Concordância entre avaliadores na seleção de artigos em revisões sistemáticas

Agreement among raters in the selection of articles in a systematic review

Natália Sanchez Oliveira
Julicristie Machado de Oliveira
Denise Pimentel Bergamaschi
Departamento de Epidemiologia da Faculdade de Saúde Pública da Universidade de São Paulo.

Auxílio financeiro: FAPESP – Processo 05/50462-4
Correspondência: Natália Sanchez Oliveira. Departamento de Epidemiologia, Faculdade de Saúde Pública/USP. Av. Dr. Arnaldo, 715 CEP 01246-904 - Cerqueira César, São Paulo, SP. E-mail: nataliasanchez@usp.br

Resumo

O objetivo do estudo é apresentar aspectos metodológicos relativos à concordância entre avaliadores, na seleção inicial de artigos em estudo de revisão sistemática com ou sem metanálise. Foram utilizados como exemplo dados da fase inicial do estudo "Suplementação de vitamina A em lactantes: revisão sistemática". Estes constituem o resultado da leitura, realizada por dois avaliadores, de resumos de artigos selecionados, criteriosamente, em bases bibliográficas eletrônicas. Para cada resumo, foram respondidas as questões: "O estudo envolve mulheres no pós-parto?"; "Trata-se de um estudo com suplementação de vitamina A?"; "O estudo é ensaio clínico?"; seguidas da decisão (inclusão/exclusão) do trabalho. Os dados foram inseridos em planilha Excel, com dupla digitação e uso de procedimento de validação. Utilizou-se o índice de concordância kappa para os aspectos: população, intervenção, tipo de estudo e decisão. Foram identificados 2.553 trabalhos. Os índices de concordância kappa foram, quanto à adequação da população de estudo: k=0,46; do tipo de intervenção: k=0,59; do tipo de estudo: k=0,59 e, sobre a decisão pela inclusão/exclusão: k=0,44. Com base nas concordâncias razoável (tipo de estudo e intervenção) e pobre (população de estudo) observadas entre avaliadores, enfatiza-se a necessidade da leitura inicial dos trabalhos por pelo menos dois avaliadores. As reuniões para consenso realizadas nos casos discordantes foram úteis para resolver diferenças de interpretação entre os avaliadores, possibilitando nova leitura e maior reflexão, e colaboraram, portanto, para diminuir as chances de exclusão de um artigo que, na verdade, deveria ser incluído, com conseqüente maior controle sobre possível vício de seleção.

Palavras-chave: Metanálise. Concordância entre avaliadores. Índice de concordância kappa. Viés de seleção.

Abstract

The objective of this study is to present the methodological aspects of inter-rater agreement in the initial selection of studies for a systematic review with or without meta-analysis. As an example, we used data from the initial phase of the study called "Vitamin A supplementation for breastfeeding mothers: systematic review". The data are the result of a reading carried out by two raters of article abstracts selected judiciously among electronic bibliography databases. For each study we posed the questions: "Does the study involve post-partum females?"; "Is it a vitamin A supplementation study?"; "Is it a clinical trial?", followed by a decision (inclusion/exclusion) concerning the study. The data were keyed twice into an Excel spreadsheet and a validation procedure was followed. The kappa agreement rate was applied to the following aspects: population, intervention, study type, and decision. We identified 2,553 studies. The kappa agreement rates were: k=0.46 for suitability of the population studied; k= 0.59 for intervention type; k=0.59 for study type; and k=0.44 for the inclusion/exclusion decision. Given the fair (intervention and study type) and slight (population studied) agreement between raters, we emphasize the need for the studies to be read initially by at least two raters. The consensus meetings carried out in the presence of disagreement were useful to solve differences of interpretation between raters, provided new understanding and deeper reflection, contributed to reduce the chance of non-inclusion of necessary studies, and thus enhanced control over a possible selection bias.

Keywords: Meta-Analysis. Interrater agreement. Kappa statistics. Selection Bias.

Introdução

Estudos de revisão sistemática devem ser conduzidos segundo metodologia bem definida[1,2] e possuem vantagens sobre as revisões bibliográficas tradicionais, em que os trabalhos são identificados de maneira extensiva e nem sempre objetiva, podendo apresentar vieses de seleção. Estudos de revisão sistemática com metanálise incluem não só a identificação e seleção criteriosa de trabalhos, mas também a combinação estatística dos resultados provenientes dos mesmos, com o intuito de produzir uma única estimativa do efeito de uma intervenção de saúde[1,2].

A estratégia de identificação e a seleção de trabalhos que tratam do assunto de investigação constituem importantes etapas metodológicas em um estudo de revisão sistemática com ou sem metanálise e são, de modo geral, compostas por: identificação de trabalhos, seleção inicial e análise de qualidade dos artigos.

A primeira seleção é realizada no início do estudo; nesta, procede-se à leitura de resumos de trabalhos oriundos de bases de dados bibliográficos, identificados com a utilização de estratégias de busca pré-definidas e testadas, que envolvem palavras-chave referentes ao objeto de estudo. Para a primeira seleção, deve ser conduzida a análise de elegibilidade de artigos que, em um momento posterior, serão avaliados e selecionados segundo critérios de qualidade. A elegibilidade é definida com base em pelo menos três características: a população de estudo, o tipo de intervenção e o tipo de desenho metodológico[2,3].

O objetivo do presente estudo é discorrer sobre aspectos metodológicos referentes à análise de concordância entre avaliadores na seleção inicial de artigos em revisão sistemática com ou sem metanálise, utilizando como exemplo o estudo "Suplementação de vitamina A em lactantes: revisão sistemática"[4].

Metodologia

Por meio de buscas criteriosas nas bases bibliográficas eletrônicas Cochrane Library – Cochrane Controlled Trials Register, Electronic Reference Library (ERL), Web of Science, LILACS e PubMed, e a partir de estratégias com palavras-chave referentes à população de estudo (mulheres no pós-parto e lactantes), tipo de estudo (ensaio aleatorizado) e tipo de intervenção (suplementação com vitamina A) foram identificados artigos científicos que constituíram a amostra inicial (n_1). A estratégia utilizou operadores booleanos, sem restrição de idioma ou tipo de publicação. Informações (título, autores, referências e resumo) sobre os artigos identificados foram armazenadas no programa gerenciador de referências EndNote 7.0[5], para análise de elegibilidade. Decidiu-se considerar o trabalho no estudo de elegibilidade, mesmo que o resumo não estivesse disponível. Nestes casos, as decisões foram tomadas com base nos títulos dos trabalhos e, na impossibilidade da tomada de decisão por falta de clareza, os artigos foram lidos na íntegra, com maior enfoque nos aspectos metodológicos.

Para os n_1 artigos, dois avaliadores preencheram, de forma independente, formulários que continham as questões: "O estudo envolve mulheres no pós-parto?"; "Trata-se de um estudo com suplementação de vitamina A?"; "O estudo é ensaio clínico?", com as alternativas: sim, não e "não claro". Com base nestas, foram tomadas decisões sobre a inclusão/exclusão do trabalho. Os trabalhos incluídos foram considerados elegíveis (n_2) e constituíram a nova seleção da próxima etapa do estudo, a análise de qualidade. Frente a discordâncias entre os avaliadores sobre a elegibilidade, as decisões foram tomadas com base em discussão e consenso. Os dados dos formulários de elegibilidade foram armazenados na planilha Excel[6], com dupla digitação de modo independente, sendo realizado procedimento de validação da mesma.

Para análise estatística foi utilizado o programa Stata 9.0[7], com a construção de tabelas e cálculo da concordância bruta, do índice de concordância kappa[8,9] e respectivos intervalos de confiança de 95% (IC95%), sem ajuste para viés e prevalência, segundo aspectos estudados e decisão final sobre a inclusão/exclusão, utilizando-se dois bancos de dados: artigos com e sem resumo e artigos somente com resumo.

O estudo original "Suplementação de vitamina A em lactantes: revisão sistemática"[4] foi aprovado pelo Comitê de Ética em Pesquisa da Faculdade de Saúde Pública da Universidade de São Paulo.

Resultados

Foram identificados na busca inicial 2.553 artigos, que, em sua grande maioria (2,318; 90,8%), tiveram seus resumos lidos pelos dois avaliadores. Para uma parte dos artigos (241; 9,4%) não foi possível a leitura dos resumos, pois estes não estavam disponíveis. Nestes casos, as decisões foram tomadas com base nos títulos e, nos casos com decisões discordantes (49; 1,9%), os artigos foram lidos na íntegra.

Segundo a análise de elegibilidade, observou-se que o avaliador 1 decidiu pela inclusão de 30 trabalhos e o avaliador 2 pela inclusão de 51, sendo que para 26 (1,0%) trabalhos foi observada decisão concordante sobre a inclusão. Os resultados de não inclusão foram concordantes para 2.416 trabalhos (94,6%). Para 16 trabalhos observou-se discordância sobre a inclusão dos mesmos, sendo que para 4 artigos o avaliador 1 decidiu pelo sim e o 2 pelo não, e para 12 trabalhos as decisões foram invertidas. Os avaliadores decidiram pela categoria "não claro" em números semelhantes de artigos (distribuições marginais); entretanto, as distribuições conjuntas merecem atenção: observa-se concordância para 10 trabalhos. As demais caselas guardam resultados discordantes: 43 trabalhos avaliados como não elegíveis pelo avaliador 1 foram classificados como "não claros" pelo avaliador 2, e 13 traba-

lhos identificados como "não claros" pelo avaliador 1 foram considerados elegíveis pelo avaliador 2. As comparações para os demais aspectos avaliados são também apresentadas na Tabela 1.

A proporção de concordância global observada para a decisão final (inclusão, exclusão e "não claro"), quando utilizado o banco de dados que incluía os artigos com e sem resumos disponíveis, foi de 0,960 (ou 96,0%), e o índice de concordância kappa foi igual a 0,44 (IC95%: 0,33 – 0,56). Os valores destas estatísticas, quando considerado o banco de dados que continha somente artigos com resumos, foram concordância bruta de 97,4% e kappa igual a 0,51 (IC95%: 0,38 – 0,64). Estes valores são detalhados segundo aspectos abordados e apresentados na Tabela 2. Observa-se semelhança nas concordâncias, pela análise dos intervalos de classe, para os aspectos tipo de intervenção e de estudo, segundo valores calculados em ambos os bancos de dados.

Discussão

Vícios de seleção em estudos de revisão sistemática com ou sem metanálise precisam ser controlados, uma vez que a inclusão ou exclusão de um único trabalho tem potencial para alterar a decisão final sobre o efeito da intervenção sob estudo[1,2].

A existência de discordância na seleção de artigos para revisões sistemáticas constitui importante aspecto metodológico evidenciado pelo presente artigo, exemplificado com base em estudo que utilizou dois avaliadores na análise de elegibilidade. Os dados apresentados são particulares, pois a estratégia de busca utilizada foi ampla, fazendo com que muitos trabalhos fossem identificados inicialmente, sem que de fato apresentassem os critérios de elegibilidade. Isto se reflete na concordância bruta de apenas 1% para a inclusão do estudo, e de 95% para não inclusão. A utilização de estratégia de busca ampla pode

Tabela 1 - Distribuição de trabalhos segundo critérios de elegibilidade e respostas dos avaliadores.
Table 1 - *Articles distributed according to eligibility criteria and rater answers.*

	Avaliador 1	Avaliador 2			
		Sim	Não	Não claro	Total
O estudo envolve mulheres no pós-parto?	Sim	85	35	7	127
	Não	32	2145	16	2193
	"Não claro"	17	174	42	233
	Total	134	2354	65	2553
Trata-se de um estudo com suplementação de vitamina A?	Sim	432	79	9	520
	Não	189	1604	45	1838
	"Não claro"	28	109	58	195
	Total	649	1792	112	2553
O estudo é ensaio clínico?	Sim	438	170	16	624
	Não	81	1524	52	1657
	"Não claro"	19	155	98	272
	Total	538	1849	166	2553
Estudo deve ser incluído?	Sim	26	4	0	30
	Não	12	2416	43	2471
	"Não claro"	13	29	10	52
	Total	51	2449	53	2553

Tabela 2 - Valores da concordância bruta, do índice de concordância kappa e IC95% segundo aspectos abordados e situação do artigo quanto à disponibilidade do resumo.
Table 2 - *Gross agreement, kappa and 95% CI rates according to aspects addressed and abstract availability.*

Artigos	Característica avaliada	Concordância bruta (%)	Kappa	IC95%
Com e sem resumos (n= 2553)	Adequação da população de estudo	89,0	0,46	0,39 – 0,52
	Adequação do tipo de intervenção	82,0	0,59	0,55 – 0,62
	Adequação do tipo de estudo	80,7	0,59	0,55 – 0,63
	Decisão pela inclusão/exclusão	96,0	0,44	0,33 – 0,56
Somente com resumo (n = 2312)*	Adequação da população de estudo	92,2	0,48	0,40 – 0,56
	Adequação do tipo de intervenção	85,3	0,62	0,58 – 0,67
	Adequação do tipo de estudo	84,0	0,61	0,57 – 0,65
	Decisão pela inclusão/exclusão	97,4	0,51	0,38 – 0,64

*excluindo-se os artigos que apresentavam somente o título

ser uma boa alternativa quando o interesse é diminuir a possibilidade de não inclusão de trabalhos. Entretanto, se for ampla demais, como ocorreu no presente estudo, a identificação de trabalhos elegíveis pode passar a ser um evento raro. Segundo Bartko[10], quando a característica de interesse ocorre raramente, a concordância pela ausência desta característica pode ser melhor quantificada pela concordância geral. Assim, a concordância geral de 96,0% pode ser a melhor estatística descritiva a ser utilizada.

No entanto, caso se deseje utilizar uma estatística que extraísse a parcela de concordância explicada pelo acaso, o índice de concordância kappa seria indicado, pois a variável de interesse é qualitativa nominal e os avaliadores são igualmente habilitados na identificação das características de interesse[7,8,11].

A existência de artigos sem resumo disponível apresentou uma dificuldade metodológica extra, que talvez pudesse ser contornada pelo encaminhamento direto destes artigos para revisão do texto completo e tomada de decisão a partir deste procedimento, em substituição ao adotado no estudo - utilizar somente o título para a análise de elegibilidade. A comparação dos indicadores de concordância utilizando-se os dois bancos de dados indica pouca melhora nos valores destes após exclusão dos trabalhos que continham somente título, possivelmente em função do pequeno número relativo destes trabalhos. Um outro estudo poderia avaliar o ganho potencial em termos de concordância caso esta situação se repetisse e o encaminhamento direto para leitura fosse a estratégia utilizada. Evidentemente, uma decisão final sobre a pertinência dessa medida dependeria de uma análise do custo envolvido com esta estratégia.

Com base nos índices de concordância, pode-se suspeitar que houve menor dificuldade em identificar nos resumos os tipos de estudo e de intervenção, em contraposição à população de estudo, para ambos os avaliadores, indicando que grande parte da identificação destas características ocorreu devido à confiabilidade dos resumos na apresentação dos dados necessários. Para a população de estudo, o kappa apresentou menor valor, indicando que para esta característica os resumos possivelmente não eram tão completos. Analisando-se os percentuais de classificação na categoria "não claro" para ambos os bancos – 10,0; 9,8 e 13,3% para tipo de população, de intervenção e de estudo (2.553 artigos) e 4,7; 3,2 e 5,2% (2.312 artigos) – observam-se menores valores quando são avaliados somente artigos comple-

tos, como esperado. Entretanto, tanto nesta situação quanto na avaliação considerando-se todos os artigos, os percentuais de classificação "não claro" parecem maiores para o tipo de estudo.

Alcançar altos índices de concordância na seleção inicial de trabalhos pode ser uma tarefa difícil, considerando que nem sempre os resumos de trabalhos apresentam claramente o tipo de intervenção e o desenho metodológico do estudo, com ausências ainda maiores para a população de estudo.

Existem, na literatura, interpretações para os valores de kappa e, segundo Byrt[12], estas dependem da prevalência do efeito de interesse e da existência de vício – classificações diferentes entre os avaliadores. Este autor, em carta ao Editor da revista *Epidemiology*[13], propôs uma tabela de interpretação ajustando os valores de kappa pela prevalência e vício de classificação, na qual valores deste índice entre 0,41 e 0,60 são considerados como concordância razoável e, entre 0,21 e 0,40, concordância pequena. Nossos dados indicam, se considerados os limites inferiores dos IC, concordância pobre para a característica população e razoável para as outras duas.

Há poucos estudos na literatura que avaliam a concordância entre avaliadores na seleção inicial de artigos. Edwards *et al.*[14] avaliaram a concordância entre pares de avaliadores na seleção de ensaios clínicos aleatorizados para uma revisão sistemática. Os avaliadores procederam à leitura, de forma independente, de 22.571 registros contendo título e resumo, e decidiram sobre a inclusão dos trabalhos com base em critérios de elegibilidade previamente discutidos. A concordância entre dois avaliadores variou entre k=0,59 (IC95%: 0,56 – 0,62) e k=0,93 (IC95%: 0,90 – 0,96), sem ajuste para prevalência e vício de classificação. Assim como no presente estudo, houve concordância de apenas 1% (n=301) na inclusão de trabalhos. Os autores, utilizando a técnica de captura-recaptura, estimaram que a chance de excluir um artigo potencialmente elegível diminuiu de 22%, quando a seleção foi feita por apenas um avaliador, para 4% quando esta foi realizada por pares de avaliadores. Cooper *et al.*[15] avaliaram a concordância entre pares de revisores na seleção de estudos feita de duas maneiras: apenas pelo título (n=185) ou pela leitura dos resumos completos (n=90), preenchendo-se formulários com questões referentes ao desenho metodológico, população de estudo, intervenção, duração do estudo e resultados, seguidas pela decisão de inclusão/exclusão do estudo, e com resolução de casos discordantes por consenso. A concordância entre os avaliadores foi razoável, sendo os índices de concordância kappa entre 0,53 (IC95%: 0,35 – 0,71) e 0,60 (IC95%: 0,43 – 0,77), quando a seleção foi realizada a partir de leitura dos resumos completos, e entre 0,47 (IC95%: 0,36 – 0,58) e 0,66 (IC95%: 0,57 – 0,76), quando se considerou somente os títulos para decisão sobre a elegibilidade. Apesar de os avaliadores terem sido treinados, grande parte das discordâncias foram atribuídas à subjetividade inerente ao processo de seleção de artigos por diferentes avaliadores.

A presença de discordância entre os avaliadores confirma a necessidade de participação de pelo menos duas pessoas nesta fase de estudos de revisão sistemática, pois seriam selecionados conjuntos de trabalhos diferentes caso houvesse a participação de apenas um avaliador.

As reuniões para consenso realizadas nos casos discordantes foram úteis para resolver diferenças de interpretação entre os avaliadores, possibilitando nova leitura e maior reflexão, e colaboraram, portanto, para diminuir as chances de exclusão de um artigo que, na verdade, deveria ser incluído na revisão sistemática.

Referências

1. Egger M, Smith GD, Altman DG. Systematic reviews in health care. In: *Meta-analysis in context*. London: BMJ; 2001.

2. Alderson P, Green S, Higgins JPT, editors. Cochrane reviewer's handbook 4.2.2 [Updated March 2004]. In: *The Cochrane Library*, Issue 1, 2004. Chichester, UK: John Wiley & Sons; Ltd.

3. Egger M, Smith GD, Sterne JAC. Uses and abuses of meta-analysis. *Clin Med* 2001; 6: 478-84.

4. Oliveira JM. *Suplementação de vitamina A em lactantes: revisão sistemática* [minuta da dissertação de Mestrado em processo de pré-banca]. São Paulo: Faculdade de Saúde Pública da USP; 2006.

5. Thomson Scientific. *EndNote*. Version 7.0. ISI [computer program]. Philadelphia; 2003.

6. Microsoft Corporation**.** *Microsoft® Office Excel* [spreadsheet software]. Versão 5.1.2600. IBM; 2003.

7. Stata Corporation. *Statistical software for professionals/ STATA* [computer program]. Versão 9. Texas: College Station; 2005.

8. Cohen J. A coefficient of agreement for nominal scales. *Educ Psycho Meas*. 1960; 20: 37-46.

9. Fleiss JL. Statistical methods for rates and proportions. In: *The measurement of interrater agreement*. 2nd ed. New York: John Wiley & Sons; 1981. pp. 212-35.

10. Bartko JJ, Carpenter WT. On the methods and theory of reliability. *J Nerv Ment Dis* 1976; 163(5): 307-17.

11. Maclure M, Willet WC. Misinterpretation and misuse of the kappa statistics. *Am J Epidemiol* 1987; 126(2): 161-9.

12. Byrt T, Bishop J., Carlin JB. Bias, prevalence and kappa. *J Clin Epidemiol* 1993; 46: 423-9.

13. Byrt T. How good is that agreement? [letter]. *Epidemiology* 1996; 7(5): 561.

14. Edwards P, Clarke M, DiGuiseppi C, Pratap S, Roberts I, Wentz R. Identification of randomized controlled trials in systematic reviews: accuracy and reliability of screening records. *Statist Med* 2002; 21(11): 1635-40.

15. Cooper M, Ungar W, Zlotkin S. An assessment of inter-rater agreement of the literature filtering process in the development of evidence-based dietary guidelines. *Public Health Nutr* 2006; 9(4): 494-500.

recebido em: 06/03/06
versão final reapresentada em: 02/08/06
aprovado em: 03/08/06