RESEACH

大うつ病のスクリーニングにおける Patient Health Questionnaire-9(PHQ-9) の精度

最新のシステマティックレビュー及び個別参加者データのメタアナリシス (IPDMA : Individual Participant Data Meta-Analysis)

Zelalem F Negeri,[1,2] Brooke Levis,[3] Ying Sun,[1] Chen He,[1] Ankur Krishnan,[1] Yin Wu,[1,4] Parash Mani Bhandari,[1,2] Dipika Neupane,[1,2] Eliana Brehaut,[1] Andrea Benedetti,[2,5,6] Brett D Thombs,[1,2,4,5,7,8,9] DEPRESSD PHQ グループを代表して

抄録

目的

従来の個別参加者データによるメタアナリシスを更新することと、一般診療でうつ病のスクリーニング評価ツールとして最もよく使用されている Patient Health Questionnaire-9（PHQ-9）について、大うつ病の研究全体あるいは参加者のサブグループ別に検出し、その診断精度を確定させること。

デザイン

システマティックレビュー及び個別参加者データメタアナリシス。

データソース

2018 年 5 月 9 日までの文献を Ovid Medline（Medline In-Process 及び Other Non-Indexed Citations を含む）、PsycINFO、Web of Science を用いて検索した。

レビュー方法

PHQ-9 を実施し、半構造化面接（臨床医による実施用にデザインされた）、構造化面接（一般の者による実施用にデザインされた）、または精神疾患簡易構造化面接法（MINI）を用いて大うつ病患者の状態を分類している研究を適格とした。二変量ランダム効果モデルによるメタアナリシスを行い、半構造化面接（例えば、Structured Clinical Interview for Diagnostic and Statistical Manual（DSM 用構造化臨床面接））、構造化面接（例えば、WHO 統合国際診断面接（CIDI））、MINI をそれぞれ用いた研究において、カットオフ値 5 ～ 15 点でプールされた PHQ-9 の感度及び特異度の点推定と区間推定を得た。メタ回帰を利用して、PHQ-9 の精度が参照基準カテゴリーや参加者の特性と関連しているかどうかを検討した。

結果

適格基準を満たす研究 127 件（追加された研究は 42 件）のうち 100 件（79%、全参加者のうち 86%）から、合計 44,503 名（更新により 27,146 名が追加）の参加者データが得られた。半構造化面接の参照基準を用いた研究の中では、感度と特異度との組合せが最大となる標準的なカットオフ値は≧ 10 点で、プールされた PHQ-9 の感度と特異度（95% 信頼区間）はそれぞれ 0.85（0.79 ～ 0.89）、0.85（0.82 ～ 0.87）であった。特異度は参照基準間で同程度であったが、感度については、カットオフ値間で、半構造化面接を用いた研究の場合に構造化面接よりも 7 ～ 24%（中央値 21%）高く、

このトピックについて既に知られていること

Patient Health Questionnaire-9（PHQ-9）は、初期医療機関及び一般の医療機関で最も広く用いられているうつ病スクリーニングツールであり、カットオフ値≧ 10 点が大うつ病を同定する基準として使用されている。

2015 年 2 月までに実施された PHQ-9 の大うつ病検出精度に関する 58 件の研究（17,357 名の参加者）を対象とした個別参加者データのメタアナリシスでは、PHQ-9 は、半構造化面接を用いた場合に、他の参照基準を用いた場合よりも精度が高いこと、参加者の年齢が高いほど特異度がわずかに高くなること、標準的なカットオフ値≧ 10 点で感度と特異度の合計が最大になることなどが明らかにされている。

この研究が新たに加えた知見

2018 年 5 月までに実施された研究を検索した本研究の更新において、サンプルに 42 件の研究（27,146 名の追加の参加者）が追加されたにもかかわらず、全体的な感度と特異度の推定値は堅牢であり、以前の推定値と一致することが示された。

PHQ-9 の特異度は、診断を受けていない、または治療を受けていないことが確認され、実際にスクリーニングで選ばれた参加者のみを対象として推定した場合、わずかに良好であった。

すべての可能なカットオフ値において、半構造化面接の場合、特異度は高齢者の方が 0 ～ 12%（中央値 5%）高く、これはスクリーニングツールの精度が高齢者では低いかもしれないという仮定と矛盾するものである。

本研究で得られた知見に基づくナレッジ・トランスレーションツール（www.depressionscreening100.com/phq）は、地域における有病率の想定に基づき、異なるカットオフ値ごとにスクリーニングのアウトカムを得るために用いることができる。

また MINI よりも 2 ～ 14%（中央値 11%）高かった。参照基準とカットオフ値の全体にわたり、特異度は男性で 0 ～ 10%（中央値 3%）高く、60 歳以上の人々では 0 ～ 12%（中央値 5%）高くなった。

結論

研究者や臨床医は、www.depressionscreening100.com/phq のナレッジ・トランスレーション（利害関係者が健康や医療に関する意思決定のためにリサーチ・エビデンスを認識し、社会で利用できるようにすること）ツールを用いて、臨床現場ごとに異なる PHQ-9 のカットオフ値における陽性判定数及び偽陽性判定数の総数といったアウトカムを決定するために本研究の結果を活用できるであろう。

研究登録

PROSPERO CRD 42014010673

序論

うつ病は、他のどの病状よりも長い年月、健康的な生活を失う原因となる[1)-4)]。米国予防医療専門委員会（United States Preventive Services Task Force）によって、うつ症状があると認識されていない人々を特定するためのスクリーニングが推奨されている[5)]。ただし、カナダ予防医療専門委員会（Canadian Task Force on Preventive Health Care）[6)]や英国国家スクリーニング委員会（UK National Screening Committee）[7)]はそうした推奨は行っていない。うつ症状の質問紙調査は多くの目的で用いることができる。例えば、うつ病を有するかどうか不確かな患者の症状を評価・検討することや、治療効果や再発の発見をモニタリングすること、そしてうつ病のスクリーニングといった目的である[8)]。うつ病のスクリーニングでは、あらかじめ定めたカットオフ値を用いて質問紙調査を実施し、他の方法ではうつ病の可能性があるとは特定されていない患者をふるいにかけて陽性または陰性に分類し、さらに陽性の方にふるい分けられた患者が大うつ病の診断基準を満たしているかどうかを判断する[8)-13)]。

Patient Health Questionnaire-9（PHQ-9）[14)-16)]は、米国予防医療専門委員会などによってスクリーニングに用いることが推奨されている[11),17),18)]。精神疾患の診断・統計マニュアル（DSM）の大うつ病の診断基準である 9 種類の各症状[19)-22)]に一つずつ対応する質問項目が含まれており、回答者が過去 2 週間、どれくらいの頻度で悩まされていたかが、それぞれ 0 ～ 3 点のスコアで反映される。標準的なカットオフ値としては 10 点となっており、すなわち合計スコアが 10 点以上の場合、大うつ病の検出となる[14)-16),23)-25)]。

以前、我々が共同研究で実施した、個別参加者データ（計 58 件の研究から得られた計 17,357 名の参加者のデータ）を用いたメタアナリシス（IPDMA: Individual Participant Data Meta-Analysis）[25)]では、半構造化診断面接、構造化診断面接、精神疾患簡易構造化面接法（MINI: Mini International Neuropsychiatric Interview）ごとに PHQ-9 の精度を比較した。というのも、診断面接のタイプによってその性質やパフォーマンスに重要な違いがあるためである[26)-29)]。半構造化面接は、訓練を受けた臨床医が研究の場で臨床的な診断手順をできる限り忠実に再現できるよう設計されているのに対して、構造化面接は臨床的な判断を要することなく素人が実施できるように設計されている[30)-33)]。MINI は、迅速に実施でき包括的な内容となるよう設計された、簡潔な構造化面接である[34),35)]。その IPDMA では、半構造化面接を用いた場合の比較において、PHQ-9（カットオフ値：合計スコア 10 点以上）の感度は 0.88（95% 信頼区間：0.83 ～ 0.92）、特異度は 0.85（0.82 ～ 0.88）であることが示されている。感度については、年齢により統計的に（しかし最小限の）有意な差が認められたものの、他に参加者あるいは研究レベルのサブグループによる違いは認められなかった。

我々の目的は、より大規模な最新のデータセットを用いて PHQ-9 に関する IPDMA を更新することである。そのために、まず半構造化面接（一次分析）、（MINI 以外の）構造化面接、MINI それぞれの診断面接の比較の上で PHQ-9 のスクリーニング精度を評価し、次に参加者ごとの精度を調べ、特性の異なるサブグループについても検討する。本研究で行った更新では、PHQ-9 に関して我々が以前実施した IPDMA[25)] 以降に行われた 42 件の研究（参加者 27,146 名）の追加データが得られた。

方法

本研究で実施した IPDMA は、実施予定のシステマティックレビューの事前登録用国際レジストリ PROSPERO に登録し（登録番号 CRD42014010673）、そのプロトコルは公開されている[36)]。診断テストの精度に関する報告は、システマティックレビュー及びメタアナリシスのための優先的報告項目（PRISMA）の指針[37)]に沿って行った。また、個別参加者データに関しても PRISMA の指針[38)]に従った。この方法は、我々の以前の IPDMA[25)] と一致したものである。なお、PHQ-2[39)] 及び PHQ-8[40)] に関しては、別の IDPMA がすでに公表されている。

研究の適格基準

18 歳以上の参加者を対象として、PHQ-9 を実施後 2 週間以内に有効な半構造化または構造化面接を行い、DSM[19)-22)] または国際疾病分類（ICD）[41)] の基準に基づき大うつ病性障害または大うつ病エピソードの診断分類を行った研究を対象にした。18 歳未満の若年者や精神科の患者、既にうつ病の症状があると診断されている者を参加者としている研究は除外した。未診断のうつ病患者を特定するために実施されるのがスクリーニングであるため、精神科の患者や、既にうつ病の症状が確認されている患者は対象外とした。あらゆる言語のデータセットを対象とした。

一部の参加者が適格基準を満たさないデータセットについても、一次データを用いて適格な参加者を選択できる場合にはそれらを含めた。大うつ病の定義に関しては、DSM と ICD の両方から結果が得られた場合、または大うつ病性障害と大うつ病エピソードの両方の結果が得られた場合には、ICD よりも DSM を優先し、大うつ病性障害よりも大うつ病エピソードを優先した。これは、ほとんどの研究で DSM の分類が使用されていることと、スクリーニングが主にうつ病エピソードの検出を目的としているためである（エピソードが大うつ病性障害、双極性障害、持続性抑うつ障害などに関連しているかを判断するためには追加の面接が必要である）。

データベース分析と研究選択

Ovid Medline、Ovid Medline In-Process、Ovid Other Non-Indexed Citations、PsycINFO、Web of Science を介して、査読済み[42)] の検索戦略（補足方法 A を参照）を使用し、最初の検索では 2000 年 1 月 1 日から 2015 年 2 月を対象期間とし、今回の更新では 2018 年 5 月 9 日までを新たに含めて検索を行った。最初の検索対象期間を 2000 年からとしたのは、PHQ-9 が 2001 年に発表されたためである[14)]。さらに、関連するレビューの参考文献リストを精査し、未公表の研究に関する情報を得るため、著者に問い合わせを行った。検索結果は RefWorks（RefWorks-COS、米国メリーランド州ベセスダ）にアップロードした後、重複を排除し、残った引用文献は DistillerSR（Evidence Partners、カナダ、オタワ）にアップロードした。

2 名の独立した研究者がタイトルと抄録を精査した。いずれかの研究者が適格基準に該当しうると判断した研究について、2 名の研究者が独立して全文レビューを行い、意見の相違は話し合いのもと合議の上で解決し、必要な場合は第三者の研究者に相談した。チームメンバーが堪能な言語以外の場合は翻訳者に相談した。

データ提供、抽出、統合

適格なデータセットを持つ研究者に、個人情報を非特定化した一次データの提供を依頼した。責任著者に、必要に応じて最大 3 通の電子メールを送信した。これが成功しない場合は、共著者に電子メールを送り、さらには責任著者に電話での連絡を試みた。収集した個別参加者データは、標準化した形式に変換し、研究レベルデータと共に一つのデータセットに統合した。公表されている結果と提供された生データセットに不一致が見られた場合は、研究責任者に相談して解決した。

2 名の研究者が独立して、研究実施国、研究実施機関（非医療機関、プライマリーケア、入院施設、外来専門）及び診断面接に関する研究レベルのデータを公表されている論文から抽出した。不明確な点については、必要に応じて第三者の研究者への相談や、著者への問い合わせを行った。国連の人間開発指数（HDI）[43)] を基に、発表年の指数に基づき「非常に高い」「高い」「低中程度」の開発レベルに分類を行った。参加者の年齢、性別、大うつ病の面接状況、現在の精神保健診断または治療状況、そして PHQ-9 の得点を参加者レベルのデータとして含めた。一次研究 2 件で、複数の医療機関からの参加者にまたがっていたため、医療機関は参加者ごとにコード化した。可能な場合は、サンプリング手順を反映させるために、一次研究の重み付けを使用した。一次研究で重み付けが行われていないが、サンプリング手順により重み付けが正当化される場合は、逆選択確率を用いて重みを算出した。例えば、スクリーニング結果が陽性の参加者全員と陰性の参加者のランダムサブセットに診断面接を実施した研究では、重み付けが必要であった。

バイアスリスクの評価

一次論文の報告に基づいて、2 名の研究者が独立して QUality Assessment of Diagnostic Accuracy Studies-2（QUADAS-2）ツールを用いてバイアスリスクを評価した。不一致の解決には議論と合意を用い、必要に応じて第三者の研究者に相談した。補足方法 B は、本研究で使用した QUADAS-2 のコーディングルールを示している[44)]。

統計分析

4 種類の分析を行った。第一の分析では、二変量ランダム効果モデルを用いて、カットオフ値 5 点から 15 点における PHQ-9 の感度と特異度を 95% 信頼区間で推定した。半構造化面接（一次分析；Structured Clinical Interview for DSM[45)]、Schedules for Clinical Assessment in Neuropsychiatry[46)]、Depression Interview and Structured Hamilton[47)]）、構造化面

接（Composite International Diagnostic Interview[48]、Clinical Interview Schedule-Revised[49]、Diagnostic Interview Schedule[50]）、MINI[34),35)] の各診断面接を用いた研究ごとに分析を行った。

第二の分析では、二変量ランダム効果モデルを用いて、精神衛生上の問題を抱えているとの診断を受けていない、または治療を受けていないことが確認された参加者のみを対象に、PHQ-9 のカットオフ値における感度と特異度を推定し、全参加者の結果と比較した。この分析方法を選んだ理由は、PHQ-9 が様々な目的（スクリーニング、治療中の症状モニタリング、再発チェックなど）に使用可能であるにもかかわらず、主に未診断の大うつ病患者の同定のためのスクリーニングに用いられるからである。 すでに診断された患者や治療を受けている患者は精神医療以外の領域での一次研究に含まれることはあっても、実際にはスクリーニングの対象にはならない[51),52)]。しかし、全ての一次研究で過去のうつ病診断に関するデータが提供されているわけではないため、全参加者と、未診断または未治療であることが確認できた参加者の結果を比較した。 クラスター化されたブートストラップ法を用いて、精神衛生上の問題を抱えているとの診断やその治療を受けていない参加者と全参加者の間で、カットオフ値 5 点～ 15 点における感度と特異度の差に関する 95% 信頼区間を算出した[53),54)]。

第三の分析では、メタ回帰を用いて、参照基準カテゴリーと参加者のサブグループ間での感度と特異度の違いを調べた。参照基準カテゴリー（基準：半構造化）と PHQ-9 の精度係数（logit（感度）及び logit（1 －特異度））の交互作用を含む一段階の多変量メタ回帰モデルを適用し、二変量ランダム効果モデルによる結果と比較した。さらに、PHQ-9 の logit（感度）及び logit（1 －特異度）を、連続的に測定された参加者の特性、すなわち年齢、性別（基準：女性）、国の人間開発指数（基準：非常に高い）、参加者採用機関（基準：プライマリーケア）と交互作用させ、各参照基準カテゴリー内で多変量メタ回帰モデルを適用した。全 44,503 名の参加者のうち、18,316 名（41%）に併存疾患のデータがなく、また疾患ごとに評価を行うには単一の疾患（例：がん）に関するデータが含まれる研究が少なすぎたため、サブグループ分析に併存疾患を含めなかった。同様に、言語や国別の結果を分析するための研究がカテゴリー間で不十分だったため、これらの分析も行わなかった。

第四の分析では、Kent らの推奨[55)] に従い、第三の分析のメタ回帰モデルから、カットオフ値 5 点～ 15 点のすべてまたは大部分で 3 種類すべての参照基準カテゴリーに対して感度または特異度に有意な関連を示す各参加者特性を抽出し、それらに基づいて各サブグループ間に二変量ランダム効果モデルを適用した。この分析では、我々の以前の研究[25)] と同様に、年齢を 60 歳未満と 60 歳以上に分けて考慮した。大うつ病患者を含まない、またはうつ病を持たない患者が含まれない一次研究は、そのような研究を含めると二変量ランダム効果モデルが適用できなくなるため、各サブグループの解析から除外した。

すべての分析において、検査の感度と特異度の相関を考慮した二変量ランダム効果モデルを、ラプラス近似に対応する、評価点を 1 個としたガウス・エルミート適応求積法[56)] を使用して PHQ-9 データに適用し、全体の感度、特異度及び関連する 95% 信頼区間を得た。また、プールされた感度及び特異度の推定値に基づいて、各参照基準について経験的な受信者操作特性（ROC）曲線を作成し、その曲線下面積（AUC）を算出した。

各参照基準カテゴリーでの研究間及び各カテゴリー内の参加者サブグループ間の統計学的異質性を定量化するため、感度と特異度のフォレストプロットを作成した。診断検査精度のメタアナリシスにおける異質性のレベルを定量化するための確立された方法は存在しないが[37),57)]、我々は τ^2（感度と特異度に関するランダム効果の推定分散）、R（ランダム効果モデルを用いた場合の全体的な感度（または特異度）の推定標準偏差を対応する固定効果モデルを用いた場合の推定標準偏差で除した比）[58)] および新しい研究の未知の感度と特異度に関する 95% 予測区間を用いて異質性を定量化した。

さらに、第一の分析で得られた標準カットオフ値 10 点における感度及び特異度の推定値及び大うつ病の想定有病率を用いてノモグラムを作成し、PHQ-9 の陽性及び陰性的中率を推定した。

感度分析として、各参照基準カテゴリーに対して QUADAS-2 のシグナリングクエスチョンに基づく複数のメタ回帰モデルを適用し、バイアスリスクに基づいて、サブグループ間での結果の精度を比較した。これらの分析では、QUADAS-2 のシグナリングクエスチョンを、すべてのクエスチョンに対して logit（感度）及び logit（1 －特異度）と交互作用させた。バイアスリスクが「低」に分類された研究と、「高」または「不明確」に分類された研究の比較の上で、大うつ病の参加者が 100 名以上、うつ病でない参加者が 100 名以上いた研究を対象とした。さらに、主要な IPDMA の結果に対する影響を評価するため、追加の感度分析を行った。この分析では、データ提供がなかったが適格な精度のデータを公表している研究を含めた。

R[59)]（バージョン 4.0.0）と RStudio[60)]（バージョン 1.2.5042）を使用し、R パッケージ lme4[61)] 内の glmer 関数を使用してすべての分析を実行した。

患者及び一般市民の関与

研究課題、アウトカム指標、研究デザインの開発には患者は関与していない。研究開始後、Sarah Markham 博士が患者協力者として DEPRESSD グループに加わり，原稿のレビューを行った。

結果

検索結果及び対象として含めたデータセット

データベースの検索により、最初の検索と今回更新して行った検索を合わせて、9,670 報の論文のタイトルと抄録を特定した。これらの中から 9,199 報をタイトルと抄録のレビューの段階で除外し、残りを全文レビューした結果、さらに 297 報を除外した。結果として、適格基準を満たす 174 報の論文が残り、その中で 56 報はサンプルが重複していた。残りの 118 件の研究のうち 91 件（77%）が参加者のデータを提供しており、9 つの未公表研究データが追加で著者から提供されたため、合計 100 件の研究から参加者データが得られた。これらの研究には、参加者が合計 44,503 名含まれ、そのうち大うつ病を抱える者が 4,541 名（有病率は 10%。図 1）であった。今回更新した検索結果により、27,146 名の参加者を含む 42 件の研究が追加された。対象に含めたこれらの研究の特性と、適格基準を満たすもののデータが得られなかった研究の特性を、補足資料の表 B に示す。

表 1 に、参照基準ごとの参加者データを示す。対象として含めた 100 件の研究のうち、47 件（参加者 11,234 名、うち大うつ病 1,528 名）が半構造化面接を使用し、20 件（参加者 17,167 名、大うつ病 1,352 名）が構造化面接（MINI 以外）を使用し、33 件（参加者 16,102 名、大うつ病 1,661 名）が MINI を使用していた。最もよく使用された半構造化面接は Structured Clinical Interview for DSM（DSM 用構造化臨床面接）であり、最もよく使用された構造化面接は Composite International Diagnostic Interview（WHO 統合国際診断面接）であった。参加者をサブグループにより分類した結果を表 2 に示す。

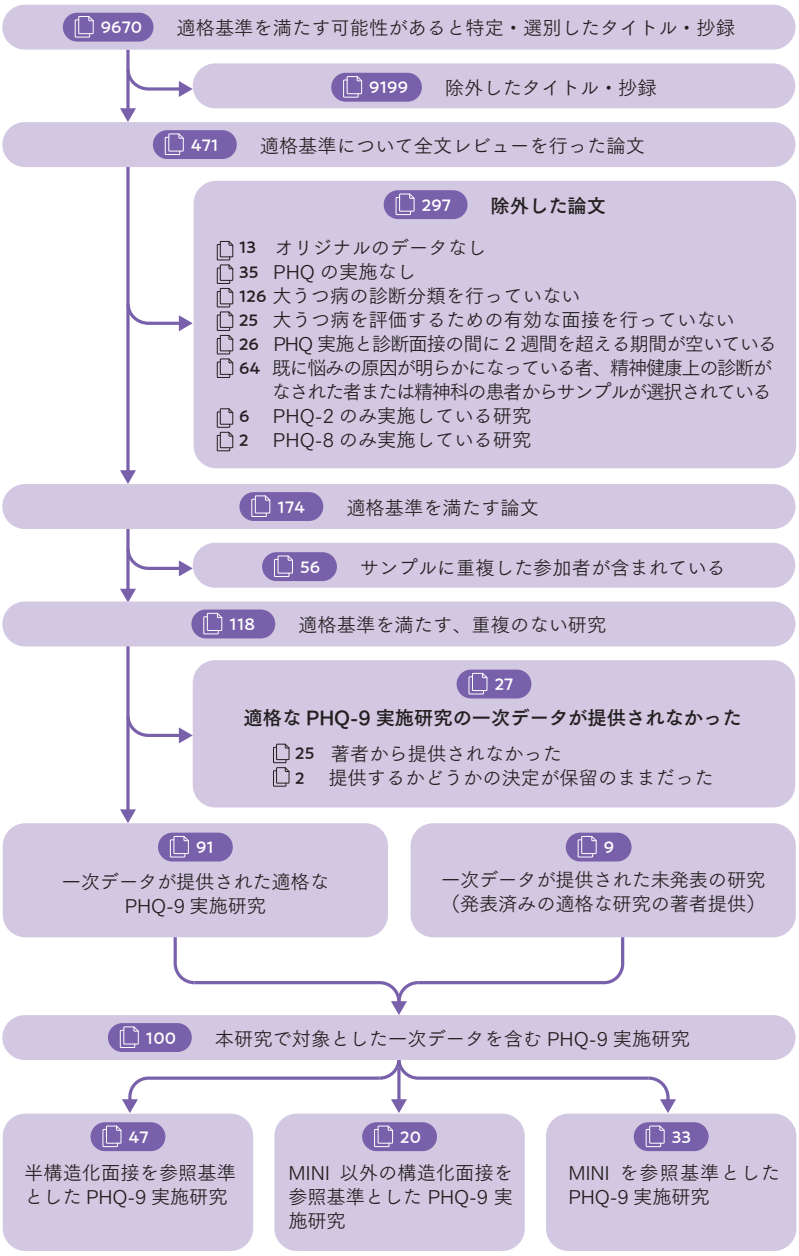


図 1　研究選択プロセスのフロー図

MINI = Mini International Neuropsychiatric Interview（精神疾患簡易構造化面接法）
PHQ = Patient Health Questionnaire

参照基準カテゴリーごとの PHQ-9 の感度及び特異度

表 3 に、各参照基準カテゴリーに対する PHQ-9 の感度と特異度の推定値を示す。半構造化面接、構造化面接（MINI を除く）、MINI の各参照基準に対して感度と特異度が最大化したのは、カットオフ値をそれぞれ≧ 10 点、≧ 8 点、≧ 8 点とした場合であった。標準的なカットオフ値≧ 10 点では、感度の推定値（95% 信頼区間）は半構造化面接で 0.85（0.79 ～ 0.89）、構造化面接で 0.64（0.53 ～ 0.74）、MINI で 0.74（0.67 ～ 0.79）であり、対応する特異度の推定値（95% 信頼区間）はそれぞれ 0.85（0.82 ～ 0.87）、0.88（0.83 ～ 0.92）、0.89（0.86 ～ 0.91）であった。PHQ-9 の感度は、すべてのカットオフ値において、参照基準として半構造化面接を用いた場

表 1　診断面接別の参加者データの分布

診断面接	研究数	参加者数	大うつ病の参加者数 (%)
半構造化面接：	47	11234	1528 (14)
Structured Clinical Interview for DSM	44	9242	1389 (15)
Schedules for Clinical Assessment in Neuropsychiatry	2	1892	130 (7)
Depression Interview and Structured Hamilton	1	100	9 (9)
構造化面接：	20	17167	1352 (8)
Composite International Diagnostic Interview	17	15759	1067 (7)
Diagnostic Interview Schedule	1	1006	221 (22)
Clinical Interview Schedule-revised	2	402	64 (16)
MINI	33	16102	1661 (10)
計	100	44503	4541 (10)

DSM = Diagnostic and Statistical Manual

合、構造化面接を用いた場合と比較して 7 ～ 24%（中央値 21%）高く、MINI を用いた場合と比較して 2 ～ 14%（中央値 11%）高かった。PHQ-9 の特異度はすべてのカットオフ値と参照基準にわたって同様であった。**図 2** は、各参照基準に対する受信者操作特性プロットと曲線下面積を示している。曲線下面積は、PHQ-9 を半構造化面接と併用した場合が最も高く（0.90）、次いで MINI（0.88）、そして MINI 以外の構造化面接（0.84）の順になった。

研究間の異質性は中程度の範囲から大きい範囲まで変動したが、いくつかのサブグループでは減少した（フォレストプロットは補足資料の図 A を参照。τ^2、R、予測区間は補足資料の表 C を参照）。カットオフ値≧ 10 点では、半構造化面接の場合、τ^2 値は感度については 0 ～ 6.97、特異度については 0 ～ 1.65 の範囲であり、構造化面接（MINI 除外）の場合、感度については 0.32 ～ 1.28、特異度については 0.32 ～ 1.48、MINI の場合、感度については 0.21 ～ 1.44、特異度については 0.07 ～ 0.71 の範囲であった。感度と特異度の 95% 予測区間は、補足資料表 E の対応する 95% 信頼区間よりもはるかに広く、研究間の中～高程度の異質性を同様に反映していた。

カットオフ値≧ 10 点における、仮想的な大うつ病有病率 5 ～ 25% の範囲での PHQ-9 の陽性的中率及び陰性的中率のノモグラムを**図 3** に示す。これらの理論的な有病率値において、陽性的中率は半構造化面接で 23% ～ 65%、構造化

表 2　サブグループ別の参加者データの分布 *

参加者のサブグループ	半構造化面接			構造化面接			MINI		
	研究数	参加者数	大うつ病の参加者数（%）	研究数	参加者数	大うつ病の参加者数（%）	研究数	参加者数	大うつ病の参加者数（%）
全参加者	47	11234	1528 (14)	20	17167	1352 (8)	33	16102	1661 (10)
精神衛生上の問題を抱えているとの診断を受けていない、または治療を受けていない参加者	26	3687	603 (16)	5	4001	289 (7)	15	8365	578 (7)
60 歳未満	42	7349	1131 (15)	20	13784	1087 (8)	31	10489	1119 (11)
60 歳以上	39	3860	397 (10)	15	3374	265 (8)	27	5585	533 (10)
女性	46	6986	1040 (15)	20	9603	793 (8)	32	9574	1126 (12)
男性	39	4168	488 (12)	18	7554	557 (7)	30	6511	534 (8)
人間開発指数が非常に高い国	38	9156	1047 (11)	16	15422	1149 (7)	21	10484	1108 (11)
人間開発指数が高い国	5	811	215 (27)	0	0	0	7	3753	237(6)
人間開発指数が低い、または中程度の国	4	1267	266 (21)	4	1745	203 (12)	5	1865	316 (17)
非医療機関	2	567	105 (19)	4	8219	371 (5)	9	7802	117 (15)
プライマリーケア	14	4566	683 (15)	7	4746	425 (9)	9	5063	543 (11)
入院施設	12	2355	257 (11)	2	593	72 (12)	3	473	106 (22)
外来専門	21	3746	483 (13)	7	3609	484 (13)	12	2634	511 (19)

* 研究レベルでコード化した変数もあれば、参加者レベルでコード化した変数もある。したがって、各参照基準について、研究数が必ずしも合計されるわけではない。

面接で 22% ～ 64%、MINI で 26% ～ 69% の範囲であった。対応する陰性的中率はそれぞれ 94% ～ 99%、88% ～ 98%、91% ～ 99% であった。

PHQ-9 の精度と 3 つの参照基準カテゴリー間の相関を示す複数のメタ回帰分析結果を補足資料表 D に示す。カットオフ値 5 点～ 15 点にわたって、参照基準と PHQ-9 の感度の間には有意な相関が見られた。半構造化面接を用いた場合では、構造化面接の場合よりも感度が 5 ～ 23%（中央値 19%）高く、MINI の場合よりも 1 ～ 15%（中央値 10%）高かった。すべてのカットオフ値にわたって、メタ回帰に基づき推定された差の大きさは、二変量ランダム効果モデルに基づく推定値と 2 ～ 3% 以内の差に収まっていた。いずれかのサブグループ分析から大うつ病の患者がいなかった、または大うつ病を有さない者がいなかったために除外された参加者の最大数は 63 であった。

個別参加者データが得られなかった 27 件の研究のうち、13 件が適格な精度のデータを公開していた（補足資料表 B2）。6 件は半構造化面接、2 件は構造化面接、5 件は MINI を使用していたが、大うつ病の有無の参加者数を公開していなかった 2 件の半構造化面接を用いた研究は感度分析から除外した。残りの 11 件の研究から得られた公表結果を含めても、補足資料表 D14 から D16 に示される結果は変わらなかった。

精神衛生上の問題で診断されていない、または治療を受けていない参加者を対象とした PHQ-9 と全参加者を対象とした PHQ-9 の精度の比較

精神衛生上の問題で診断されていない、または治療を受けていない参加者を対象とした PHQ-9 と、全参加者を対象とした PHQ-9 の精度の比較については、どの参照基準カテゴリーでも感度の推定値に有意な違いはなかった。しかし、特異度の推定値については、半構造化面接と MINI 面接で統計的に有意な差があった。精神衛生上の問題で診断されていない、または治療を受けていない参加者では、全参加者と比較して、半構造化面接を使用した場合の特異度がすべてのカットオフ値にわたって 1 ～ 4%（中央値 4%）高く、MINI を使用した場合は 1 ～ 6%（中央値 3%）高くなることが示された（補足資料表 E）。構造化面接を使用した場合の特異度には、有意な違いはなかった。

各サブグループにおける PHQ-9 の感度及び特異度並びにバイアスリスク

PHQ-9 の感度と特異度の参照基準との交互作用、そして参照基準内で層別化されたその他のサブグループ変数との交互作用を検討するメタ回帰分析の結果を、補足資料の表 D に示す。カットオフ値 5 点～ 15 点に対して、参照基準ごと、サブグループ変数ごとに、プールした感度及び特異度と関連する 95% 信頼区間を補足資料の表 E に示す。受信者操作特性（ROC）曲線とそれに対応する曲線下面積は、補足資料

表 3　半構造化面接、構造化面接、MINI の各参照基準における感度（95% 信頼区間）及び特異度（95% 信頼区間）の推定値の比較

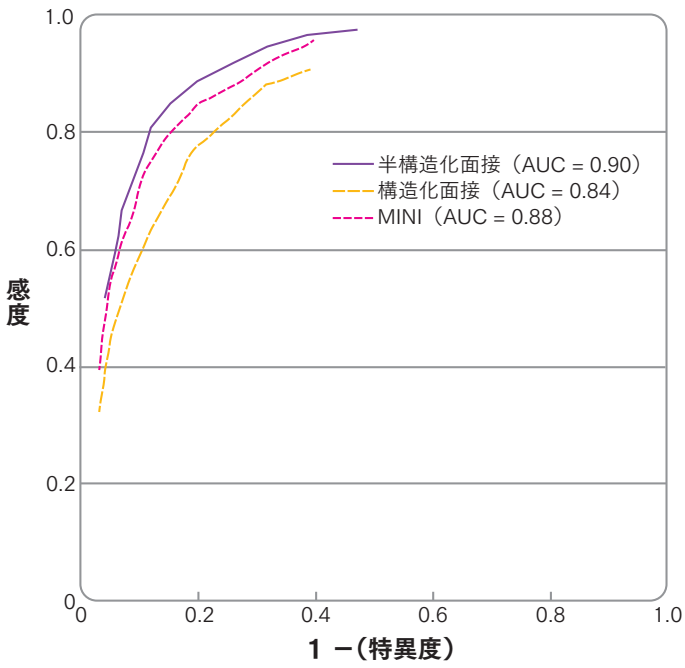
カットオフ値	参照基準：半構造化面接※		参照基準：構造化面接†		参照基準：MINI ‡	
	感度（95% 信頼区間）	特異度（95% 信頼区間）	感度（95% 信頼区間）	特異度（95% 信頼区間）	感度（95% 信頼区間）	特異度（95% 信頼区間）
5	0.98（0.95 ～ 0.99）	0.53（0.49 ～ 0.58）	0.91（0.85 ～ 0.95）	0.61（0.51 ～ 0.69）	0.96（0.93 ～ 0.97）	0.60（0.55 ～ 0.64）
6	0.97（0.94 ～ 0.98）	0.61（0.57 ～ 0.65）	0.88（0.80 ～ 0.93）	0.69（0.60 ～ 0.76）	0.92（0.89 ～ 0.95）	0.68（0.63 ～ 0.72）
7	0.95（0.92 ～ 0.98）	0.68（0.64 ～ 0.72）	0.82（0.73 ～ 0.89）	0.75（0.67 ～ 0.82）	0.88（0.83 ～ 0.92）	0.74（0.70 ～ 0.78）
8	0.92（0.88 ～ 0.95）	0.74（0.70 ～ 0.77）	0.77（0.66 ～ 0.86）	0.81（0.74 ～ 0.86）	0.85（0.79 ～ 0.89）	0.80（0.76 ～ 0.83）
9	0.89（0.84 ～ 0.92）	0.80（0.76 ～ 0.82）	0.69（0.59 ～ 0.78）	0.85（0.79 ～ 0.90）	0.80（0.73 ～ 0.85）	0.85（0.82 ～ 0.88）
10	0.85（0.79 ～ 0.89）	0.85（0.82 ～ 0.87）	0.64（0.53 ～ 0.74）	0.88（0.83 ～ 0.92）	0.74（0.67 ～ 0.79）	0.89（0.86 ～ 0.91）
11	0.81（0.75 ～ 0.86）	0.88（0.85 ～ 0.90）	0.57（0.46 ～ 0.67）	0.91（0.87 ～ 0.94）	0.67（0.60 ～ 0.73）	0.91（0.89 ～ 0.93）
12	0.75（0.69 ～ 0.80）	0.90（0.88 ～ 0.92）	0.52（0.41 ～ 0.63）	0.93（0.89 ～ 0.95）	0.61（0.54 ～ 0.68）	0.93（0.91 ～ 0.95）
13	0.67（0.61 ～ 0.72）	0.93（0.91 ～ 0.94）	0.45（0.35 ～ 0.56）	0.95（0.92 ～ 0.97）	0.55（0.47 ～ 0.62）	0.95（0.93 ～ 0.96）
14	0.61（0.55 ～ 0.67）	0.94（0.93 ～ 0.96）	0.39（0.30 ～ 0.50）	0.96（0.94 ～ 0.97）	0.47（0.41 ～ 0.54）	0.96（0.95 ～ 0.97）
15	0.52（0.46 ～ 0.58）	0.96（0.94 ～ 0.97）	0.32（0.24 ～ 0.41）	0.97（0.95 ～ 0.98）	0.40（0.35 ～ 0.46）	0.97（0.96 ～ 0.98）

MINI=Mini International Neuropsychiatric Interview
※ 研究数＝ 47、参加者数＝ 11234、大うつ病の参加者数＝ 1528
† 研究数＝ 20、参加者数＝ 17167、大うつ病の参加者数＝ 1352
‡ 研究数＝ 33、参加者数＝ 16102、大うつ病の参加者数＝ 1661

の図 B に示す。

参加者の年齢と性別のみ、3 つの参照基準カテゴリーすべてにおいて特異度と統計的に有意な相関がある変数であった（補足資料表 D）。60 歳以上の参加者の特異度は、より若い参加者よりも高かった。半構造化面接では 2 ～ 12%（中央値 6%）、構造化面接では 3 ～ 11%（中央値 6%）、MINI では 0 ～ 8%（中央値 2%）高く、すべての参照基準にわたって 0 ～ 12%（中央値 5%）の範囲であった。また、男性参加者の特異度は女性よりも高かった。メタ回帰から得られた差は、半構造化面接、構造化面接、MINI でそれぞれ 1 ～ 10%（中央値 4%）、1 ～ 7%（中央値 3%）、0 ～ 7%（中央値 3%）であり、すべての参照基準にわたって 0 ～ 10%（中央値 3%）の範囲であった。年齢と性別について、メタ回帰に基づく差の大きさは、二変量ランダム効果メタ分析モデルに基づく推定値と 1 ～ 5% 以内の差であった。

半構造化面接を用いた研究の結果に基づくと、いくつかのサブグループに対して感度と特異度の両方を最大化するカットオフ値は、≧ 10 点からわずかに変動した。年齢では、60 歳未満ではカットオフ値≧ 11 点、60 歳以上では≧ 10 点で最大値が得られた。性別では、女性では≧ 11 点、男性では≧ 9 点であった。しかし、これらの最大値はすべて、カットオフ値≧ 10 点で得られた値と 1 ～ 2% 以内の差であった。構造化面接や MINI についても同様の結果であった（補足資料表 E）。



AUC = 曲線下面積
MINI = Mini International Neuropsychiatric Interview（精神疾患簡易構造化面接法）

図 2　各カットオフ値及び各参照基準カテゴリーにおける感度及び特異度の推定値を示す受信者操作特性曲線

対象に含めたすべての一次研究の QUADAS-2 評価を、補足資料表 F に示す。395 の研究レベル項目のうち、12 が高リスクバイアス、130 が不明確、253 が低リスクバイアスと評価された。PHQ-9 の感度や特異度と一貫して関連する QUADAS-2 のシグナリングクエスチョンはなかった（補足資料表 D）。

考察

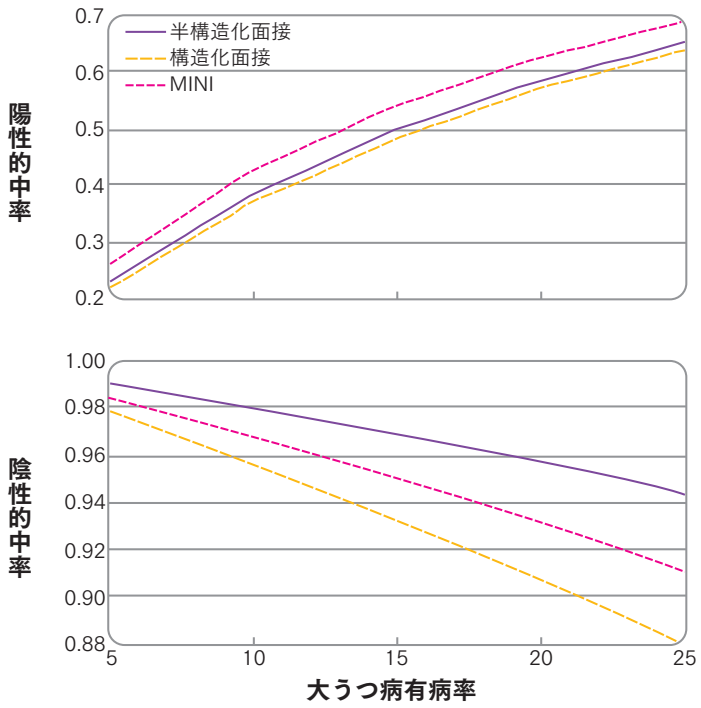
主な結果について

本研究では、大うつ病のスクリーニングのための PHQ-9 の精度を評価した。診断手順を最も忠実に再現できるようにデザインされた半構造化面接を用いた研究において、PHQ-9 の感度（85%）と特異度（85%）は標準のカットオフ値（≧ 10 点）で最大となることが示された。診断されていない、または治療を受けていない者のみを対象とした場合は、カットオフ値≧ 10 点で感度は変化せず、特異度は 89% に改善された。

性別と年齢は、3 種類のすべての診断面接において、PHQ-9 の特異度と有意な相関が見られた。PHQ-9 は 60 歳未満の参加者よりも 60 歳以上の参加者で、また女性よりも男性において特異的であった。感度については、年齢や性別と相関が見られなかった。サブグループによる精度の違いにより、感度や特異度を最大化するカットオフ値は異なるが、標準のカットオフ値である≧ 10 点との差はすべてにおいて最小であり、異なる特性の患者に対して異なるカットオフ値を使用することを正当化できるほど大きい差異ではない。

他の研究との比較

本研究の IPDMA は、以前の PHQ-9 の IPDMA[25] に比べて約 2 倍の数の一次研究、約 2.5 倍の数の参加者から得られたデータを含んでいた。多くの結果は類似するものであった。両研究とも、感度については、構造化面接及び MINI の各参照基準と比べて、半構造化面接の方が大幅に高かった。この結果は、半構造化面接と比べて、構造化面接及び MINI の各参照基準は、参加者や研究の特性を管理して相当数の偽陽性診断を生む[26)-29)] という知見と一致する。すなわち、診断結果が増すごとに、大うつ病を有する者の検出において、PHQ-9 の感度が低くなると考えられるが、それが正しいことを見いだした。半構造化面接を用いた研究では、感度と特異度はカットオフ値≧ 10 点で最大となった。高齢であることは、ほとんどのカットオフ値に対して、すべての参照基準にわたって統計的に有意に（しかし最小限に）特異度が高いことと関連していた。これは、高齢者におけるスクリーニン



MINI = Mini International Neuropsychiatric Interview（精神疾患簡易構造化面接法）

図３　大うつ病有病率５～25% に対して PHQ-9（カットオフ値 10）とともに半構造化面接、構造化面接または MINI を用いた場合の陽性・陰性的中率のノモグラム

グツールの精度が低いという推定に反しており、PHQ-9 の精度が同等かそれ以上であることを示唆している。

以前の IPDMA とは対照的に、PHQ-9 の特異度は、参照基準を問わず、女性よりも男性の方が高い可能性があることがわかった。また、半構造化面接または MINI を参照基準とした研究のうち、精神衛生上の問題について診断されていない、または治療を受けていない参加者のデータのみを検討した場合、特異度は有意に高いことがわかった（半構造化面接を採用した研究に基づくと、約 4% も高かった）。先行研究[51),52)]では、このような参加者を含めると精度にバイアスが生じる可能性があると予測されており、おそらく感度の向上によるバイアスが生じるであろうことが示唆されていた。代わりに、診断を受けているか治療を受けている可能性のある参加者を考慮すると特異度が下がることがわかり、このグループでは偽陽性のスクリーニングが発生することが示唆された。しかし、バイアスは小さく、この結果は構造化面接を実施した研究ではみられなかった。

本研究結果がもたらす意義

多くの研究では、感度と特異度の合計値が最大となるカットオフ値を報告しており、本研究においてもカットオフ基準値を設定するためにそれを行った。しかし、臨床試験や診療において、この基準に基づいてカットオフ値を選択する臨床的理由はない。カットオフ値が高ければ、うつ病でない参加者をより多く除外することができるが、大うつ病の基準を満たす参加者をより少なく検出することになる。反対に、カットオフ値が低ければ、診断基準を満たす参加者をより多く検出できるが、代償として大うつ病ではない人の偽陽性のスクリーニングが多くなる。理想としては、臨床的な意思決定は正しいスクリーニング結果と誤ったスクリーニング結果から生じる正味の利益や費用、害を考慮することである[62)]。このように、カットオフ値を選択することは、地域の価値観や資源、異なるカットオフ値におけるスクリーニング検査の陽性と陰性による結果に関する仮定に依存する。

理想としては、臨床試験において、異なるカットオフ値を用いた場合の結果について研究者や医師に知らせることができることである。しかし、我々の知る限りでは、うつ病スクリーニング試験[59)]のうち、参加者をスクリーニング群と、非スクリーニング群に無作為に割り当て、うつ病の可能性のある参加者の同定に PHQ-9 または 8 項目の PHQ-8（これらはほぼ同じものである）を用いた試験[40)]は 1 件しかない。Kronish ら[63)]は、PHQ-8 でカットオフ値≧ 10 点を用いて、過去 12 ヵ月間に急性冠症候群を発症した参加者を、プライマリーケア医にスクリーニング結果を通知する群、臨床医に通知した上で段階的なうつ病ケアを行う群、通知せずに通常のケアを行う群に振り分けた。既にうつ病の治療を受けている患者は除外した。しかし、標準的なカットオフ値≧ 10 点のスクリーニング検査で陽性であった参加者は 7% に過ぎなかった。スクリーニング検査で陽性であった参加者は各群で 40 名未満であり、そのうちの何名かは介入を完了していなかった可能性があり、スクリーニングの有益性は示されなかった。しかしながら、介入を提供できた可能性のあるスクリーニング陽性の参加者の数が少ないため、カットオフ値の妥当性について結論を出すことは困難である。より低いカットオフ値を用いれば、陽性者のスクリーニングが増加し、うつ病患者が増加する可能性があったかもしれないが、これはまたより多くの偽陽性のスクリーニングをもたらし、不必要な評価（設定により異なるが一般的には精神医療の専門家による）によって消費される資源が介入に使用されるかもしれない希少な資源を流用するため、より大変な評価の負担となったであろう。さらに、もし真陽性として検出される人数が増えていれば、介入から利益を得る可能性の低い、軽い症状の患者がスクリーニングされた可能性がある。もう一つの可能性として、カットオフの閾値を高くすることである。明らかな欠点としては、介入の対象となる人を同定するためにスクリーニングを受けなければならない人数が膨大になるこ

とである。

したがって、最も適切なカットオフ値を選択するための簡単な回答はなく、PHQ-9 を用いたスクリーニングを希望する研究者や臨床医は異なるカットオフ値における PHQ-9 の可能性を検討する必要がある。臨床医の実施を助けるために、我々はこの研究結果に基づくウェブベースのツール（depressionscreening100.com/phq）を作成した。このツールは異なる有病率と異なるカットオフ値に基づいて予想される陽性・陰性数、真のスクリーニング結果と偽のスクリーニング結果を推定するものである（下記「本研究の結果と実践について」参照）。

研究の強みと限界

従来のメタアナリシスと比較した際の、我々の IPMDA による PHQ-9 の精度評価法の強みは、(a) PHQ-9 と参照基準のデータを収集しつつ、精度の結果を公表していない研究データを統合したこと、(b) 適格な参加者と不適格な参加者（例えば、既に精神医療の治療を受けている人）を含む研究において、適格な参加者のみを選択することにより、適格な参加者データを含めたこと、(c) 大うつ病の有無に基づいてコード化することにより、複合参照基準（例えば、あらゆる精神疾患）に基づいて結果を公表した研究を含めたこと、(d) 参加者または研究の特性によるサブグループ分析を実施する能力（必要なデータ量のために、サブグループ分析を試みた一次研究はほとんどない）、(e) 対象として含めたすべての研究において、すべての関連するカットオフ値について分析を実施し、選択的なカットオフ値を報告することでバイアスを減らすことができることである[64),65)]。本研究の IPDMA には 44,503 名の参加者のデータが含まれ、我々の以前の IPDMA（N = 17,357）[25)] よりも 27,146 名が増加した。

本研究にはいくつかの限界がある。1 つ目に、127 件の適格基準を満たす研究のうち 27 件の研究（21%：適格参加者の 14%）のデータを含めることができなかった。とは言え、感度及び特異度を公表しているがデータを提供していない適格研究のデータを含めても、IPDMA の結果は変わらなかった。2 つ目に、サブグループ分析では少しは減少したものの、かなりの異質性が認められた。検査精度のメタアナリシスにおける異質性の推定の解釈の方法は十分に確立されておらず、本研究で用いた定量的指標の結果を解釈するための確立されたガイドラインはない。3 つ目に、参加者の 41% は医学的併存疾患のデータが欠落しており、ほとんどの言語と国では研究が少なかったため、これらのサブグループ分析を実施することができなかった。4 つ目に、一次研究は使用された診断面接によって分類されたが、面接者が必ずしも意図した通りに面接を行ったとは限らず、結果に影響を与えた可能性がある。5 つ目に、発表された研究で使用されたデータセットが適格であるかどうかの判断、著者への参加の呼びかけ、データ転送契約などデータ転送の手配、品質管理とデータハーモナイゼーション手順の実施に時間がかかるため、本 IPDMA に含まれる研究は 2018 年 5 月までに発表されたものであり、より最近の研究は含めることができなかった。

結論

PHQ-9 の感度と特異度は、半構造化面接を用いた場合、共に 85% であった。スクリーニングを行うのに実際上適格な者のみを考慮した場合、感度は変わらなかったが、特異度はより高かった（89%）。特異度は 60 歳以上の参加者や男性で高くなるようであるが、その差はサブグループごとにカットオフ値のしきい値を検討するほど大きくはない。PHQ-9 をスクリーニングに用いる臨床医は、感度と特異度、真陽性と偽陽性のスクリーニングについて、自らの優先事項やリソースの最適なバランスが得られるカットオフ値を選択するべきである。

本研究の結果と実践について

▶ PHQ-9 におけるカットオフ値≧ 10 点を含め、うつ病スクリーニングに一般的に使用される標準的なカットオフ値は、一般的に感度と特異度の合計を最大化するように選択される。

▶しかし、感度と特異度を最大にすることは、必ずしも患者の利益の可能性を最大にしたり、コストや害を最小にしたり、陽性者の評価を行える能力など地域ごとの懸念を反映したりするわけではない。

▶研究者及び臨床医は、真陽性と偽陽性、真陰性と偽陰性など、異なるアウトカムを用いて生じるスクリーニング結果を比較することで、臨床上の優先順位と地域のリソースに基づいてカットオフ値を選択することができる。

▶本研究で得られた知見に基づくナレッジ・トランスレーションツール（www.depressionscreening100.com/phq）を用いて、地域における有病率の想定に基づき、異なるカットオフ値ごとにスクリーニング結果を作成することができる。

著者の所属

1 Lady Davis Institute for Medical Research, Jewish General Hospital, Montréal, QC, Canada
2 Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montréal, WC, Canada
3 Centre for Prognosis Research, School of Medicine, Keele University, Keele, UK
4 Department of Psychiatry, McGill University, Montréal, QC, Canada
5 Department of Medicine, McGill University, Montréal, QC, Canada
6 Respiratory Epidemiology and Clinical Research Unit, McGill University Health Centre, Montréal, QC, Canada

参考文献

1) Moussavi S, Chatterji S, Verdes E, Tandon A, Patel V, Ustun B. Depression, chronic diseases, and decrements in health: results from the World Health Surveys. *Lancet* 2007;370:851-8. doi:10.1016/ S0140-6736(07)61415-9
2) Lopez AD, Mathers CD, Ezzati M, Jamison DT, Murray CJ. Global and regional burden of disease and risk factors, 2001: systematic analysis of population health data. *Lancet* 2006;367:1747-57. doi:10.1016/S0140-6736(06)68770-9
3) Mathers CD, Lopez AD, Murray CJL. The burden of disease and mortality by condition: Data, methods, and results for 2001. In: Lopez AD, Mathers CD, Ezzati M, Jamison DT, Murray CJL, eds. *Global Burden of Disease and Risk Factors*. The International Bank for Reconstruction and Development/The World Bank Group, 2006:45- 93.
4) Whiteford HA, Degenhardt L, Rehm J, et al. Global burden of disease attributable to mental and substance use disorders: findings from the Global Burden of Disease Study 2010. *Lancet* 2013;382:1575-86. doi:10.1016/S0140-6736(13)61611-6
5) Siu AL, Bibbins-Domingo K, Grossman DC, et al, US Preventive Services Task Force (USPSTF). Screening for depression in adults: US Preventive Services Task Force recommendation statement. *JAMA* 2016;315:380-7. doi:10.1001/jama.2015.18392
6) Joffres M, Jaramillo A, Dickinson J, et al, Canadian Task Force on Preventive Health Care. Recommendations on screening for depression in adults. *CMAJ* 2013;185:775-82. doi:10.1503/cmaj.130403
7) Allaby M. *Screening for depression: A report for the UK National Screening Committee (Revised report)*. UK National Screening Committee, 2010.
8) Thombs BD, Markham S, Rice DB, Ziegelstein RC. Does depression screening in primary care improve mental health outcomes?*BMJ* 2021;374:n1661. doi:10.1136/bmj.n1661
9) Palmer SC, Coyne JC. Screening for depression in medical care: pitfalls, alternatives, and revised priorities. *J Psychosom Res* 2003;54:279-87. doi:10.1016/S0022-3999(02)00640-2
10) Gilbody S, Sheldon T, Wessely S. Should we screen for depression?*BMJ* 2006;332:1027-30. doi:10.1136/bmj.332.7548.1027
11) Thombs BD, Coyne JC, Cuijpers P, et al. Rethinking recommendations for screening for depression in primary care. *CMAJ* 2012;184:413-8. doi:10.1503/cmaj.111035
12) Thombs BD, Ziegelstein RC. Does depression screening improve depression outcomes in primary care?*BMJ* 2014;348:g1253. doi:10.1136/bmj.g1253
13) Thombs BD, Ziegelstein RC, Roseman M, Kloda LA, Ioannidis JP. There are no randomized controlled trials that support the United States Preventive Services Task Force Guideline on screening for depression in primary care: a systematic review. *BMC Med* 2014;12:13. doi:10.1186/1741-7015-12-13
14) Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. *J Gen Intern Med* 2001;16:606-13. doi:10.1046/j.1525-1497.2001.016009606.x
15) Kroenke K, Spitzer RL. The PHQ-9: a new depression diagnostic and severity measure. *Psychiatr Ann* 2002;32:1-7. doi:10.3928/0048- 5713-20020901-06
16) Spitzer RL, Kroenke K, Williams JB, Primary Care Evaluation of Mental Disorders. Validation and utility of a self-report version of PRIME-MD: the PHQ primary care study. Primary Care Evaluation of Mental Disorders. Patient Health Questionnaire. *JAMA* 1999;282:1737-44. doi:10.1001/jama.282.18.1737
17) Maurer DM, Raymond TJ, Davis BN. Depression: Screening and Diagnosis. *Am Fam Physician* 2018;98:508-15.
18) American Academy of Family Physicians. Clinical preventive service recommendation. Depression. https://www.aafp.org/family-physician/patient-care/clinical-recommendations/all-clinical-recommendations/depression.html.
19) *Diagnostic and statistical manual of mental disorders: DSM-III*. 3rd ed, revised. American Psychiatric Association, 1987.
20) *Diagnostic and statistical manual of mental disorders: DSM-IV*. 4th ed. American Psychiatric Association, 1994.
21) *Diagnostic and statistical manual of mental disorders: DSM-IV*. 4th ed, text revised. American Psychiatric Association, 2000.
22) *Diagnostic and statistical manual of mental disorders: DSM-V*. 5th ed. American Psychiatric Association, 2013.
23) Wittkampf KA, Naeije L, Schene AH, Huyser J, van Weert HC. Diagnostic accuracy of the mood module of the Patient Health Questionnaire: a systematic review. *Gen Hosp Psychiatry* 2007;29:388-95. doi:10.1016/j.genhosppsych.2007.06.004
24) Gilbody S, Richards D, Brealey S, Hewitt C. Screening for depression in medical settings with the Patient Health Questionnaire (PHQ): a diagnostic meta-analysis. *J Gen Intern Med* 2007;22:1596-602. doi:10.1007/s11606-007-0333-y
25) Levis B, Benedetti A, Thombs BD, DEPRESsion Screening Data (DEPRESSD) Collaboration. Accuracy of Patient Health Questionnaire-9 (PHQ-9) for screening to detect major depression: individual participant data meta-analysis. *BMJ* 2019;365:l1476. doi:10.1136/bmj.l1476
26) Wu Y, Levis B, Ioannidis JPA, Benedetti A, Thombs BD, DEPRESsion Screening Data (DEPRESSD) Collaboration. Probability of Major Depression Classification Based on the SCID, CIDI, and MINI Diagnostic Interviews: A Synthesis of Three Individual Participant Data Meta-Analyses. *Psychother Psychosom* 2021;90:28-40. doi:10.1159/000509283
27) Levis B, Benedetti A, Riehm KE, et al. Probability of major depression diagnostic classification using semi-structured versus

fully structured diagnostic interviews. *Br J Psychiatry* 2018;212:377-85. doi:10.1192/bjp.2018.54

28) Levis B, McMillan D, Sun Y, et al. Comparison of major depression diagnostic classification probability using the SCID, CIDI, and MINI diagnostic interviews among women in pregnancy or postpartum: An individual participant data meta-analysis. *Int J Methods Psychiatr Res* 2019;28:e1803. doi:10.1002/mpr.1803

29) Wu Y, Levis B, Sun Y, et al. Probability of major depression diagnostic classification based on the SCID, CIDI and MINI diagnostic interviews controlling for Hospital Anxiety and Depression Scale - Depression subscale scores: An individual participant data meta-analysis of 73 primary studies. *J Psychosom Res* 2020;129:109892. doi:10.1016/j.jpsychores.2019.109892

30) Brugha TS, Jenkins R, Taub N, Meltzer H, Bebbington PE. A general population comparison of the Composite International Diagnostic Interview (CIDI) and the Schedules for Clinical Assessment in Neuropsychiatry (SCAN). *Psychol Med* 2001;31:1001-13. doi:10.1017/S0033291701004184

31) Brugha TS, Bebbington PE, Jenkins R. A difference that matters: comparisons of structured and semi-structured psychiatric diagnostic interviews in the general population. *Psychol Med* 1999;29:1013-20. doi:10.1017/S0033291799008880

32) Nosen E, Woody SR. Diagnostic Assessment in Research. In: McKay, D. *Handbook of research methods in abnormal and clinical psychology*. Sage, 2008:109-24.

33) Kurdyak PA, Gnam WH. Small signal, big noise: performance of the CIDI depression module. *Can J Psychiatry* 2005;50:851-6. doi:10.1177/070674370505001308

34) Lecrubier Y, Sheehan DV, Weiller E, et al. The Mini International Neuropsychiatric Interview (MINI). A short diagnostic structured interview: reliability and validity according to the CIDI. *Eur Psychiatry* 1997;12:224-31. doi:10.1016/S0924-9338(97)83296-8

35) Sheehan DV, Lecrubier Y, Sheehan KH, et al. The validity of the Mini International Neuropsychiatric Interview (MINI) according to the SCID-P and its reliability. *Eur Psychiatry* 1997;12:232-41. doi:10.1016/S0924-9338(97)83297-X

36) Thombs BD, Benedetti A, Kloda LA, et al. The diagnostic accuracy of the Patient Health Questionnaire-2 (PHQ-2), Patient Health Questionnaire-8 (PHQ-8), and Patient Health Questionnaire-9 (PHQ-9) for detecting major depression: protocol for a systematic review and individual patient data meta-analyses. *Syst Rev* 2014;3:124. doi:10.1186/2046-4053-3-124

37) Salameh JP, Bossuyt PM, McGrath TA, et al. Preferred reporting items for systematic review and meta-analysis of diagnostic test accuracy studies (PRISMA-DTA): explanation, elaboration, and checklist. *BMJ* 2020;370:m2632. doi:10.1136/bmj.m2632

38) Stewart LA, Clarke M, Rovers MP, et al, RISMA-IPD Development Group. Preferred Reporting Items for Systematic Review and Meta-Analyses of individual participant data: the PRISMA-IPD Statement. *JAMA* 2015;313:1657-65. doi:10.1001/jama.2015.3656

39) Levis B, Sun Y, He C, et al, Depression Screening Data (DEPRESSD) PHQ Collaboration. Accuracy of the PHQ-2 Alone and in Combination With the PHQ-9 for Screening to Detect Major Depression: Systematic Review and Meta-analysis. *JAMA* 2020;323:2290-300. doi:10.1001/jama.2020.6504

40) 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. doi:10.1017/S0033291719001314

41) The ICD-10 Classifications of Mental and Behavioural Disorder. *Clinical Descriptions and Diagnostic Guidelines Geneva*. World Health Organization, 1992.

42) PRESS – Peer Review of Electronic Search Strategies: 2015 Guideline Explanation and Elaboration (PRESS E&E). Ottawa: CADTH; 2016 Jan.

43) United Nations. International Human Development Indicators. http://hdr.undp.org/en/countries.

44) Whiting PF, Rutjes AW, Westwood ME, et al, QUADAS-2 Group. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011;155:529-36. doi:10.7326/0003-4819-155-8-201110180-00009

45) First MB. *Structured clinical interview for the DSM (SCID)*. John Wiley & Sons, Inc, 1995.

46) World Health Organization. *Schedules for clinical assessment in neuropsychiatry: manual*. Amer Psychiatric Pub Inc, 1994.

47) Freedland KE, Skala JA, Carney RM, et al. The Depression Interview and Structured Hamilton (DISH): rationale, development, characteristics, and clinical validity. *Psychosom Med* 2002;64:897-905.

48) Robins LN, Wing J, Wittchen HU, et al. The Composite International Diagnostic Interview. An epidemiologic Instrument suitable for use in conjunction with different diagnostic systems and in different cultures. *Arch Gen Psychiatry* 1988;45:1069-77. doi:10.1001/archpsyc.1988.01800360017003

49) Lewis G, Pelosi AJ, Araya R, Dunn G. Measuring psychiatric disorder in the community: a standardized assessment for use by lay interviewers. *Psychol Med* 1992;22:465-86. doi:10.1017/S0033291700030415

50) Robins LN, Helzer JE, Croughan J, Ratcliff KS. National Institute of Mental Health Diagnostic Interview Schedule. Its history, characteristics, and validity. *Arch Gen Psychiatry* 1981;38:381-9. doi:10.1001/archpsyc.1981.01780290015001

51) Thombs BD, Arthurs E, El-Baalbaki G, Meijer A, Ziegelstein RC, Steele RJ. 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. doi:10.1136/bmj.d4825

52) Rice DB, Thombs BD. Risk of bias from inclusion of currently diagnosed or treated patients in studies of depression screening tool accuracy: A cross-sectional analysis of recently published primary studies and meta-analyses. *PLoS One* 2016;11:e0150067. doi:10.1371/journal.pone.0150067

53) van der Leeden R, Busing FMTA, Meijer E. *Bootstrap methods for two-level models. Technical Report PRM 97-04*. Leiden University, Department of Psychology, 1997.

54) van der Leeden R, Meijer E, Busing FMTA. Resampling multilevel

models. In: Leeuw J, Meijer E, eds. *Handbook of multilevel analysis New York*. Springer, 2008: 401-33. doi:10.1007/978-0-387- 73186-5_11

55) Kent DM, Rothwell PM, Ioannidis JP, Altman DG, Hayward RA. Assessing and reporting heterogeneity in treatment effects in clinical trials: a proposal. *Trials* 2010;11:85. doi:10.1186/1745-6215-11-85

56) Riley RD, Dodd SR, Craig JV, Thompson JR, Williamson PR. Meta-analysis of diagnostic test studies using individual patient data and aggregate data. *Stat Med* 2008;27:6111-36. doi:10.1002/ sim.3441

57) Macaskill P, Gatsonis C, Deeks JJ, et al. Analysing and Presenting Results. In: Deeks JJ, Bossuyt PM, Gatsonis C, eds. *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 1.0*. Cochrane Collaboration, 2010.

58) Higgins JP, Thompson SG. Quantifying heterogeneity in a meta-analysis. *Stat Med* 2002;21:1539-58. doi:10.1002/sim.1186

59) R Core Team. (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

60) RStudio Team. (2020). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA URL https://www.rstudio.com/.

61) Bates D, Maechler M, Bolker B, et al. Fitting Linear Mixed-Effects Models Using lme4. *J Stat Softw* 2015;67:1-48. doi:10.18637/jss.v067.i01

62) Smits N, Smit F, Cuijpers P, De Graaf R. 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. doi:10.1002/mpr.230

63) Kronish IM, Moise N, Cheung YK, et al. Effect of depression screening after acute coronary syndromes on quality of life: the CODIACS-QoL randomized clinical trial. *JAMA Intern Med* 2020;180:45-53. doi:10.1001/jamainternmed.2019.4518

64) Levis B, Benedetti A, Levis AW. Selective cutoff reporting in studies of diagnostic test accuracy: a comparison of conventional and individual-patient-data meta-analysis of the Patient Health Questionnaire-9 depression screening tool. *Am J Epidemiol* 2017;185:954-64.

65) Neupane D, Levis B, Bhandari PM, et al . Selective cutoff reporting in studies of the accuracy of the Patient Health Questionnaire-9 and Edinburgh Postnatal Depression Scale: Comparison of results based on published cutoffs versus all cutoffs using individual participant data meta-analysis. *Int J Methods Psychiatr Res* 2021;30:e1873.

このパンフレットは、下記の英語論文の日本語翻訳である。

（英語論文内に記載されている補足資料（図、表）については、掲載していない）

Zelalem F Negeri et al.Accuracy of the Patient Health Questionnaire-9 for screening to detect major depression: updated systematic review and individual participant data meta-analysis. BMJ 2021;374:n2183. http://dx.doi.org/10.1136/bmj.n2183

DEPRESSD PHQ Group に所属する村松公美子が、研究代表者の Brett D Thombs 博士から、日本語に翻訳をする許諾を得て作成された。

AMED の課題番号 23rea522113h（※）の支援を得た。

※ R5 年度 予防・健康づくりの社会実装に向けた研究開発基盤整備事業：PRO のエビデンスの整理と関連医学会との連携基盤の構築

＜日本語翻訳作成者＞

村松公美子　新潟青陵大学・短期大学部保健管理センター長／福祉心理子ども学部特任教授　muramatu@n-seiryo.ac.jp

齋藤恵美　新潟青陵大学　福祉心理子ども学部准教授

小林大介　新潟青陵大学大学院　臨床心理学研究科助教

鴨志田冴子　山形大学保健管理センター助教

小岩広平　北海道教育大学大学院准教授

前田駿太　東北大学大学院教育学研究科准教授

PHQ-9（Patient Health Questionnaire-9）日本語版（2018）

この２週間、次のような問題にどのくらい頻繁に悩まされていますか？	全くない	数日	半分以上	ほとんど毎日
（A）物事に対してほとんど興味がない、または楽しめない	□	□	□	□
（B）気分が落ち込む、憂うつになる、または絶望的な気持ちになる	□	□	□	□
（C）寝付きが悪い、途中で目がさめる、または逆に眠り過ぎる	□	□	□	□
（D）疲れた感じがする、または気力がない	□	□	□	□
（E）あまり食欲がない、または食べ過ぎる	□	□	□	□
（F）自分はダメな人間だ、人生の敗北者だと気に病む、または自分自身あるいは家族に申し訳がないと感じる	□	□	□	□
（G）新聞を読む、またはテレビを見ることなどに集中することが難しい	□	□	□	□
（H）他人が気づくぐらいに動きや話し方が遅くなる、あるいは反対に、そわそわしたり、落ちつかず、ふだんよりも動き回ることがある	□	□	□	□
（Ｉ）死んだ方がましだ、あるいは自分を何らかの方法で傷つけようと思ったことがある	□	□	□	□

あなたが、いずれかの問題に１つでもチェックしているなら、それらの問題によって仕事をしたり、家事をしたり、他の人と仲良くやっていくことがどのくらい困難になっていますか？

全く困難でない	やや困難	困難	極端に困難
□	□	□	□

出典： Muramatsu K, Miyaoka H, Kamijima K et al.
Performance of the Japanese version of the Patient Health Questionnaire-9 (J-PHQ-9) for depression in primary care.
General Hospital Psychiatry. 52: 64-69, 2018.
新潟青陵大学大学院臨床心理学研究 , 第 7 号 ,p35-39, 2014

PHQ-9 日本語版（2018）の使用方法

PHQ-9 日本語版 2018 は、「大うつ病性障害」および「その他のうつ病性障害」に関する評価をするためのツールです。さらに評価項目の点数を合計することで、症状レベルを把握することができます。

1. この 2 週間の症状を、PHQ-9 日本語版 2018 用いて確認します。
2. 質問 A から H にチェックされた数から評価します。（質問票の網かけ部分がこの評価の対象となります）

大うつ病性障害	【半分以上】、【ほとんど毎日】に 5 つ以上のチェックがある場合（そのうちの 1 つは質問 1 または 2）
その他のうつ病性障害	【半分以上】、【ほとんど毎日】に 2 ～ 4 つのチェックがある場合（そのうちの 1 つは質問 1 または 2）

3. 質問 H は【数日】、【半分以上】、【ほとんど毎日】のいずれかにチェックがあった場合も 1 つと考えます。
4. 「大うつ病性障害」、「その他のうつ病性障害」は、躁病エピソードの既住、身体疾患、薬物に伴うものを除外して評価します。
5. 最下段の質問から、おおよその生活機能全般の困難度を評価します。

症状評価の目安

1 ～ 9 の各質問に対して、
それぞれ点数をつけます。

全くない	0 点
数日	1 点
半分以上	2 点
ほとんど毎日	3 点

総得点（0-27 点）により、
症状レベルを判定します。

0 - 4 点	軽微
5 - 9 点	軽度
10 - 14 点	中等度
15 - 19 点	中等度～重度
20 - 27 点	重度

・References

1）Muramatsu K, Miyaoka H, Kamijima K et al.
Performance of the Japanese version of the Patient Health Questionnaire-9 (J-PHQ-9) for depression in primary care.
General Hospital Psychiatry. 52: 64-69, 2018.

2）村松公美子. Patient Health Questionnaire (PHQ-9,PHQ15) B 本語版および Generalized Anxiety Disorder-7 日本語版—up to date.
新潟青陵大学大学院臨床心理学研究 , 第 7 号 ,p35-39, 2014

3）村松公美子 .Patient Health Questionnaire 日本語版シリーズ (PHQ, GAD) - うつと不安のメンタルヘルスアセスメント - 3-6 , 2021.
https:/ /n-seiryo.repo.nii.ac.jp/records/2014

2024 年 3 月 31 日発行 ／ 発行者・村松公美子 ／ 制作・金木犀舎