameters. Given the same data, different priors can lead to different results and it is therefore mandatory to check how the results of the analysis change when different priors are used. We used the priors derived by our prior predictive checking in this sensitivity analysis.

Prior sensitivity analysis reveals that the posterior rate of the prevalence is affected by prior choices for both the prevalence and sensitivity/specificity, but the general conclusions of our analysis remain the same regardless of the prior assumptions. Figure 5 shows the different prior distributions used in the main analysis and the prior sensitivity analysis.

The uniform prior on the prevalence, which was used in the main analysis, resulted in larger prevalence estimates than the more informative Beta(1,10)-prior on the prevalence if the observed positive rates were rather large. For small observed positive rates, the prevalence
prior was less important, because the specificity was estimated to be close to 1 regardless of
the prevalence prior and therefore became the main driver of posterior estimation. In line
with this observation, adjusting the prior specificity to have a mean of 95% increased all
estimates considerably as well as their uncertainty. This can be accounted to a much higher
expected true positive rate due to the high expected specificity.

We additionally investigated a hypothetical scenario in which very large estimates with high
precision for sensitivity and specificity are available. Even in this scenario of a quite accurate
diagnostic tool, which currently seems very unrealistic, the uncertainty of the estimates
remained relatively large and observed positive rates were considerably overestimating the
prevalence. In general, the estimated prevalence will always be considerably below the
observed positive rate, except for situations with extreme priors that assume a high
prevalence or a close to perfect specificity.

Based on a reviewer’s suggestion we additionally estimated a model which takes into
account uncertainty of the mean logit sensitivity and specificity by imposing a normal
hyperprior on the mean logit sensitivity and specificity. This leads to slightly broader credible
intervals (see Figure 7).
Figure 6: Results of prior sensitivity analysis
Figure 7: Prevalence estimates of the main analysis compared to a model incorporating uncertainty of the mean logit sensitivity and specificity.
**Appendix 11: A note on the width of the credible intervals**

Unlike credible/confidence intervals of the positive test rates, where the width would mainly depend on sample size, the credible intervals of the prevalence reflects the functional association between prevalence, sensitivity and specificity. As these three interact, a wide range of combinations of parameters are in line with the observed positive test rates. Small observed test rates constrain this parameter space, mainly through the specificity, as we can’t observe more false positives than positives. Hence, the width of the credible intervals for prevalence here depends substantially on the positive test rate.

Consider the following example: Imagine a country with a positive test rate of 2%. Then, the specificity must actually be very high, because there are very few false-positives (assuming prevalence is not extremely high, e.g. >50%). Thus, there are two possible scenarios how this low positive test rate can be explained. Either (1) the prevalence is rather small or (2) the prevalence is substantially higher than the positive test rate, but sensitivity is extremely low, so we have very few true positives. Since we assume - based on prior studies - that sensitivity and specificity are both at least moderately high, the first scenario is much more plausible. When the positive test rate is 10% instead, this is compatible with a relatively large prevalence together with a moderately large sensitivity, while low prevalence rates are equally plausible since specificity is not perfect.

In other words: if we assume at least moderately high sensitivity and specificity, and the true positive test rate is 2%, then we can be quite sure, that many of them will be false-positives, and since we do not believe that we have an extraordinary additional number of false-negatives, the prevalence is bounded somewhere between 0% and approx. 2%. If the true positive rate is 10%, again, many of these could be false-positives, which suggests prevalence close to 0%. But it could also be explained by a mixture of false-positives and true-positives, resulting in more uncertainty regarding the prevalence estimation.
Appendix 12: A note on the rank order of prevalence estimates and prevalence differences between countries

It has been noted that the overall rank order of countries remains unchanged after correcting the positive test rate for the diagnostic imperfectness of the PHQ-8.

The credible intervals of prevalence shown in Figure 1 represent all possible rates that are consistent with all plausible combinations of sensitivity and specificity. For example, higher prevalence rates close to the positive test rate result from iterations with high sensitivity and few false positives, while lower prevalence rates near zero come from iterations with low specificity and many false positives. The width of the credible intervals primarily reflects our uncertainty about the diagnostic accuracy of the PHQ-8.

The impact on expected prevalence differences across countries is, however, substantial. Our model assumes that sensitivity and specificity are independent across countries. It is therefore possible that extreme sensitivity and specificity values are assumed for a given prevalence comparison and that the rank order of countries is not necessarily the same as the average rank order in each iteration. Hence, this leads to situations where the estimated prevalence for the Czech Republic is relatively high (at the upper bound of its credible interval, since sensitivity and specificity in the Czech Republic might be close to 1), whereas it is at its lower bound for Iceland (since sensitivity and specificity might be close to their lower bound). Hence, the observed data is in line with the scenario that prevalence in Iceland is lower than in the Czech Republic.
References


