Confiabilidade de instrumentos diagnósticos: estudo do inventário de sintomas psiquiátricos do DSM-III aplicado em amostra populacional

Reliability of diagnostic instruments: investigating the psychiatric DSM-III checklist applied to community samples

Sérgio Baxter Andreoli [1]
Sergio Luis Blay [1]
Naomar de Almeida Filho [2]
Jair de Jesus Mari [1]
Cláudio Torres de Miranda [1]
Evandro da Silva Freire Coutinho [3]
Josimar França [4]
Ellis D'Arrigo Busnello [5]

[1] *Núcleo de Estatística e Metodologia Aplicadas, Departamento de Psiquiatria, Escola Paulista de Medicina, Universidade Federal de São Paulo. Rua Dr. Bacelar 334, São Paulo, SP 04026-001, Brasil. andreoli@psiquiatria.epm.br*
[2] *Instituto de Saúde Coletiva, Universidade Federal da Bahia. Rua Padre Feijó 29, 4º andar, Salvador, BA 40110-170, Brasil.*
[3] *Departamento de Epidemiologia e Métodos Quantitativos em Saúde, Escola Nacional de Saúde Pública, Fundação Oswaldo Cruz. Rua Leopoldo Bulhões 1480, Rio de Janeiro, RJ 21041-210, Brasil.*
[4] *Departamento de Clínica Médica, Universidade de Brasília. Campus Universitário Darcy Ribeiro, Asa Norte, Brasília, DF 70910-090, Brasil.*
[5] *Faculdade de Medicina, Universidade Federal do Rio Grande do Sul. Rua Ramiro Barcelos 2600, Porto Alegre, RS 90035-003, Brasil.*

Abstract *This study focused on the reliability of the DSM-III inventory of psychiatric symptoms in representative general population samples in three Brazilian cities. Reliability was assessed through two different designs: inter-rater reliability and internal consistency. Diagnosis of lifetime (k = 0.46) and same-year generalized anxiety (k = 1.00), lifetime depression (k = 0.77), and lifetime alcohol abuse and dependence (k = 1.00) was consistently reliable in the two methods. Lifetime diagnosis of agoraphobia (k = 1.00), simple phobia (k = 0.77), non-schizophrenic psychosis (k = 1.00), and psychological factors affecting physical health (1.00) showed excellent reliability as measured by the kappa coefficient. The main reliability problem in general population studies is the low prevalence of certain diagnoses, resulting in small variability in positive answers and hindering kappa estimation. Therefore it was only possible to examine 11 of 39 diagnoses in the inventory. We recommend test and re-test methods and a short time interval between interviews to decrease the errors due to such variations.*

Key words *Psychiatric; Mental Disorders; Psychiatric Status Rating Scales*

Resumo *O objetivo foi estudar a confiabilidade do inventário de sintomas psiquiátricos do DSM-III aplicado em amostras representativas da população geral de três cidades brasileiras. Foram utilizados os métodos do entrevistador-observador e de consistência interna para medir a confiabilidade. Os diagnósticos de ansiedade generalizada, na vida (k = 0,46), no ano (k = 1,00), depressão na vida (k = 0,77) e os diagnósticos de abuso e dependência de álcool na vida (k = 1,00) foram confiáveis de forma consistente nos dois métodos empregados. Os diagnósticos de agorafobia (k = 1,00), fobia simples (k = 0,77), transtorno psicótico não esquizofrênico (k = 1,00) e o de fatores psicológicos que afetam o físico (1,00), todos feitos para a vida, apresentaram confiabilidade excelente medidos por meio do Kappa. O principal problema de medir a confiabilidade em estudos populacionais é a baixa prevalência de alguns diagnósticos que resulta em uma pequena variabilidade nas respostas positivas, o que impossibilita o cálculo do Kappa. Por causa disso, apenas 11 dos 39 diagnósticos que compõem o inventário puderam ser examinados. Recomenda-se a utilização do método de teste e re-teste com um tempo curto entre as entrevistas para diminuir esse problema.*

Palavras-chave *Psiquiatria; Transtornos Mentais; Escalas de Graduação Psiquiátrica*

Introdução

A confiabilidade de instrumentos diagnósticos no campo da psiquiatria tem sido muito estudada, entretanto as peculiaridades do desempenho destes, aplicados em estudos epidemiológicos com amostragem populacional, deixaram de ser alvo de pesquisas nos últimos anos. Poucos têm se preocupado em incluir um estudo de confiabilidade durante o desenvolvimento da coleta de campo, o que, a nosso ver, pode representar erros de interpretação, sobretudo quando se trata de diagnóstico psiquiátrico. Objetivou-se, portanto, discutir essas peculiaridades apresentando as principais dificuldades e as possíveis soluções para os problemas freqüentemente encontrados. Para tanto, utilizou-se o estudo de confiabilidade do inventário de sintomas do *Diagnostic and Statistical Manual of Mental Disorders* – DSM-III – desenhado para o Estudo Multicêntrico de Morbidade Psiquiátrica do Adulto (Almeida Filho et al., 1997) realizado em três áreas urbanas brasileiras.

O Estudo Multicêntrico de Morbidade Psiquiátrica é o primeiro estudo epidemiológico de transtornos psiquiátricos realizado no Brasil que utiliza metodologia de identificação de caso em duas etapas, com instrumentos padronizados. O objetivo deste estudo foi estimar a prevalência de morbidade psiquiátrica em amostras representativas da população acima de 15 anos residentes em três áreas metropolitanas brasileiras (Brasília, São Paulo e Porto Alegre). Para a identificação de casos psiquiátricos, foram utilizadas duas escalas psiquiátricas aplicadas em duas etapas. Na primeira etapa, o Questionário de Morbidade Psiquiátrica do Adulto (QMPA) foi aplicado para rastreamento de casos psiquiátricos em 6.740 indivíduos; na segunda fase, 30% de prováveis casos e 10% de prováveis não-casos (775 indivíduos), identificados pelo QMPA, foram selecionados e submetidos ao inventário de sintomas do DSM-III.

Método

Estudo de confiabilidade

O estudo de confiabilidade do inventário de sintomas do DSM-III foi abordado por dois métodos: o método de entrevistador-observador, conduzido em São Paulo e o método de consistência interna, que possibilitou o estudo do inventário, aplicado nas cidades de Brasília, São Paulo e Porto Alegre.

Inventário de sintomas do DSM-III

O inventário de sintomas do DSM-III foi desenvolvido no Departamento de Psiquiatria da Universidade de Washington, Saint Louis, Estados Unidos, para verificar os diagnósticos gerados pelo *Diagnostic Interview Schedule* (DIS) (Robins et al., 1981). Esse inventário investiga 39 diagnósticos do DSM-III.

O inventário pode funcionar como um padrão válido para a avaliação de sintomas psiquiátricos e, conseqüentemente, identificação de caso psiquiátrico, desde que aplicado por clínicos treinados e familiarizados com os critérios do DSM-III (Helzer et al., 1985). Todos os diagnósticos que aparecem no instrumento são definidos e têm seus critérios estabelecidos no DSM-III. O instrumento apresenta, para alguns dos diagnósticos, uma lista de sintomas e, para ser classificado como positivo em um ou mais de um diagnóstico, o indivíduo deve apresentar um número mínimo de sintomas e preencher critérios definidos no DSM-III, tais como: pertencer a uma dada faixa de idade determinada, ter apresentado tais sintomas durante um período de tempo determinado, não apresentar outro diagnóstico que exclua a sua presença ou qualquer outra condição de exclusão.

Para a codificação dos sintomas são utilizados três níveis: (1) o sintoma está ausente, ou não há evidência clínica relevante; (5) o sintoma está presente e é relevante do ponto de vista clínico psiquiátrico; (9) incerteza na distinção entre presença e ausência do sintoma. Para a codificação diagnóstica são utilizados quatro níveis: (1) ausente; (3) preenche os critérios do DSM-III e não é excluído por outro diagnóstico do DSM-III; (5) preenche os critérios do DSM-III somente se as regras de exclusão forem ignoradas; (9) incerto. Todos os diagnósticos são feitos tomando como base o tempo de vida do entrevistado; assim, para os indivíduos que preenchem os critérios para um diagnóstico, o período de aparecimento deste é codificado em: (1) nas últimas duas semanas; (2) de duas a menos de um mês; (3) de 1 mês a 6 meses; (4) de 6 meses a 1 ano; (5) menos de 1 ano (não sabe quando); (6) mais de 1 ano atrás.

O inventário de sintomas do DSM-III está organizado para cada diagnóstico de formas diferentes. Existem diagnósticos sem lista e com lista de sintomas e critérios, e estes últimos podem apresentar a lista de sintomas hierarquizada ou não. A hierarquização, nesse caso, significa que um ou mais sintomas devem ser preenchidos de modo afirmativo para que um grupo subseqüente de sintomas seja investigado. Quando não hierarquizado, o entre-

vistador deve investigar todos os sintomas do diagnóstico. No caso dos diagnósticos sem lista de sintomas e critérios, o entrevistador deve indicar o código diagnóstico de acordo com os critérios da classificação diagnóstica do DSM-III.

Método de entrevistador-observador

O método no qual dois entrevistadores, um entrevistador e outro observador, utilizando o inventário de sintomas do DSM-III, avaliaram um mesmo indivíduo simultaneamente e de forma independente foi utilizado em sujeitos (n = 21) selecionados de forma aleatória dentro da uma subamostra da Cidade de São Paulo. Nessa fase da análise, foi avaliada a concordância alcançada pelos entrevistadores com relação à formulação diagnóstica final. Como se trata de variáveis com níveis de medidas categorizados (três níveis no caso dos sintomas e quatro no caso da codificação diagnóstica), o coeficiente Kappa (Almeida Filho, 1989) foi escolhido para a medida da confiabilidade.

O coeficiente Kappa pode variar de 1 a -1, indicando concordância ou discordância completa, e o valor 0 indica o acaso. Para uma interpretação dos valores de Kappa, utilizou-se a caracterização em faixas de valores para os graus de concordância feita por Landis & Koch (1977). Esses autores sugerem que os valores acima de 0,75 representam concordância excelente, valores abaixo de 0,40 uma concordância pobre e os valores entre 0,40 e 0,75 representariam concordâncias de suficiente a boa. O cálculo do coeficiente Kappa e os respectivos intervalos de confiança a 95% foram calculados segundo às fórmulas apresentadas por Bartko & Carpenter (1976).

Método da consistência interna

Um estudo como esse, que envolve grupos de entrevistadores diferentes em cada cidade, que reúne informações de universos sócio-culturais diversos, poderia suscitar a seguinte pergunta: é possível que um estudo de confiabilidade realizado em uma das cidades (São Paulo) garanta a confiabilidade das entrevistas das outras cidades? A propósito dessa questão, Mitchell (1979) afirma que a qualidade dos dados coletados durante um estudo pode não ser a mesma qualidade dos dados coletados durante o estudo de confiabilidade ou treinamento; então o pesquisador está obrigado a mostrar que seu instrumento de medida é confiável, ou seja, que existe pouco erro de medida e que as medidas individuais mostram estabilidade, consistência e dependência do aspecto, característica ou comportamento estudado.

Assim, para responder à pergunta sobre a confiabilidade, não só em São Paulo, mas também em Brasília e Porto Alegre, calculou-se o coeficiente alfa de Cronbach (α) do conjunto de critérios de cada diagnóstico do inventário de sintomas do DSM-III em cada uma das três cidades estudadas. Os a foram calculados apenas para os diagnósticos que puderam ser analisados com o coeficiente Kappa e que tinham uma lista de sintomas ou critérios.

O coeficiente alfa de Cronbach (1951) foi desenvolvido para calcular a confiabilidade de um teste naquelas situações em que o pesquisador não tem a oportunidade de fazer outra entrevista com o indivíduo; contudo, precisa obter uma estimativa apropriada da magnitude do erro da medida. Nessas situações de pesquisa, também pode ser usado o método de partir ao meio (*Split-half method*), no qual os escores de duas subdivisões do instrumento são comparados para determinar sua confiabilidade. Esse método, entretanto, tem sido criticado por confundir erro randômico dos sujeitos com diferenças entre as subdivisões do instrumento (Mitchell, 1979).

A interpretação do alfa de Cronbach, todavia, está relacionada à interpretação que é dada para as estimativas de confiabilidade baseadas no método *split-half*. Isso porque o alfa é uma média de todos os coeficientes *split-half* para um dado instrumento (Carmines & Zeller, 1979; Cronbach, 1951). Em geral, escalas com valor do alfa menor do que 0,70 são evitadas, por outro lado, o valor de α aumenta com o número de questões da escala; assim, escalas com vinte questões freqüentemente apresentam valores de α próximo de 0,90 (Streiner, 1993). Valores de α altos, no entanto, são necessários, mas não suficientes, uma vez que é uma estimativa "otimista" da confiabilidade (Streiner, 1993).

Procedimentos do estudo de campo

O estudo empregou um total de 25 psiquiatras e psicólogos com treinamento clínico para a condução da fase de confirmação diagnóstica. Os profissionais foram treinados para aplicar o inventário de sintomas do DSM-III em cursos sobre os critérios diagnósticos do DSM-III, vídeo-teipes de entrevistas psiquiátricas e entrevistas supervisionadas.

Com um intervalo que variou entre 1 e 4 semanas da primeira etapa do estudo, uma subamostra de indivíduos foi entrevistada por meio do inventário de sintomas do DSM-III. Os re-

sultados do exame de confirmação diagnóstica eram registrados de acordo com o nível de certeza da presença de patologia, duração do transtorno e grau de gravidade.

Em São Paulo, foram coletados dados para um estudo de confiabilidade diagnóstica entre oito entrevistadores treinados na aplicação do inventário de sintomas do DSM-III. Participaram desse estudo duplas de entrevistadores, um psiquiatra e um psicólogo, que entrevistaram simultaneamente um mesmo indivíduo. Apenas um examinador conduzia a entrevista; porém, ambos preenchiam o inventário de sintomas do DSM-III sem se consultarem. Os sujeitos que participaram do estudo de confiabilidade foram escolhidos de forma aleatória no decorrer do estudo por meio de sorteio dos indivíduos previstos para serem entrevistados na semana. Os entrevistadores revezaram as duplas, mantendo sempre um psiquiatra e um psicólogo. Para este estudo de confiabilidade foram previstas cinqüenta entrevistas, mas apenas 21 foram realizadas por problemas operacionais. O principal deles foi a da não-coincidência de horários entre os dois entrevistadores e o entrevistado.

Procedimentos de digitação crítica dos dados

À medida que os inventários de sintomas eram preenchidos, um supervisor de campo discutia os critérios diagnósticos com o entrevistador e os códigos eram anotados. Posteriormente, os inventários foram digitados e os critérios foram novamente conferidos e ajustados para as regras estritas do instrumento. No processo de digitação crítica dos dados, algumas inconsistências foram encontradas, assim, por exemplo, no diagnóstico de ansiedade generalizada houve oito casos cujos critérios anotados apontavam para um código diagnóstico, no entanto, outro código foi anotado; no diagnóstico de depressão foram vinte casos na mesma situação e nos diagnósticos de abuso e dependência de álcool foram 19 casos.

Resultados

Método do entrevistador-observador

No estudo de confiabilidade do inventário de sintomas do DSM-III, cujo método de estudo foi o do entrevistador-observador, dos 39 diagnósticos investigados apenas 11 apresentaram condições para o cálculo do coeficiente de confiabilidade Kappa (Tabela 1). Os motivos pelos quais não foi possível o cálculo de Kappa para 28 diagnósticos foi a ausência de resposta positiva. Devido às características da fórmula para o cálculo de Kappa só é possível o cálculo para aqueles diagnósticos que apresentam variabilidade de resposta positiva. Além disso, é preciso pelo menos uma resposta positiva concordante, por isso, mesmo entre os 11 diagnósticos estudados, ficaram sem estimativa de confiabilidade os diagnósticos de agorafobia, depressão, transtorno distímico, transtorno psicótico não esquizofrênico e fatores psicológicos que afetam o físico, todos feitos para o último ano.

A maior parte desses 11 diagnósticos, feitos para a vida, apresenta coeficiente Kappa acima de 0,75 (Tabela 1). Os diagnósticos que apresentam concordância excelente são os de agorafobia ($k = 1,00$), transtorno distímico ($k = 1,00$), outros transtornos psicóticos ($k = 1,00$), abuso e dependência de álcool ($k = 1,00$) e o diagnóstico de dependência de tabaco ($k = 0,80$; $p < 0,01$). O diagnóstico de ansiedade generalizada ($k = 0,46$; $p < 0,05$) apresenta uma concordância suficiente e o diagnóstico de outros transtornos de ansiedade ($k = 0,34$; $p < 0,05$) apresenta concordância pobre.

Para os diagnósticos que foram feitos em relação ao último ano, foi possível calcular o coeficiente de confiabilidade Kappa de seis diagnósticos (Tabela 1). Destes, o diagnóstico de ansiedade generalizada ($k = 1,00$) apresenta concordância excelente, o diagnóstico de fobia simples ($k = 0,64$) apresenta concordância boa, os diagnósticos de outros transtornos de ansiedade ($k = 0,34$), abuso de álcool ($k = 0,35$) e dependência de álcool ($k = 0,35$) apresentam concordância pobre. O diagnóstico de dependência de tabaco ($k = 0,31$) apresenta concordância pobre não significante, ou seja, não podemos afirmar que é diferente do acaso. Os diagnósticos de depressão e transtornos psicóticos não esquizofrênicos, que apresentam concordância excelente quando feitos para a vida, não apresentam variabilidade de resposta positiva quando feitos para o último ano, não permitindo o cálculo do Kappa; entretanto, apresentam uma proporção de concordância igual a 100%.

Método de consistência interna

Os alfas de Cronbach foram calculados para os diagnósticos de dependência de tabaco, ansiedade generalizada, depressão, transtorno distímico, abuso de álcool e dependência de álcool nos dados dos indivíduos que tiveram todos os critérios para o respectivo diagnóstico preenchidos.

Os critérios diagnósticos para a maioria dos diagnósticos estudados apresentam coeficientes com valores acima de 0,70 (Tabela 2), o que representa uma excelente confiabilidade, exceto para o diagnóstico de transtorno distímico que apresenta valores menores do que 0,42.

Os valores dos coeficientes mantêm-se estáveis entre as cidades para a maioria dos diagnósticos (Tabela 2) com diferenças nos valores inferiores a 0,03. Os diagnósticos de dependência de tabaco e transtorno distímico são as exceções. Os critérios diagnósticos de dependência de tabaco apresentaram uma diferença de 0,09 entre os valores encontrados nas cidades de Brasília e Porto Alegre ($\alpha = 0,79$) e o valor encontrado na Cidade de São Paulo ($\alpha = 0,87$). Os critérios diagnósticos de transtorno distímico apresentam as maiores diferenças; a menor diferença entre os valores de a está entre as cidade de Brasília e Porto Alegre ($\alpha = 0,16$) e a maior diferença está entre as cidades de São Paulo e Porto Alegre ($\alpha = 0,30$).

Comparação entre resultados dos dois métodos

O sumário dos resultados das medidas de confiabilidade dos dois métodos empregados (Tabela 3) mostra os diagnósticos de ansiedade generalizada, na vida, no ano, depressão na vi-

Tabela 1

Concordância positiva, negativa e Kappa da aplicação do inventário de sintomas do DSM-III feita por dois entrevistadores em entrevistas simultâneas e com avaliações independentes, realizadas em 21 indivíduos na Cidade de São Paulo.

São Paulo	Diagnóstico na vida				Diagnóstico no último ano			
	Concordância +	Concordância -	Kappa	IC 95%	Concordância +	Concordância -	Kappa	IC 95%
Dependência de tabaco	6	13	0,80	0,41-1,00	2	14	0,31	-0,06-0,68
Ansiedade generalizada	1	18	0,46	0,12-0,83	1	20	1,00	–
Outros transtornos de ansiedade	1	17	0,34	0,03-0,65	1	17	0,34	0,03-0,65
Agorafobia	1	20	1,00	–	0	20	–	–
Fobia simples	2	18	0,77	0,36-1,00	1	19	0,64	0,25-1,00
Depressão	2	18	0,77	0,36-1,00	0	21	–	–
Transtorno distímico	1	20	1,00	–	0	20	–	–
Outros transtornos psicóticos	1	20	1,00	–	0	21	–	
Abuso de álcool	4	17	1,00	–	1	17	0,35	0,04-0,66
Dependência de álcool	4	17	1,00	–	1	17	0,35	0,04-0,66
Fatores psicológicos que afetam o físico	2	18	0,77	0,36-1,00	0	20	–	–

Tabela 2

Consistência interna, medida pelo Alfa de Cronbach, dos critérios diagnósticos do inventário de sintomas do DSM-III, aplicado por entrevistador nas amostras de Brasília, São Paulo e Porto Alegre.

Diagnósticos do DSM-III	Brasília		São Paulo		Porto Alegre	
	α	n	α	n	α	n
Dependência de tabaco	0,79	80	0,87	71	0,79	97
Ansiedade generalizada	0,94	157	0,93	194	0,93	241
Depressão	0,95	233	0,95	176	0,92	246
Distúrbio distímico	0,26	4	0,12	9	0,42	20
Abuso e dependência de álcool	0,90	234	0,90	219	0,89	263

O número ao lado de cada valor de Alfa de Cronbach é o de indivíduos que tiveram todos os critérios diagnósticos respondidos para o respectivo diagnóstico e que serviram de base de dados para a análise de confiabilidade.

Tabela 3

Sumário dos resultados do estudo de confiabilidade do inventário de sintomas do DSM-III aplicados em amostras populacionais. Brasília, São Paulo e Porto Alegre, Brasil, 1990.

Diagnóstico	**Kappa**			**Consistência interna** (α)			
	Excelente	Suficiente	Pobre	Excelente	Pobre	Estável	Instável
Dependência de tabaco							
Vida	X			X			X
Ano			X	X			X
Ansiedade generalizada							
Vida		X		X		X	
Ano	X			X		X	
Outros transtornos de ansiedade							
Vida			X	*	*	*	*
Ano			X	*		*	
Agorafobia							
Vida	X			*		*	
Ano	*	*	*	*	*	*	*
Fobia simples							
Vida	X			*		*	
Ano		X		*		*	
Depressão							
Vida	X			X		X	
Ano	*	*	*	X		X	
Transtorno distímico							
Vida	x				X		X
Ano	*	*	*		X		X
Transtorno Psicóticos não esquizofrênicos							
Vida	X			*	*	*	
Ano	*	*	*	*	*	*	*
Abuso e dependência de álcool							
Vida	X			X		X	
Ano			X	X		X	
Fatores psicológicos que afetam o físico							
Vida	X			*	*	*	*
Ano	*	*	*	*	*	*	*

X = nível de confiança; * = situações na qual o método não pode ser utilizado para medir a confiabilidade do diagnóstico.

da e os diagnósticos de abuso e dependência de álcool na vida como os mais confiáveis, porque apresentam todos os indicadores apontando para isso. Os diagnósticos de agorafobia, fobia simples, fobia social, transtorno psicótico não esquizofrênico e o de fatores psicológicos que afetam o físico, todos feitos para a vida, apresentaram confiabilidade excelente, medidos por meio do Kappa, mas não puderam ser verificados por meio do método de consistência interna. Para esses diagnósticos podemos afirmar sobre a confiabilidade na Cidade de São Paulo, mas não sobre as outras cidades.

Discussão

Nos estudos de confiabilidade dos instrumentos diagnósticos existe uma variedade de fontes de erro que pode resultar em baixa confiabilidade. Segundo Grove et al. (1981), esses erros podem ser em função de problemas esta-

tísticos, com o desenho e/ou na execução do estudo. Neste estudo, foram tomados alguns cuidados para controle de erros e cuidados na escolha dos coeficientes para medir a confiabilidade do inventário de sintomas de DSM-III.

Problemas estatísticos

Confiabilidade em termos psicométricos é a reprodução das distinções feitas entre aspectos das pessoas (Bartko, 1991). A confiabilidade, definida como o grau com que múltiplas medidas de um sujeito concordam, é, em termos conceitual e computacional, uma função de duas variações: a variação entre sujeitos e a intra-sujeitos (Bartko, 1991).

Para medir a confiabilidade de um instrumento, é importante que a característica a ser medida, nesse caso, os sintomas, critérios diagnósticos e diagnósticos, varie (variação entre sujeitos). É importante notar que a mera replicação sem discriminação não é o bastante (Shrout et al., 1987). Por exemplo: o diagnóstico de esquizofrenia não foi encontrado na subamostra do estudo de confiabilidade, nesse caso, apesar de os entrevistadores terem sido perfeitamente concordantes ao codificarem esse diagnóstico como negativo, não podemos afirmar que a confiabilidade do instrumento para o diagnóstico de esquizofrenia foi averiguada, isto porque, no estudo em questão, não houve variação entre os sujeitos examinados quanto ao referido diagnóstico.

O segundo componente da variação expressa o erro ou a variação intra-sujeitos. Essa variação expressa o grau com que os entrevistadores concordaram ao avaliar. Por exemplo: no diagnóstico de dependência de álcool na vida a variação intra-sujeitos foi zero, ou seja, a confiabilidade foi um, que quer dizer perfeita.

A melhor situação para um estudo de confiabilidade ocorre quando: (1) os entrevistadores são apresentados para um número grande e variado de características (variação entre sujeitos diferente de zero; em termos psicométricos, isto é conhecido como a variação do escore positivo) e (2) quando a maioria dos entrevistadores concordam entre eles para cada uma das diferentes características (baixa variação intra-sujeitos) (Bartko, 1991).

Em estudos populacionais, a tarefa de garantir a variação do escore positivo se torna complicada, uma vez que as amostras geralmente são homogêneas, ou seja, como a prevalência é baixa na população geral para a maioria das doenças, uma amostra randômica dessa população resulta em uma maioria de indivíduos sem doença e uma minoria doente. De fato, essa situação se apresenta neste estudo por dois motivos. O primeiro pela amostra selecionada ser pequena, apenas 21 indivíduos quando estavam sendo investigados 39 diagnósticos, e o segundo é que, na grande maioria, os diagnósticos estudados são raros na população geral. Por estes motivos, pudemos estudar a confiabilidade de apenas 11 dos 39 diagnósticos que compunham o inventário de sintomas do DSM-III, isso porque os outros diagnósticos não foram encontrados nas amostras.

Na escolha da medida estatística, o Kappa é a melhor medida de confiabilidade para dados categóricos (Bartko, 1991; Shrout et al., 1987), entretanto muito se tem discutido a propósito da dependência entre o Kappa e as taxas de prevalência. Essa dependência fica bem ilustrada quando comparamos os valores de Kappa que foram encontrados na dependência de tabaco na vida ($k = 0,80$) e aquele encontrado no diagnóstico de ansiedade generalizada na vida ($k = 0,46$); apesar de a proporção de concordância observada ser exatamente a mesma ($Po = 0,90$), a distribuição desigual na proporção de concordância de respostas positivas observadas ($Po+ = 0,28$ e $Po+ = 0,05$) determinou uma variação muito grande do Kappa. Esse comportamento do Kappa muitas vezes dificulta a interpretação dos resultados.

Devido à dependência da prevalência e a conseqüente dificuldade de interpretação do Kappa, algumas propostas de interpretação e de medidas de confiabilidade alternativas têm sido discutidas na literatura. Logo, quem primeiro nomeou o problema foram Carey & Gottesman (1978) que concluem que o cálculo da confiabilidade para uma determinada prevalência não pode ser comparado com o cálculo para uma outra prevalência muito diferente, nessas circunstâncias, a interpretação das medidas de confiabilidade passaria por um consenso ou um relaxamento de critérios. Grove et al. (1981) afirmam que a generalização da confiabilidade para uma outra taxa de prevalência diferente daquela observada num determinado estudo pode ou não pode ser válida e que a interpretação das medidas de confiabilidade deveria levar em conta fatores como: o desenho, taxa de prevalência, sensibilidade e especificidade, se possível, e levar em conta a própria escolha do coeficiente utilizado para essa medida.

Até aqui o problema da taxa de prevalência nos estudos de confiabilidade vinha sendo colocado como um problema basicamente de procedimentos; Spitznagel & Helzer (1985) recolocam o problema em termos estatísticos, ou seja, o assim nomeado “problema da taxa de

prevalência nos estudos de confiabilidade" era na verdade um problema do coeficiente Kappa e sugerem a medida de associação Yule's, que teria a propriedade de ser independente da prevalência e, portanto, mais adequada para estudos populacionais.

Shrout et al. (1987) em seu artigo intitulado *Quantification of Agreement in Psychiatric Diagnosis Revisited* afirmam que, contrário aos argumentos de Spitznagel & Helzer (1985), não existe um problema com o coeficiente Kappa e que considerando qualquer taxa de prevalência, o máximo valor de Kappa é sempre 1,0; indicando concordância perfeita. A argumentação de Shrout et al. (1987) se baseia na baixa possibilidade de interpretação do Yule's; essa baixa possibilidade de interpretação seria decorrente do Yule's ser função da raiz quadrada da *odds ratio*, o que impediria uma transformação linear que, por sua vez, impediria uma apelação intuitiva. Segundo estes autores, o Yule's poderia ser interpretado como um verdadeiro coeficiente de confiabilidade quando: (1) cada clínico tivesse a mesma sensibilidade e especificidade relativa ao critério diagnóstico, (2) a sensibilidade fosse igual à especificidade e (3) a taxa de prevalência fosse igual a 50%. Os autores chamam a atenção para o fato de que essas condições por si seriam uma restrição bastante importante para sua aplicabilidade.

A despeito de todas essas considerações, Shrout et al. (1987) afirmam que a consagração do Kappa na comunidade científica, ao longo de 18 anos, provou ser esse de extrema utilidade e versatilidade no teste e no desenvolvimento dos procedimentos e critérios diagnósticos e que uma mudança no uso padronizado da estatística de confiabilidade poderia resultar em confusão na literatura científica.

O coeficiente de consistência interna a de Cronbach também se mostrou útil para o estudo da confiabilidade sobretudo para se ter acesso às medidas nas cidades onde não foi desenhado um estudo específico para medir a confiabilidade, entretanto, só pôde ser utilizado nos diagnósticos que continham listas de sintomas e critérios. Ao introduzir tal coeficiente no estudo, o número de diagnósticos passíveis de serem estudados ficou restrito, 5 dos 39 iniciais.

Cuidados com desenho e execução

As fontes de erro podem ocorrer durante a etapa de entrevista (variação de informação): o entrevistado pode dar informação incorreta por falta de compreensão, falta de concentração, resistência intencional, ou o entrevistador pode errar ao registrar os dados. Outras possíveis fontes de erro seriam a instabilidade do fenômeno clínico (variação ocasional) e a variação dos critérios diagnósticos empregados por cada entrevistador (variação de critério). Finalmente, os erros de medida podem ser o resultado da falta de cuidado, da inconsistência ou incompetência dos entrevistadores.

Dentro do desenho desse estudo, as únicas fontes de erro passíveis de controle foram as que determinam as variações ocasionais e de critério. A variação ocasional foi controlada já que os sujeitos foram entrevistados uma única vez; entretanto, em estudos populacionais isso pode se constituir em um problema de respostas positivas nos diferentes diagnósticos, como os encontrados neste estudo. Em outras palavras, um desenho do tipo entrevistador-observador necessita de dois entrevistadores ao mesmo tempo, então para realizá-lo, garantindo um número de indivíduos com os diagnósticos (respostas positivas), teríamos que trabalhar a maior parte do tempo com dois entrevistadores, o que tornaria o estudo inviável. Parece que o método de teste e re-teste seria o mais adequado para estudos populacionais, visto que seria possível investigar a confiabilidade para todos os diagnósticos, à medida que fossem aparecendo. Embora esse método apresente problemas de erro devido à variação ocasional (mudança de sintomas entre as entrevistas), isso poderia ser controlado com a diminuição de tempo entre as entrevistas.

Outro motivo para a recomendação do método teste/re-teste é a possível fonte de erro devido à variação de informação introduzida ao adotarmos um instrumento composto de questões hierarquizadas. Desse modo, dentro de um desenho entrevistador-observador, o primeiro pode induzir ao segundo respostas às questões subseqüentes sem que este último tivesse concordado *a priori* com a questão que determinou o prosseguimento. Mais grave ainda é a situação na qual o entrevistador salta as questões subseqüentes sem que o observador concorde com a falsidade da questão que determinou o pulo. Neste último caso, as questões ficam irremediavelmente perdidas, pois o observador não pode retomá-las.

Quanto à variação, devido aos critérios adotados para a codificação diagnóstica, alguns cuidados foram tomados antes, durante e depois da coleta dos dados. Antes do início da coleta de dados, os entrevistadores foram treinados por intermédio de aulas sobre os critérios diagnósticos do DSM-III; treinados para a aplicação do inventário de sintomas do DSM-III, e os dados foram conferidos por um supervisor

durante a aplicação do inventário no campo. Depois da coleta, os inventários de sintomas foram digitados integralmente (sintomas, critérios diagnósticos e diagnósticos) de forma crítica, ou seja, foram conferidos todos os critérios, o que permitiu uma padronização dos critérios nas três cidades estudadas.

O processo de digitação crítica do inventário do DSM-III revelou um aspecto importante na identificação de caso adotado nesse estudo, qual seja, os entrevistadores nem sempre seguiram os critérios adotados para o estudo, houve algumas mudanças entre as anotações feitas para os sintomas e o que foi efetivamente anotado como código diagnóstico. Os critérios adotados para a elaboração diagnóstica (critérios do DSM-III) não foram seguidos em alguns casos, isso porque, entre o que foi anotado como sintomas e o que foi anotado como diagnóstico existiu um raciocínio clínico, decorrente provavelmente da interação entre o entrevistador e o entrevistado, ao qual não é possível ter acesso e nem controle dentro de um estudo epidemiológico em larga escala. Na tentativa de controlar esse importante viés e manter uniformes os critérios adotados no estudo, ajustaram-se os códigos diagnósticos desses casos aos critérios estritos do instrumento.

O procedimento de ajuste dos códigos diagnósticos aos critérios do DSM-III, se por um lado tranqüiliza, por garantir uma padronização dos resultados, por outro lado inquieta, por deixar a questão da validade em aberto porque deixa de lado o raciocínio clínico. Esse estudo, entretanto, não tem como objetivo questionar a validade dos critérios do DSM-III, mas toma-os como válidos com a finalidade de delimitar um objeto de estudo.

Confiabilidade do inventário de sintomas do DSM-III

Os resultados do estudo de confiabilidade do inventário de sintomas do DSM III no estudo multicêntrico brasileiro mostram, de um modo geral, que o instrumento foi adequado dentro dos seus propósitos no estudo. Marcadamente para os diagnósticos de ansiedade generalizada, na vida, no ano, depressão na vida e os diagnósticos de abuso e dependência de álcool na vida porque se mostram confiáveis de forma consistente. Este resultado é importante, já que estes foram os diagnósticos encontrados como os mais prevalentes na população brasileira (Almeida Filho et al., 1997).

Os maiores problemas de confiabilidade encontrados nos 11 diagnósticos estudados foram aqueles decorrentes da formulação diagnóstica para último ano. Os mesmos diagnósticos com confiabilidade excelente ou boa para a formulação feita para a vida tiveram confiabilidade pobre para a formulação no último ano. É importante notar que a única diferença entre um e outro é uma questão que indaga sobre a última vez em que apresentou tais sintomas; mesmo assim os entrevistadores, presentes na mesma entrevista, escutando a mesma resposta, foram discordantes nas suas anotações. É importante que mais atenção seja dada para este tipo de questão.

Conclusão

Nos estudos epidemiológicos em larga escala como este, a confiabilidade do instrumento utilizado para identificação de caso é um aspecto essencial. Desta forma alguns cuidados devem ser tomados. No desenho do estudo, a escolha do método entrevistador-observador garante o controle das fontes de erro ocasionais; no entanto, penaliza o estudo com pouca variabilidade de respostas positivas e introduz uma fonte de erro de informação nos casos de instrumentos hierarquizados. Neste sentido, o método teste/re-teste seria uma alternativa, desde que o tempo entre as entrevistas pudesse ser o menor possível. Os cuidados com as fontes de erro por variação de critérios foram adequados neste estudo. Além das tradicionais que são o treinamento dos entrevistadores, supervisões de campo e o ajuste dos critérios durante o processo de digitação.

O coeficiente de confiabilidade Kappa, como vimos, é influenciado pela prevalência do fenômeno estudado e, além disso, necessita de pelo menos uma resposta positiva na qual os dois entrevistadores concordem para que possa ser calculado. Essa característica do coeficiente torna o estudo da confiabilidade dos instrumentos utilizados nos estudos epidemiológicos muitas vezes inviável. De fato, foi o que aconteceu neste estudo em que, por este motivo, foi possível analisar apenas 11 dos 39 diagnósticos investigados. Além do Kappa, método de consistência interna se mostrou viável e auxiliou na medida da confiabilidade nas cidades onde um estudo específico na foi desenhado; porém, a análise ficou limitada a cinco diagnósticos apenas.

Os diagnósticos de ansiedade generalizada, na vida, no ano, depressão na vida e os diagnósticos de abuso e dependência de álcool na vida foram confiáveis de forma consistente nos dois métodos empregados. Os diagnósticos de

agorafobia, fobia simples, fobia social, transtorno psicótico não esquizofrênico e o de fatores psicológicos que afetam o físico, todos feitos para a vida, apresentaram confiabilidade excelente, medidos por meio do Kappa, mas não puderam ser verificados por meio do método de consistência interna.

Agradecimentos

Estudo financiado pela Fundação de Amparo à Pesquisa do Estado de São Paulo (processo nº 94/0526-0).

Referências

ALMEIDA FILHO, N., 1989. Desenvolvimento de instrumentos na pesquisa epidemiológica. In: *Epidemiologia Sem Números: Uma Introdução Crítica à Ciência Epidemiológica* (N. Almeida Filho, org.), pp. 39-54, Rio de Janeiro: Editora Campus.

ALMEIDA FILHO, N.; MARI, J. J.; COUTINHO, E.; FRANÇA, J. F.; FERNANDES, J. G.; ANDREOLI, S. B. & BUSNELLO, E. D., 1997. Brazilian multicentric study of psychiatric morbidity: Methodological feature and prevalence estimates. *British Journal of Psychiatry*, 171:524-529.

BARTKO, J. J., 1991. Measurement and reliability: Statistical thinking considerations. *Schizophrenia Bulletin*, 17:483-489.

BARTKO, J. J. & CARPENTER Jr., W. T., 1976. On the methods and theory of reliability. *Journal of Nervous and Mental Disease*, 163:307-317.

CAREY, G. & GOTTSMAN, I. I., 1978. Reliability and validity in binary rating. *Archives of General Psychiatry*, 35:1454-1459.

CARMINES, E. G. & ZELLER, R. A., 1979. Assessing reliability. In: *Assessing Reliability: Reliability and Validity Assessment* (J. L. Sullivan, ed.), Quantitative Applications in the Social Science 17, pp. 37-49, Beverly Hills/London: Sage Publications.

CRONBACH, L. J., 1951. Coefficient alpha and the internal structure of test. *Psychometrika*, 16:297-334.

GROVE, W. M.; ANDREASEN, N. C.; McDONALD-SCOTT, P.; KELLER, M. B. & SHAPIRO, R. W., 1981. Reliability studies of psychiatric diagnosis. *Archives of General Psychiatry*, 38:408-413.

HELZER, J. E.; ROBINS, L. N.; McEVOY, L. T.; SPITZNAGEL, E. L.; STOLTZMAN, R. K.; FARMER, A. & BROCKINGTON, I. F., 1985. A comparison of clinical and diagnostic interview schedule diagnoses. Physician reexamination of lay-interviewed cases in the general population. *Archives of General Psychiatry*, 42:657-666.

LANDIS, J. R. & KOCH, G. G., 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33:159-174.

MITCHELL, S. K., 1979. Interobserver agreement, reliability, and generalizability of data collected in observation studies. *Psychological Bulletin*, 86: 376-390.

ROBINS, L. N.; HELZER, J. E.; CROUGHAN, J. & RATCLIFF, K. S., 1981. The NIMH diagnostic interview schedule: Its history, characteristics, and validity. *Archives of General Psychiatry*, 38:381-389.

SHROUT, P. E.; SPITZER, R. L. & FLEISS, J. L., 1987. Quantification of agreement in psychiatric diagnosis revisited. *Archives of General Psychiatry*, 44:172-177.

SPITZNAGEL, E. L. & HELZER, J. E., 1985. A proposed solution to the base rate problem in the kappa statistic. *Archives of General Psychiatry*, 42:725-728.

STREINER, D. L., 1993. A checklist for evaluating the usefulness of rating scales. *Canadian Journal of Psychiatry*, 38:140-148.