

Sensibilidade e especificidade do *Patient Health Questionnaire-9* (PHQ-9) entre adultos da população geral

Sensitivity and specificity of the *Patient Health Questionnaire-9* (PHQ-9) among adults from the general population

Sensibilidad y especificidad del Cuestionario de Salud del Paciente-9 (PHQ-9) entre adultos de la población general

Iná S. Santos ¹
 Beatriz Franck Tavares ²
 Tiago N. Munhoz ¹
 Laura Sigaran Pio de Almeida ²
 Nathália Tessele Barreto da Silva ²
 Bernardo Dias Tams ²
 André Machado Patella ²
 Alicia Matijasevich ¹

Abstract

This population-based study focused on the validity of the Patient Health Questionnaire-9 (PHQ-9) for screening major depressive episodes in the general population in Pelotas, Rio Grande do Sul State, Brazil. Households were selected by multi-stage sampling, and adults (≥ 20 years) were invited to participate. The gold standard was the structured diagnostic Mini International Neuropsychiatric Interview (MINI), applied by psychiatrists and psychologists. Both the PHQ-9 and the MINI were applied in the subjects' homes. In a total of 447 participants (191 men and 256 women), the continuous analysis identified ≥ 9 as the cutoff point with the highest sensitivity (77.5%; 61.5-89.2) and specificity (86.7%; 83.0-89.9). Use of the test's algorithm decreased the sensitivity to 42.5% (27.0-59.1), while the specificity increased to 95.3% (92.8-97.2). The PHQ-9 proved appropriate for screening major depressive episodes. For greater sensitivity, the continuously scored PHQ-9 proved more adequate than the algorithm for screening major depressive episodes in the community.

Depression; Sensitivity and Specificity; Validity of Tests; Questionnaires; Adult

Resumo

Com o objetivo de estudar a validade do Patient Health Questionnaire-9 (PHQ-9) no rastreio de episódio depressivo maior na população geral, conduziu-se um estudo de base populacional em Pelotas, Rio Grande do Sul, Brasil. Os domicílios foram sorteados por amostragem em múltiplos estágios, sendo os adultos (≥ 20 anos) convidados a participar. O padrão ouro foi a entrevista diagnóstica estruturada Mini International Neuropsychiatric Interview (MINI) aplicada por psiquiatras e psicólogos. Tanto o PHQ-9 quanto o MINI foram aplicados no domicílio. Em um total de 447 participantes (191 homens e 256 mulheres) a análise contínua identificou o ponto de corte ≥ 9 como de máxima sensibilidade (77,5%; 61,5-89,2) e especificidade (86,7%; 83,0-89,9). Usando o algoritmo do teste, houve diminuição da sensibilidade para 42,5% (27,0-59,1), enquanto que a especificidade aumentou para 95,3% (92,8-97,2). O PHQ-9 mostrou-se apropriado para rastreamento de episódio depressivo maior. Pela maior sensibilidade, o PHQ-9 pontuado de forma contínua mostrou-se mais adequado do que o algoritmo para rastreamento de episódio depressivo maior na comunidade.

Depressão; Sensibilidade e Especificidade; Validade dos Testes; Questionários; Adulto

¹ Programa de Pós-graduação em Epidemiologia, Universidade Federal de Pelotas, Pelotas, Brasil.
² Faculdade de Medicina, Universidade Federal de Pelotas, Pelotas, Brasil.

Correspondência
 T. N. Munhoz
 Programa de Pós-graduação em Epidemiologia, Universidade Federal de Pelotas.
 Rua Marechal Deodoro 1160, Pelotas, RS 96020-220, Brasil.
 tyagomunhoz@hotmail.com

Introdução

No Brasil, estudos de base populacional para avaliar a prevalência de depressão em populações não específicas são escassos. A maioria dos estudos tiveram como população alvo adolescentes, idosos ou mulheres no pós-parto. Na população geral, um inquérito conduzido pela Organização Mundial da Saúde (OMS), com 5.037 residentes na área metropolitana de São Paulo, mostrou que a prevalência de episódio depressivo maior, nos últimos 12 meses, entre indivíduos com 18 anos ou mais de idade, era de $10,4 \pm 0,6\%$, superior à média de $5,9\%$ obtida com a análise da prevalência de mais sete países de baixa e média renda, de diferentes continentes¹. Em Pelotas, Rio Grande do Sul, um estudo transversal de base populacional entre indivíduos com 20-69 anos de idade mostrou alta prevalência-ponto de sintomas depressivos como tristeza (29%), ansiedade (58%), falta de energia (37%), falta de disposição (40%), pensar no passado (34%) e preferir ficar em casa (54%)².

Estudos têm mostrado que a depressão é duas vezes mais frequente entre as mulheres do que entre os homens^{2,3} e entre os separados ou divorciados, em comparação aos que estão casados^{4,5}. A direção da associação com a idade depende do local onde o estudo foi realizado: em países de alta renda, a prevalência diminui com a idade⁶, enquanto que em países de baixa e média renda, parece ser mais elevada entre os mais velhos^{4,5}. Prevalências maiores têm sido também descritas entre os mais pobres^{7,8} e entre portadores de outras doenças crônicas⁹.

Estudos de base populacional apontam para uma ampla variação na prevalência de depressão entre diferentes países^{4,5}. Tal variação deve-se à combinação de vários fatores, entre eles o delineamento dos estudos, vulnerabilidade genética e fatores de risco ambientais. Fatores relativos ao instrumento de mensuração de depressão, como propriedades psicométricas e de validade dos testes, além de diferenças culturais na compreensão dos itens dos questionários, somam-se a essas razões.

A padronização do instrumento de medida é importante para o monitoramento da prevalência de doenças e fatores de risco. Entre os instrumentos usados para identificar indivíduos em risco de depressão, encontra-se o *Patient Health Questionnaire-9* (PHQ-9)¹⁰, que embora tenha tido suas propriedades para rastreamento testadas entre mulheres usuárias de atenção primária em Uberaba, Minas Gerais¹¹, ainda não foi validado na população geral no Brasil. O PHQ-9 é derivado do PRIME-MD¹², que foi originalmente desenvolvido para identificar

cinco transtornos mentais comuns em atenção primária à saúde: depressão, ansiedade, abuso de álcool, transtornos somatoformes e transtornos da alimentação. O PHQ-9 caracteriza-se por ser um instrumento de aplicação relativamente rápida, contendo nove questões, o que seria uma vantagem em estudos epidemiológicos, em comparação a outros atualmente validados para o Brasil, como por exemplo, o *Beck Depression Inventory* (BDI).

O conhecimento das propriedades dos testes, quanto à capacidade de identificar corretamente indivíduos em risco de apresentar depressão, é imprescindível para sua aplicação em estudos epidemiológicos, uma vez que permite corrigir as estimativas de prevalência da doença em função dos erros de classificação (falsos positivos e falsos negativos), decorrentes da acurácia imperfeita do teste. O objetivo desse estudo foi avaliar a sensibilidade e especificidade do PHQ-9 como instrumento de rastreamento de indivíduos adultos da população geral em maior risco de apresentar episódio depressivo maior.

Material e métodos

Um estudo transversal de base populacional foi realizado de fevereiro a junho de 2012, para avaliar a saúde de adolescentes, adultos e idosos, moradores da zona urbana do Município de Pelotas, uma cidade localizada na Região Sul do Brasil, com 328.275 habitantes. O processo de amostragem foi realizado em múltiplos estágios. Primeiramente foram selecionados os conglomerados, utilizando dados do *Censo Demográfico* de 2010, fornecidos pelo Instituto Brasileiro de Geografia e Estatística (IBGE. <http://www.ibge.gov.br>). Os 495 setores censitários da cidade foram ordenados por sua numeração, sendo essa estratégia baseada na localização geográfica. Os setores são numerados em formato de espiral, do centro para a periferia, em sentido horário. Com isso, a participação de diversos bairros da cidade, com diferentes níveis socioeconômicos, fica garantida na amostra.

Foram selecionados sistematicamente 130 setores censitários e, em cada um, cerca de 12 domicílios foram visitados para a pesquisa. Todos os residentes nos domicílios sorteados com dez anos ou mais de idade eram elegíveis. Os participantes foram entrevistados em casa, por entrevistadoras treinadas, através da aplicação de um questionário estruturado com itens sobre sua condição econômica, escolaridade, estado conjugal, cor da pele, trabalho, saúde e comportamentos. Todos os participantes responderam o questionário PHQ-9.

O estudo de validação foi realizado apenas entre os adultos. O processo de amostragem do estudo de validação foi conduzido semanalmente, a partir das entrevistas realizadas pelo estudo principal. Por sorteio aleatório simples, um de cada três domicílios incluídos no estudo principal era selecionado para o subestudo de validação. O encarregado do sorteio não tinha conhecimento do resultado do teste PHQ-9 aplicado no estudo principal. Em cada domicílio selecionado para o subestudo, todos os moradores com vinte anos ou mais de idade eram convidados a receber uma segunda visita para uma entrevista complementar. Essa segunda entrevista era conduzida por um profissional de saúde mental (psiquiatra, psicólogo ou residente de psiquiatria), previamente treinado na aplicação e interpretação do instrumento padrão ouro. Para não influenciar nas respostas, os participantes não tinham conhecimento da formação profissional desses entrevistadores.

O PHQ-9 constitui-se de nove perguntas que avaliam a presença de cada um dos sintomas para o episódio de depressão maior, descritos no *Manual Diagnóstico e Estatístico dos Transtornos Mentais* (DSM-IV) ¹³. Os nove sintomas consistem em humor deprimido, anedonia (perda de interesse ou prazer em fazer as coisas), problemas com o sono, cansaço ou falta de energia, mudança no apetite ou peso, sentimento de culpa ou inutilidade, problemas de concentração, sentir-se lento ou inquieto e pensamentos suicidas. A tradução do PHQ-9 para o português foi realizada por psiquiatras brasileiros e a *back translation* por um dos autores do instrumento original, em estudo publicado previamente ¹⁴. Esta versão está disponível online (<http://www.phqscreeners.com>, acessado em 12/Set/2012). A frequência de cada sintoma nas últimas duas semanas é avaliada em uma escala Likert de 0 a 3 correspondendo às respostas “nenhuma vez”, “vários dias”, “mais da metade dos dias” e “quase todos os dias”, respectivamente. O questionário ainda inclui uma décima pergunta que avalia a interferência desses sintomas no desempenho de atividades diárias, como trabalhar e estudar.

Na fase piloto do atual estudo, observou-se que os respondentes tinham dificuldade em diferenciar as opções originais de resposta “vários dias” e “mais da metade dos dias”. Como o instrumento se propõe a medir a frequência em número de dias em que as pessoas percebem a presença dos sintomas, as opções de resposta foram assim modificadas: de “nenhuma vez” para “nenhum dia”; de “vários dias” para “menos de uma semana”; e de “mais da metade dos dias” para “uma semana ou mais”. A opção “quase todos os dias” foi mantida inalterada.

O PHQ-9 pode ser tanto autoaplicado quanto aplicado por entrevistadores treinados ^{10,15}. Diferentes estudos de validação aplicaram o PHQ-9 através de entrevistadores ^{16,17,18}. Em uma revisão sistemática, os autores do instrumento relataram que o PHQ-9 apresenta desempenho semelhante nas diferentes formas de aplicação ¹⁷. Neste estudo o instrumento foi aplicado por entrevistadoras treinadas, na mesma ordem das perguntas no instrumento. Esta forma de aplicação foi escolhida levando-se em conta a baixa escolaridade da população em estudo. A Figura 1 apresenta a versão em português do PHQ-9 que foi empregada no atual estudo.

Para o cálculo do tamanho amostral empregaram-se como parâmetros sensibilidade e especificidade de 80%, erro aceitável de 10 pontos percentuais para mais ou para menos e nível de 95% de significância, sendo necessário incluir em torno de 200 sujeitos com e 200 sem episódio de depressão maior. Com uma prevalência-ponto de cerca de 30% de sintomas depressivos na população adulta de Pelotas ², estimou-se que com uma amostra de cerca de 600 indivíduos seria possível localizar em torno de 200 com episódio de depressão maior.

Como padrão ouro foi utilizada a entrevista diagnóstica estruturada *Mini International Neuropsychiatric Interview* (MINI) ¹⁹, validada no Brasil ²⁰. Esta entrevista foi desenvolvida para ser utilizada tanto na prática clínica como em estudos epidemiológicos e avalia a presença de transtornos mentais, de acordo com o DSM-IV e 10ª revisão da Classificação Internacional de Doenças (CID-10). Para os transtornos depressivos, possui sensibilidade e especificidade de 92% ²⁰. Estudos de comparação ao *Composite International Diagnostic Interview* (CIDI) e ao *Structured Clinical Interview for DSM Disorders* (SCID) demonstraram bons índices psicométricos para o diagnóstico de episódio depressivo maior, sendo um questionário diagnóstico mais breve e de aplicação mais simples no contexto de um estudo populacional ²¹. No atual estudo foi considerado como diagnóstico pelo padrão ouro a presença de episódio depressivo maior. Todos os indivíduos considerados positivos para episódio depressivo maior respondiam a um grupo adicional de questões, que investigavam outras causas possíveis para os sintomas, como efeitos diretos de substâncias, presença de transtorno mental orgânico ou outra enfermidade médica, presença de sintomas psicóticos ou se os sintomas seriam mais bem explicados por reação a luto, com o que o diagnóstico de episódio depressivo maior era descartado.

A análise dos dados incluiu o cálculo da sensibilidade (proporção de indivíduos com episódio

Figura 1

Versão em português do *Patient Health Questionnaire-9* (PHQ-9), extraída e modificada de Fraguas Jr. et al. 14.

AGORA VAMOS FALAR SOBRE COMO O(A) SR.(A) TEM SE SENTIDO NAS ÚLTIMAS DUAS SEMANAS.
<p>1) Nas últimas duas semanas, quantos dias o(a) sr.(a) teve pouco interesse ou pouco prazer em fazer as coisas?</p> <p>(0) Nenhum dia (1) Menos de uma semana (2) Uma semana ou mais (3) Quase todos os dias</p> <p>2) Nas últimas duas semanas, quantos dias o(a) sr.(a) se sentiu para baixo, deprimido(a) ou sem perspectiva?</p> <p>(0) Nenhum dia (1) Menos de uma semana (2) Uma semana ou mais (3) Quase todos os dias</p> <p>3) Nas últimas duas semanas, quantos dias o(a) sr.(a) teve dificuldade para pegar no sono ou permanecer dormindo ou dormiu mais do que de costume?</p> <p>(0) Nenhum dia (1) Menos de uma semana (2) Uma semana ou mais (3) Quase todos os dias</p> <p>4) Nas últimas duas semanas, quantos dias o(a) sr.(a) se sentiu cansado(a) ou com pouca energia?</p> <p>(0) Nenhum dia (1) Menos de uma semana (2) Uma semana ou mais (3) Quase todos os dias</p> <p>5) Nas últimas duas semanas, quantos dias o(a) sr.(a) teve falta de apetite ou comeu demais?</p> <p>(0) Nenhum dia (1) Menos de uma semana (2) Uma semana ou mais (3) Quase todos os dias</p> <p>6) Nas últimas duas semanas, quantos dias o(a) sr.(a) se sentiu mal consigo mesmo(a) ou achou que é um fracasso ou que decepcionou sua família ou a você mesmo(a)?</p> <p>(0) Nenhum dia (1) Menos de uma semana (2) Uma semana ou mais (3) Quase todos os dias</p> <p>7) Nas últimas duas semanas, quantos dias o(a) sr.(a) teve dificuldade para se concentrar nas coisas (como ler o jornal ou ver televisão)?</p> <p>(0) Nenhum dia (1) Menos de uma semana (2) Uma semana ou mais (3) Quase todos os dias</p> <p>8) Nas últimas duas semanas, quantos dias o(a) sr.(a) teve lentidão para se movimentar ou falar (a ponto das outras pessoas perceberem), ou ao contrário, esteve tão agitado(a) que você ficava andando de um lado para o outro mais do que de costume?</p> <p>(0) Nenhum dia (1) Menos de uma semana (2) Uma semana ou mais (3) Quase todos os dias</p> <p>9) Nas últimas duas semanas, quantos dias o(a) sr.(a) pensou em se ferir de alguma maneira ou que seria melhor estar morto(a)?</p> <p>(0) Nenhum dia (1) Menos de uma semana (2) Uma semana ou mais (3) Quase todos os dias</p> <p>10) Considerando as últimas duas semanas, os sintomas anteriores lhe causaram algum tipo de dificuldade para trabalhar ou estudar ou tomar conta das coisas em casa ou para se relacionar com as pessoas?</p> <p>(0) Nenhuma dificuldade (1) Pouca dificuldade (2) Muita dificuldade (3) Extrema dificuldade</p>

depressivo maior conforme o padrão ouro que foram corretamente identificados como tal pelo PHQ-9); especificidade (proporção de indivíduos sem episódio depressivo maior conforme o padrão ouro, que foram corretamente identificados como tal pelo PHQ-9); valor preditivo positivo (proporção de verdadeiros positivos dentre todos os que pontuaram positivo pelo PHQ-9); valor preditivo negativo (proporção de verdadeiros negativos dentre todos os que pontuaram negativo pelo PHQ-9); e razão de verossimilhança positiva (sensibilidade dividida pela diferença entre 1 e a especificidade). Foram calculados intervalos de 95% de confiança (IC95%) para todas as estimativas.

Nessas análises, a pontuação total do PHQ-9 foi calculada de duas maneiras: primeiro, usando o teste de forma contínua; e segundo, usando o algoritmo recomendado por seus autores. Quando calculado de forma contínua, somaram-se os valores correspondentes a cada resposta do participante na escala Likert. Já o algoritmo define o teste como positivo na presença de cinco ou mais sintomas, desde que pelo menos um seja humor deprimido ou anedonia, e que cada sintoma corresponda à resposta 2 ou 3 (“uma semana ou mais” e “quase todos os dias”, respectivamente), com exceção do sintoma 9 (Figura 1), para o qual é aceitável qualquer valor de 1 a 3 (“menos de uma semana”, “uma semana ou mais” e “quase todos os dias”, respectivamente).

Para o teste contínuo, o melhor ponto de corte foi definido através do índice de Youden, que corresponde aquele com maior valor na equação (sensibilidade + especificidade - 1). Adicionalmente, foi construída uma *Receiver Operating Characteristic Curve* (curva ROC), que é a representação gráfica da sensibilidade e de 1-especificidade de cada um dos possíveis pontos de corte do teste. O ponto de corte com máxima sensibilidade e especificidade na curva ROC foi definido como o menor valor da equação $[(1 - \text{sensibilidade})^2 + (1 - \text{especificidade})^2]$. A acurácia do PHQ-9 (proporção de resultados corretos, tanto positivos quanto negativos para depressão) foi estimada através da área sob a curva ROC.

Os estudos principal e de validação foram aprovados pelo Comitê de Ética em Pesquisa da Faculdade de Medicina, Universidade Federal de Pelotas de acordo com os protocolos 77/2011 e 14/2012, respectivamente. O consentimento livre e esclarecido foi assinado antes da coleta das informações, durante o estudo principal. Os indivíduos diagnosticados pelo padrão ouro foram atendidos no domicílio e/ou encaminhados para os serviços de saúde.

Resultados

Nos domicílios visitados foram identificados 533 indivíduos para a entrevista padrão ouro, dos quais 447 (84%) foram avaliados. As perdas deveram-se a 29 recusas; 51 indivíduos que não foram encontrados após três tentativas ou mais, em dias e horários diferentes; e seis que mudaram de cidade. As entrevistas com os profissionais de saúde mental (padrão ouro) foram realizadas, em média, 24 dias após a aplicação do PHQ-9. O intervalo mínimo foi de 0 dias e o máximo, de 93 dias, com mediana de 17 dias.

As características sociodemográficas da amostra são apresentadas na Tabela 1. Um total de 191 homens e 256 mulheres foram incluídos no estudo. A idade média da amostra foi de 43,8 anos (15,1) e mais de três quartos dos participantes (76,5%) declararam-se com cor de pele branca. Em relação às variáveis socioeconômicas, a média de renda familiar mensal em salários mínimos foi de 5,7 (7,3) e a média de escolaridade foi de 9,3 (4,5) anos de estudo. Mais da metade dos participantes trabalhavam por ocasião da primeira entrevista (58,8%) e viviam com companheiro(a) (65,5%).

Os indivíduos perdidos para a entrevista padrão ouro assemelhavam-se aos que foram entrevistados em todas as características investigadas exceto quanto à idade e a estar trabalhando por ocasião da aplicação do PHQ-9. A média de idade na amostra foi de 43,8 (15,1) anos contra 39,2 (13,1) entre as perdas ($p = 0,009$) e 41,2% dos entrevistados contra 60,4% das perdas estavam desempregados quando o PHQ-9 foi aplicado (Tabela 1).

A entrevista padrão ouro identificou quarenta indivíduos (32 mulheres e oito homens) (8,9%; IC95%: 6,3-11,6) apresentando episódio depressivo maior. Na Tabela 2 observam-se a sensibilidade e especificidade para cada um dos pontos de corte do PHQ-9 medido em escala contínua. Tanto o índice de Youden quanto o ponto de máximas sensibilidade e especificidade (Figura 2) apontam para o valor ≥ 9 como o mais adequado para identificar indivíduos em maior risco de estar com episódio depressivo maior. Um total de 85 indivíduos (19%; IC95%: 15,4-22,7) pontuaram ≥ 9 . A sensibilidade nesse ponto foi de 77,5% (IC95%: 61,5-89,2) e a especificidade, de 86,7% (IC95%: 83,0-89,9). Os valores preditivos positivo e negativo foram, respectivamente, 57,8% (IC95%: 53,2-62,4) e 94,3% (IC95%: 92,1-96,4). A razão de verossimilhança positiva nesse ponto foi de 5,8 (IC95%: 3,6-8,0) e a área sob a curva ROC indica uma acurácia do PHQ-9 de 86%.

O cálculo das propriedades diagnósticas do PHQ-9, obtido através do algoritmo (Tabela 3),

Tabela 1

Comparação entre as características da amostra incluída na validação do *Patient Health Questionnaire-9* (PHQ-9) (N = 447) e as perdas (n = 86). Pelotas, Rio Grande do Sul, Brasil, 2012.

Variáveis	Amostra n (%)	Perdas n (%)	Valor de p
Sexo feminino	256 (57,3)	47 (54,6)	0,65
Renda familiar mensal (média em salários mínimos)	5,7 (7,3)	4,8 (5,7)	0,29
Escolaridade (média de anos completos)	9,3 (4,5)	9,2 (3,9)	0,77
Idade (média de anos completos)	43,8 (15,1)	39,2 (13,1)	0,009
Cor da pele			0,61
Branca	342 (76,5)	70 (81,4)	
Preta	67 (15,0)	10 (11,6)	
Outra	38 (8,5)	6 (7,0)	
Vive com companheiro/a (sim)	293 (65,5)	52 (60,5)	0,37
Tabagismo			0,81
Nunca fumou	251 (56,1)	46 (53,5)	
Fuma atualmente	105 (23,5)	23 (26,7)	
Ex-fumante	91 (20,4)	17 (19,8)	
Trabalha atualmente (sim)	263 (58,8)	34 (39,6)	0,0009

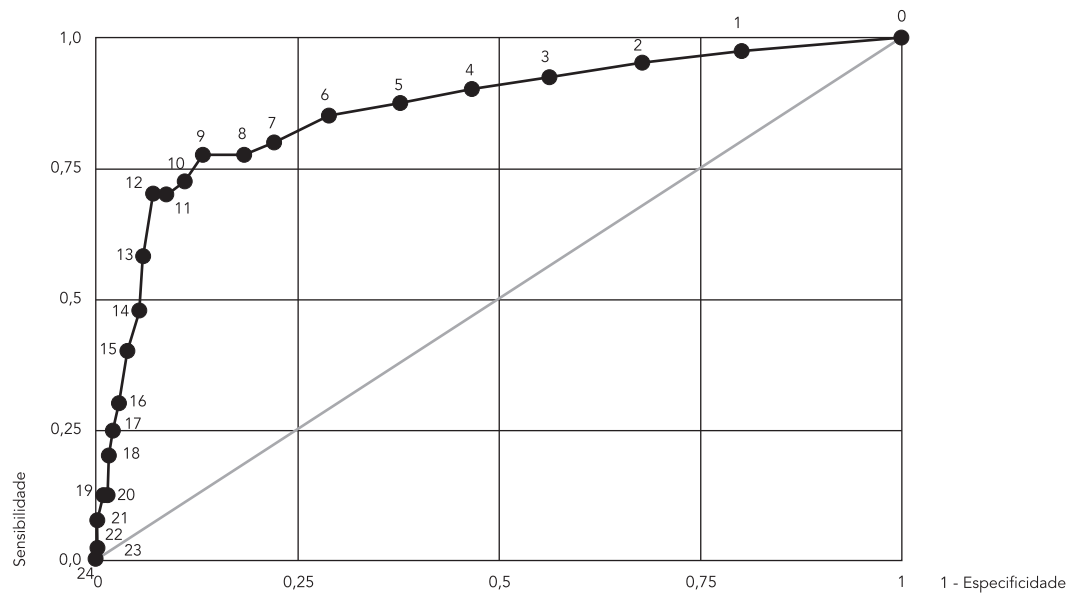
Tabela 2

Sensibilidade e especificidade conforme os pontos de corte do *Patient Health Questionnaire-9* (PHQ-9) em comparação ao *Mini International Neuropsychiatric Interview* (MINI) na população adulta (N = 447). Pelotas, Rio Grande do Sul, Brasil, 2012.

Pontos de corte	Sensibilidade (IC95%)	Especificidade (IC95%)	n (%)
≥ 1	97,5 (86,8-99,9)	19,9 (16,1-24,1)	365 (81,7)
≥ 2	95,0 (83,1-99,4)	32,4 (27,9-37,2)	313 (70,0)
≥ 3	92,5 (79,6-98,4)	43,7 (38,9-48,7)	266 (59,5)
≥ 4	90,0 (76,3-97,2)	53,3 (48,3-58,2)	226 (50,6)
≥ 5	87,5 (73,2-95,8)	62,4 (57,5-67,1)	188 (42,1)
≥ 6	85,0 (70,2-94,3)	71,0 (66,3-75,4)	152 (34,0)
≥ 7	80,0 (64,4-90,9)	77,6 (73,3-81,6)	123 (27,5)
≥ 8	77,5 (61,5-89,2)	81,6 (77,5-85,2)	106 (23,7)
≥ 9	77,5 (61,5-89,2)	86,7 (83,0-89,9)	85 (19,0)
≥ 10	72,5 (56,1-85,4)	88,9 (85,5-91,8)	74 (16,6)
≥ 11	70,0 (53,5-83,4)	91,2 (88,0-93,7)	64 (14,3)
≥ 12	70,0 (53,5-83,4)	92,9 (89,9-95,2)	57 (12,8)
≥ 13	57,5 (40,9-73,0)	94,1 (91,4-96,2)	47 (10,5)
≥ 14	47,5 (31,5-63,9)	94,6 (91,9-96,6)	41 (9,2)
≥ 15	40,0 (24,9-56,7)	96,1 (93,7-97,7)	32 (7,2)
≥ 16	30,0 (16,6-46,5)	97,1 (94,9-98,5)	24 (5,4)
≥ 17	25,0 (12,7-41,2)	97,8 (95,8-99,0)	19 (4,3)

Figura 2

Receiver operating characteristic curve (curva ROC) do Patient Health Questionnaire-9 (PHQ-9) na identificação de episódio depressivo maior, em comparação ao Mini International Neuropsychiatric Interview (MINI), na população adulta (N = 447). Pelotas, Rio Grande do Sul, Brasil, 2012.



Área sob a curva ROC = 0,8606

Tabela 3

Propriedades diagnósticas do Patient Health Questionnaire-9 (PHQ-9) com resultados obtidos através do algoritmo em comparação ao Mini International Neuropsychiatric Interview (MINI), na população adulta (N = 447). Pelotas, Rio Grande do Sul, Brasil, 2012.

PHQ-9	MINI		Total
	Positivo	Negativo	
Positivo	17	19	36
Negativo	23	388	411
Total	40	407	447

Especificidade: 95,3% (IC95%: 92,8-97,2); razão de verossimilhança positiva: 9,1 (IC95%: 5,2-16,1); sensibilidade: 42,5% (27,0-59,1); valor preditivo negativo: 94,4% (IC95%: 91,7-96,4); valor preditivo positivo: 47,2% (IC95%: 30,4-64,5).

mostrou uma sensibilidade de 42,5% (IC95%: 27,0-59,1) e especificidade de 95,3% (IC95%: 92,8-97,2). Os valores preditivos positivo e negativo foram, respectivamente, 47,2% (IC95%: 30,4-

64,5) e 94,4% (IC95%: 91,7-96,4) e a razão de verossimilhança positiva, de 9,1 (IC95%: 5,2-16,1).

Discussão

Principais achados

Esse estudo mostrou que o PHQ-9 é um teste apropriado para rastreamento de depressão entre adultos da população geral vivendo na área urbana de cidades de porte médio semelhantes a Pelotas. O uso do PHQ-9 em escala contínua mostrou-se mais útil como teste de rastreamento do que quando interpretado através do algoritmo. O uso do algoritmo corresponde aproximadamente ao ponto de corte ≥ 13 , o que significaria privilegiar a especificidade do teste, em detrimento da sensibilidade. O emprego do algoritmo foi mais eficiente, implicando o encaminhamento para avaliação especializada de um número menor de pessoas e com maior probabilidade de diagnosticar corretamente indivíduos com episódio depressivo maior do que no ponto de corte identificado na escala contínua (como indicado pela razão de verossimilhança positiva cerca de

60% maior para a positividade do teste obtida pelo algoritmo em comparação ao ponto de corte ≥ 9). Um teste de rastreamento, no entanto, deve privilegiar a sensibilidade, para identificar o maior número possível de indivíduos que necessitam de cuidados de saúde²². Pela alta especificidade, o uso do algoritmo mostrou-se mais útil para afastar a probabilidade diagnóstica de depressão, mas é pouco indicado como um teste de rastreamento, porque falhou em corretamente identificar cerca da metade dos indivíduos em risco de episódio depressivo maior.

Comparação com outros estudos

O PHQ-9 teve sua validade testada em vários níveis de atenção à saúde, como pacientes de clínicas de atenção primária, pacientes do nível secundário de atenção (com acidente vascular cerebral, clínicas ginecológicas, cardiopatas, usuários de drogas etc.), pacientes hospitalizados, pacientes ambulatoriais e pacientes de clínicas de medicina de família, em diversas línguas e contextos culturais. Uma meta-análise publicada em 2012²³, que incluiu avaliações do PHQ-9 comparativamente ao CID ou ao DSM para diagnóstico de transtorno depressivo maior, identificou 18 estudos, dos quais apenas dois foram realizados na comunidade^{24,25}.

O estudo de Adewuya et al.²⁴, entre estudantes universitários na Nigéria, com média de idade de 24 anos, a maioria homens, encontrou o valor ≥ 10 como o melhor ponto de corte para rastreamento de transtorno depressivo maior. O valor da sensibilidade nesse ponto (84,6%) foi semelhante ao encontrado em Pelotas (cujo IC95% variou de 61,5 a 89,2%), sendo a especificidade (99,4%) um pouco superior.

Gjerdengen et al.²⁵, nos Estados Unidos, entrevistaram mães em pós-parto, em consultas de puericultura. No ponto de corte ≥ 10 , a sensibilidade e especificidade no primeiro mês pós-parto foram, respectivamente, 75% e 91%. As mães foram reentrevistadas aos 2, 4, 6 e 9 meses pós-parto. A sensibilidade e especificidade para resultados do teste ≥ 10 , em qualquer dessas entrevistas foram 82% e 84%, respectivamente, sendo esses últimos mais semelhantes aos observados em Pelotas no ponto de corte ≥ 9 . Com o emprego do algoritmo, a sensibilidade foi de 67% (superior à observada em Pelotas) e a especificidade de 92% comparável à de Pelotas.

Esses dois estudos feitos na comunidade empregaram populações específicas e, portanto, mais homogêneas, o que faz com que o estudo de Pelotas seja o primeiro, até onde sabem seus autores, a avaliar as propriedades do PHQ-9 em base realmente populacional.

Limitações e aspectos positivos do estudo

Entre as limitações do estudo destaca-se a perda de 16% dos indivíduos que deveriam ter se submetido à entrevista padrão ouro. A comparação dos indivíduos do grupo de perdas com os que receberam a entrevista com os profissionais de saúde mental, no entanto, mostrou serem eles semelhantes em quase todas as características sociodemográficas investigadas, exceto quanto a serem mais jovens e a estarem desempregados na primeira entrevista, cuja prevalência foi maior. Adicionalmente, 19 dos 86 indivíduos perdidos tiveram PHQ-9 ≥ 9 , uma proporção superior à verificada na amostra estudada, possivelmente indicando uma maior prevalência de verdadeiros positivos nesse grupo. A perda desses indivíduos pode ter prejudicado a precisão da estimativa de sensibilidade do teste. De fato, extrapolando-se para as perdas os valores preditivos encontrados para o ponto de corte ≥ 9 do teste contínuo, a estimativa de sensibilidade do teste calculado pelo algoritmo passaria de 42,5% (IC95%: 27,0-59,1) para 50,9% (IC95%: 44,2-57,6), sem muita alteração no valor da especificidade (de 95,3% para 94,3%).

Uma segunda limitação é decorrente do tempo desde a aplicação do PHQ-9 até a realização da entrevista padrão ouro. Tanto o PHQ-9 quanto os critérios do DSM-IV referem-se à presença de sintomas depressivos nos quinze dias anteriores à entrevista, não tendo esse período coincidido total ou parcialmente para 53,5% dos participantes. Como a entrevista padrão ouro foi sempre realizada após a aplicação do PHQ-9 e, na maioria das vezes, em dia diferente, é possível que tenham sido erroneamente classificados como falsos alguns resultados desse último, diminuindo sua acurácia. No entanto, uma vez manifestos, os sintomas depressivos costumam persistir por semanas. Em uma coorte holandesa de base populacional que tinha entre os objetivos investigar a duração de episódio depressivo maior desde sua incidência (primeiro episódio ou início de episódio recorrente), a duração mediana foi de três meses, enquanto que 20% apresentaram um curso crônico (24 meses ou mais)²⁶. Além disso, alguns estudos relataram uma boa repetibilidade do PHQ-9 no espaço de duas semanas¹⁷ e um mês²⁴ e análises que exploraram a influência do período de tempo decorrido entre o PHQ-9 e o padrão ouro encontraram áreas similares sob a curva ROC, mesmo para intervalos maiores²⁷.

Adicionalmente, estudos epidemiológicos têm sistematicamente documentado maiores prevalências de depressão entre as mulheres do que entre os homens, o que tem sido atribuído, entre outras razões, ao papel social de gênero (di-

ferenças quanto a estressores típicos, capacidade de lidar com os estressores e maior oportunidade aos homens de expressar o estresse psicológico) e sua repercussão na forma de sentir e elaborar as experiências de vida^{2,3,28}. Embora, teoricamente, a sensibilidade e especificidade de um teste não se alterem com a prevalência do desfecho, em grupos com menores prevalências, a sensibilidade deve diminuir^{22,29}. Tais fatos justificariam uma análise do desempenho do teste estratificado por sexo. Sem perder de vista a baixa precisão das estimativas de sensibilidade decorrentes do pequeno tamanho de amostra, realizou-se uma análise exploratória separadamente para homens e mulheres. Entre as mulheres, empregando o algoritmo, a sensibilidade do PHQ-9 foi de 43,8% (IC95%: 26,4-62,3) e a especificidade, de 92,9% (IC95%: 88,7-95,9). Entre os homens, a sensibilidade e especificidade foram, respectivamente, de 37,5% (IC95%: 8,5-75,5) e 98,4% (IC95%: 95,3-99,7). Segundo o índice de Youden e o ponto de máximas sensibilidade e especificidade na curva ROC, o melhor ponto de corte para identificação de mulheres em risco de estarem deprimidas foi ≥ 12 , com sensibilidade de 75% (IC95%: 56,6-88,5) e especificidade de 88,8% (IC95%: 84,0-92,6), maior que o observado por Lima Osório et al.¹¹ entre usuárias de atenção primária (ponto de corte 10; sensibilidade 100%; especificidade 98%). Entre os homens, o melhor ponto de corte foi ≥ 6 , com sensibilidade de 87,5 (IC95%: 47,3-99,7) e especificidade de 79,8 (IC95%: 73,2-85,3). No entanto, como já chamado atenção, os valores de sensibilidade precisam ser confirmados em estudos maiores, que aumentem a precisão das estimativas. Até lá, o recomendável seria empregar o ponto de corte ≥ 9 tanto para homens quanto para mulheres.

Finalmente, a prevalência ponto esperada de episódio depressivo maior foi sobre-estimada para o cálculo do tamanho da amostra. Como ficou claro após a realização do estudo, para se localizar 200 indivíduos com episódio depressivo maior, seria necessário entrevistar em torno de 2 mil adultos da população geral. Como já mencionado anteriormente, essa limitação afetou a precisão das estimativas de sensibilidade, alargando seu intervalo de confiança.

Os aspectos positivos do estudo incluem o fato de o PHQ-9 haver sido aplicado por uma equipe de entrevistadores gerais e a entrevista diagnóstica ter sido realizada em momento diferente, no mesmo contexto (domicílio do participante) por profissionais de saúde mental. Além disso, os profissionais desconheciam o resultado do PHQ-9 de cada indivíduo, o que garantiu seu cegamento.

Aspectos adicionais

A adoção do ponto de corte identificado nesse estudo deve ser feita com cuidado ao aplicar-se o teste em outros locais, uma vez que a sensibilidade e especificidade do instrumento deve variar em contextos diferentes³⁰. Idealmente, antes da adoção de qualquer ponto de corte, deve-se ter em mente se as características dos indivíduos em quem se pretende aplicar o teste são semelhantes às daqueles em que o teste teve sua acurácia testada.

Conclusão

O PHQ-9 mostrou-se um instrumento válido para ser usado no rastreamento de episódio depressivo maior em estudos epidemiológicos. Como instrumento de rastreamento, o PHQ-9 apenas aponta os indivíduos em maior risco de estar apresentando episódio depressivo maior. O diagnóstico definitivo da doença somente poderá ser firmado através da consulta com profissionais de saúde mental. O uso de um mesmo instrumento com propriedades diagnósticas conhecidas é importante para a monitorização da prevalência de doenças com incidência crescente na comunidade, como tem alertado a OMS para o caso da depressão³¹. O PHQ-9 é um teste simples, rápido, que pode ser aplicado por entrevistadores treinados e cujas propriedades diagnósticas, demonstradas no atual estudo, permitem recomendá-lo para uso em estudos populacionais para rastreio de depressão.

Resumen

Con el objetivo de evaluar la validez del Cuestionario de Salud del Paciente-9 (PHQ-9) en la detección de un episodio depresivo mayor en la población general, se llevó a cabo un estudio de base poblacional en la ciudad de Pelotas, Río Grande do Sul, Brasil. Los hogares se seleccionaron al azar a través de muestreo multietápico, siendo solo invitados a participar los adultos (≥ 20 años). El padrón de primer orden fue la Mini Entrevista Neuropsiquiátrica Internacional (MINI), aplicada por psiquiatras y psicólogos. Ambos instrumentos fueron aplicados en la casa del participante. De un total de 447 participantes (191 hombres y 256 mujeres), el análisis continuo mostró un punto de corte ≥ 9 como el de máxima sensibilidad (77,5%; 61,5-89,2) y especificidad (86,7%; 83,0-89,9) para la identificación de un episodio depresivo mayor. Al utilizar el algoritmo del test, la sensibilidad disminuyó a un 42,5% (27,0-59,1), mientras que la especificidad aumentó a un 95,3% (92,8-97,2). El PHQ-9 mostró ser adecuado para el cribado de un episodio depresivo mayor. Debido a la sensibilidad más alta, el PHQ-9 marcado como una variable continua fue más adecuado que el algoritmo para la detección de episodio depresivo mayor en la comunidad.

Depresión; Sensibilidad y Especificidad; Validez de las Pruebas; Cuestionarios; Adulto

Colaboradores

I. S. Santos, B. F. Tavares, T. N. Munhoz e A. Matijasevich participaram da concepção e desenho do estudo, análise, interpretação dos dados, redação de todas as versões do artigo. L. S. P. Almeida, N. T. B. Silva, B. D. Tams e A. M. Patella participaram da concepção, coleta de dados e revisão crítica do artigo. Todos os autores aprovaram a versão final do manuscrito.

Agradecimentos

Ao PRONEX-CNPq pelo financiamento.

Referências

1. Bromet E, Andrade LH, Hwang I, Sampson NA, Alonso J, Girolamo G, et al. Cross-national epidemiology of DSM-IV major depressive episode. *BMC Med* 2011; 9:90.
2. Rombaldi AJ, Silva MC, Gazalle FK, Azevedo MR, Hallal PC. Prevalência e fatores associados a sintomas depressivos em adultos do sul do Brasil: estudo transversal de base populacional. *Rev Bras Epidemiol* 2010; 13:620-9.
3. Van de Velde S, Bracke P, Levecque K. Gender differences in depression in 23 European countries. Cross-national variation in the gender gap in depression. *Soc Sci Med* 2010; 71:305-13.
4. Andrade L, Caraveo-Anduaga JJ, Berglund P, Bijl RV, De Graaf R, Vollebergh W, et al. The epidemiology of major depressive episodes: results from the International Consortium of Psychiatric Epidemiology (ICPE) Surveys. *Int J Methods Psychiatr Res* 2003; 12:3-21.
5. Weissman MM, Bland RC, Canino GJ, Faravelli C, Greenwald S, Hwu HG, et al. Cross-national epidemiology of major depression and bipolar disorder. *JAMA* 1996; 276:293-9.
6. Blazer DG, Kessler RC, McGonagle KA, Swartz MS. The prevalence and distribution of major depression in a national community sample: the National Comorbidity Survey. *Am J Psychiatry* 1994; 151:979-86.
7. Everson SA, Maty SC, Lynch JW, Kaplan GA. Epidemiologic evidence for the relation between socioeconomic status and depression, obesity, and diabetes. *J Psychosom Res* 2002; 53:891-5.
8. Mendoza-Sassi R, Beria JU, Fiori N, Bortolotto A. Prevalência de sinais e sintomas, fatores socio-demográficos associados e atitude frente aos sintomas em um centro urbano no Sul do Brasil. *Rev Panam Salud Pública* 2006; 20:22-8.

9. Boing AF, Melo GR, Boing AC, Moretti-Pires RO, Peres KG, Peres MA. Associação entre depressão e doenças crônicas: um estudo populacional. *Rev Saúde Pública* 2012; 46:617-23.
10. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. *J Gen Intern Med* 2001; 16:606-13.
11. Lima Osório F, Vilela Mendes A, Crippa JA, Loureiro SR. Study of the discriminative validity of the PHQ-9 and PHQ-2 in a sample of Brazilian women in the context of primary health care. *Perspect Psychiatr Care* 2009; 45:216-27.
12. Spitzer RL, Williams JB, Kroenke K, Linzer M, Gryg 3rd FV, Hahn SR, et al. Utility of a new procedure for diagnosing mental disorders in primary care: the PRIME-MD 1000 study. *JAMA* 1994; 272:1749-56.
13. American Psychiatric Association Diagnostic and statistical manual of mental disorders: DSM-IV-TR. 4th Ed. Washington DC: American Psychiatric Association; 2000.
14. Fraguas Jr. R, Henriques Jr. SG, De Lucia MS, Iosifescu DV, Schwartz FH, Menezes PR, et al. The detection of depression in medical setting: a study with PRIME-MD. *J Affect Disord* 2006; 91:11-7.
15. Arroll B, Goodyear-Smith F, Crengle S, Gunn J, Kerse N, Fishman T, et al. Validation of PHQ-2 and PHQ-9 to screen for major depression in the primary care population. *Ann Fam Med* 2010; 8: 348-53.
16. Hyphantis T, Kotsis K, Voulgari PV, Tsifetaki N, Creed F, Drosos AA. Diagnostic accuracy, internal consistency, and convergent validity of the Greek version of the patient health questionnaire 9 in diagnosing depression in rheumatologic disorders. *Arthritis Care Res (Hoboken)* 2011; 63:1313-21.
17. Kroenke K, Spitzer RL, Williams JB, Löwe B. The Patient Health Questionnaire Somatic, Anxiety, and Depressive Symptom Scales: a systematic review. *Gen Hosp Psychiatry* 2010; 32:345-59.
18. Yu X, Tam WW, Wong PT, Lam TH, Stewart SM. The Patient Health Questionnaire-9 for measuring depressive symptoms among the general population in Hong Kong. *Compr Psychiatry* 2012; 53:95-102.
19. Sheehan DV, Lecrubier Y, Sheehan KH, Amorim P, Janavs J, Weiller E, et al. The Mini-International Neuropsychiatric Interview (M.I.N.I.): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *J Clin Psychiatry* 1998; 59 Suppl 20:22-33.
20. Azevedo Marques JM, Zuairi AW. Validity and applicability of the Mini International Neuropsychiatric Interview administered by family medicine residents in primary health care in Brazil. *Gen Hosp Psychiatry* 2008; 30:303-10.
21. Amorim P. Mini International Neuropsychiatric Interview (MINI): validação de entrevista breve para diagnóstico de transtornos mentais. *Rev Bras Psiquiatr* 2000; 22:106-15.
22. Szklo M, Javier Nieto F. *Epidemiology: beyond the basics*. 2nd Ed. Sudbury: Jones and Bartlett Publishers; 2007.
23. Manea L, Gilbody S, McMillan D. Optimal cut-off score for diagnosing depression with the Patient Health Questionnaire (PHQ-9): a meta-analysis. *CMAJ* 2012; 184:E191-6.
24. Adewuya AO, Ola BA, Afolabi OO. Validity of the patient health questionnaire (PHQ-9) as a screening tool for depression amongst Nigerian university students. *J Affect Disord* 2006; 96:89-93.
25. Gjerdingen D, Crow S, McGovern P, Miner M, Center B. Postpartum depression screening at well-child visits: validity of a 2-question screen and the PHQ-9. *Ann Fam Med* 2009; 7:63-70.
26. Spijker J, de Graaf R, Bijl RV, Beekman AT, Ormel J, Nolen WA. Duration of major depressive episodes in the general population: results from The Netherlands Mental Health Survey and Incidence Study (NEMESIS). *Br J Psychiatry* 2002; 181:208-13.
27. Zuithoff NP, Vergouwe Y, King M, Nazareth I, van Wezep MJ, Moons KG, et al. The Patient Health Questionnaire-9 for detection of major depressive disorder in primary care: consequences of current thresholds in a cross-sectional study. *BMC Fam Pract* 2010; 11:98.
28. Santos AMCC. Articular saúde mental e relações de gênero: dar voz aos sujeitos silenciados. *Ciênc Saúde Coletiva* 2009; 14:1177-82.
29. Sackett DL. *Clinical epidemiology: a basic science for clinical medicine*. 2nd Ed. Boston: Little Brown; 1991.
30. Gordis L. *Epidemiology*. 4th Ed. Philadelphia: Elsevier/Saunders; 2009.
31. World Health Organization. *The World Health Report 2001. Mental health: new understanding, new hope*. Geneva: World Health Organization; 2001.

Recebido em 01/Out/2012

Versão final reapresentada em 11/Mar/2013

Aprovado em 18/Mar/2013