Public Health Nutrition: 9(4), 494–500 DOI: 10.1079/PHN2005877

An assessment of inter-rater agreement of the literature filtering process in the development of evidence-based dietary guidelines

Marcia Cooper[1,2]*†, Wendy Ungar[3] and Stanley Zlotkin[1,4,5]

[1]Division of Gastroenterology and Nutrition, Program in Metabolism, Research Institute, The Hospital for Sick Children, Toronto, Ontario, Canada: [2]Department of Nutritional Sciences, University of Toronto, Toronto, Ontario, Canada: [3]Population Health Sciences, The Hospital for Sick Children, Toronto, Ontario, Canada: [4]Departments of Paediatrics and Nutritional Sciences, University of Toronto, Toronto, Ontario, Canada: [5]Centre for International Health, University of Toronto, Toronto, Ontario, Canada

Submitted 10 January 2005: Accepted 8 September 2005

Abstract

Objective: To determine whether the literature filtering process, a vital initial component of a systematic literature review, could be successfully completed by nutrition professionals or non-professionals.

Design: Using a diet–disease relationship as the guideline topic, inter-rater agreement for the title and abstract filtering processes between and among professionals and non-professionals was assessed and compared with an expert reference standard. Predetermined eligibility criteria were applied by all raters to 185 titles and 90 abstracts. Filtering decisions were initially made independently and then revised after a within-pair consensus meeting.

Subjects: The raters were six dietitians (RD) and six nutrition graduate students (Grad). To assess inter-rater agreement (reliability), each group was divided into three pairs.

Results: Weighted and unweighted kappa statistics and percentage agreement were calculated to determine the inter-rater agreement within pairs. Sensitivity and specificity estimates were determined by comparing responses with those of an expert reference standard. Overall, Grad pairs demonstrated greater inter-rater agreement than RD pairs for title filtering ($P < 0.05$); no differences were observed for abstract filtering. Compared with the expert reference standard, every rater and pair had false-negative responses for both title and abstract filtering.

Conclusions: After consensus meetings, both RDs and Grads were comparable in their agreement on title and abstract filtering, although important differences remained compared with the expert reference standard. This study provides preliminary findings on the value of utilising a non-expert pair in developing guidelines, and suggests that the literature filtering process is complex and quite subjective.

Keywords
Literature filtering
Dietary guidelines
Inter-rater agreement
Evidence-based

A widely accepted evidence-based approach for the establishment of practice guidelines incorporates four basic steps: (1) literature retrieval; (2) literature filtering; (3) appraisal of research evidence; and (4) assignment of graded recommendations[1,2]. The final step in the guidelines development process (recommendations) is often completed by a small group of highly trained experts who rely on the quality of steps (1)–(3). However, the preceding steps in the process (literature retrieval, filtering and appraisal) may be completed by non-experts with varying levels of training and experience. The important step of literature filtering can markedly influence the outcome of a guideline, since articles that are omitted at an early stage will not be part of the appraisal step. The systematic review technique is favoured over the simpler consensus-based methodology due to the emphasis on systematically assembled evidence, with explicit inclusion/exclusion criteria, and the attention paid to the methodological quality of the work, without prior assumptions about the findings[3]. There is little published research in the evidence-based science field concerning examination of the inter-rater agreement for different reviewers for literature filtering and assessing both randomised and non-randomised studies and who is best qualified to complete these steps.

†Present address: Nutrition Evaluation Division, Bureau of Nutritional Sciences, 3rd Floor, Frederick Banting Building, Tunney's Pasture, Health Canada, Ottawa, Ontario, Canada, K1A 0L2.

*Corresponding author: Email marcia_cooper@hc-sc.gc.ca

https://doi.org/10.1079/PHN2005877 Published online by Cambridge University Press

Our objective was to determine whether the literature filtering process, a vital initial component of a systematic literature review, could be successfully completed by non-experts. Our goal was achieved by evaluating the inter-rater agreement for the title and abstract filtering processes between and among professionals (dietitians) and non-professionals (nutrition graduate students) and comparing these assessments with an expert reference standard (PhD-trained nutritional scientists) using a dietary guideline topic as the example.

Methods

The raters

The raters were a convenience sample of six dietitians (RD) from the Toronto area and six nutrition graduate students (Grad) from the University of Toronto. Each was recruited to conduct title and abstract filtering. To assess inter-rater agreement, dietitians were paired together in three pairs, as were graduate students. Raters were recruited into the study between April and July 2002. RD_1 represents the pairing of $dietitian_1$ and $dietitian_2$, RD_2 the pairing of $dietitian_3$ and $dietitian_4$, RD_3 the pairing of $dietitian_5$ and $dietitian_6$. $Grad_1$ represents the pairing of $student_1$ and $student_2$, $Grad_2$ the pairing of $student_3$ and $student_4$, and $Grad_3$ the pairing of $student_5$ and $student_6$. Raters received a detailed description of the study protocol and provided their informed consent. The Research Ethics Board at the Hospital for Sick Children, Toronto, Ontario, approved the study.

Literature filtering tools

Literature filtering tools were designed to be specific for this research but adaptable beyond this study. A Title Selection Form was created based on the framework suggested by the American Dietetic Association (ADA) for producing evidence-based guidelines[2]. The ADA suggests that the rater should be supplied with three possible actions for classifying an article at the title filtering stage: 'YES', 'NO' or 'UNABLE TO DETERMINE'. 'YES' indicated that the rater believed the title of the article met the specified inclusion criteria and that the abstract should be read to further determine if this was an applicable article. 'NO' indicated that the rater was certain an article title did not meet the inclusion criteria and therefore should be excluded from further review. 'UNABLE TO DETERMINE' indicated that the rater was not able to decide from the title of the article whether it was an eligible study. Thus, the article would be included for further review.

The Abstract Selection Form was based on a modification of existing abstract filtering forms[1,2]. Although the ADA[2] suggests providing the rater with three actions for selecting an article at the abstract filtering stage ('YES', 'NO' or 'UNABLE TO DETERMINE'), the current research was designed with a dichotomous eligibility rating of 'INCLUDE' or 'EXCLUDE'. The rationale for a two-category final response was that the decision to assess an abstract for further review could be made with greater certainty compared with the decision for a title, and followed a more detailed assessment. Five questions regarding study inclusion criteria were included on the form; these related to the design of the study, population, diet, study length and outcomes. Four response categories were available for each question to determine if a criterion was met. 'YES' indicated that the rater believed the abstract contained the inclusion criteria. 'NO' indicated that the rater was certain the abstract did not contain the inclusion criteria. 'NOT STATED' indicated that the rater was not able to determine whether an inclusion criterion was met but did not feel confident that the criterion was not met. 'NOT APPLICABLE' indicated that a criterion did not apply for a specific study design. Based on the responses to these five criteria, raters assessed whether to 'INCLUDE' or 'EXCLUDE' an abstract from further review. Thus, a rater would 'EXCLUDE' an article if it did not meet one or more inclusion criteria.

Both the title and abstract selection forms and the instructions were developed through a formal process of pilot testing to examine the face validity of the tools. Each selection form went through three successive tests, with changes at each stage. Raters who participated in the pilot work were not the same individuals who were subjects in the inter-rater agreement study.

Rater training

All 12 raters received a one-hour training session given by one of the investigators (M.C.) for both the title and abstract filtering processes. A training manual was provided that included the purpose of the research, the research question and instructions on the inclusion/exclusion criteria for evaluating titles and abstracts. Materials included in the manual were instructions on how to use the forms, definitions, frequently asked questions (FAQs), examples, a study design booklet (for identifying the design of a study within an abstract) and tip sheets. The FAQs provided answers to questions that the rater may have contemplated while undertaking the filtering process. During the group training session, 10 examples of titles and three examples of abstracts were reviewed and any concerns or questions addressed.

Protocol for rating the titles and abstracts

Each rater independently assessed each of the provided titles and abstracts. For title filtering, the raters were instructed to read the title, the journal name and examine the number of pages in the citation to determine the potential eligibility of the article. For abstract filtering, each rater was instructed to read the abstract in its entirety and answer the five questions on the abstract rating form. Upon submitting their responses to the investigators, the degree of disagreement within the pairs of dietitian and

https://doi.org/10.1079/PHN2005877 Published online by Cambridge University Press

student raters was calculated. Each pair of raters was provided with the results of their rating and asked to meet to come to a consensus decision on those titles and abstracts for which disagreements had occurred. After this meeting, there was agreement among each rater pair.

Each rater was provided with a comment sheet to provide feedback on the ease of use and clarity of the tools, and the time spent completing the task. Additionally, comments on any general aspects of the process were solicited.

Statistical methods

The principal outcome of interest was the assessment of inter-rater agreement for the title and abstract filtering process within pairs of professionals and non-professionals. The kappa statistic (κ), which corrects for chance agreement[4–8], was calculated to assess the inter-rater agreement within pairs of raters. Since polytomous responses were required for title filtering responses (i.e. more than two categories), a weighted kappa was determined. Linear weights (W1) were calculated according to Cicchetti and Allison[9]; this method assigns weight to disagreements in a linear manner reflecting the row and column placement. The unweighted (simple) kappa statistic was calculated to assess the inter-rater agreement for abstract filtering, since the final response decision was dichotomous. Using the criteria of Landis and Koch[10], a kappa value of 0.81–0.99 was considered as 'almost perfect' agreement; 0.61–0.80 as 'substantial' agreement; 0.41–0.60 as 'moderate' agreement; 0.21–0.40 as 'fair' agreement; and <0.2 was considered as 'poor' agreement.

To test the difference in the reliability between pairs of professional and non-professional raters trained at the same time, the significance of the difference between two independent values of kappa was calculated with a Z statistic[5]. All results were considered to be statistically significant at $P < 0.05$. The SAS program (version 8.0; SAS Institute Inc., Cary, NC, USA) was used for the kappa computations using the AGREE statement. Standard errors and 95% confidence interval estimates were calculated. Percentage agreement was calculated to demonstrate the observed agreement within pairs of raters.

As a measure of validity, sensitivity, specificity, false negatives and false positives were determined for the individual raters, and the pairs of raters' responses, by comparing these with the expert reference standard[11,12]. Sensitivity is the proportion of 'YES' titles (articles whose abstracts should be evaluated) that are correctly identified by the rater compared with the expert reference standard and should be further evaluated in the filtering process. Specificity is the proportion of 'NO' titles (articles whose abstracts should not be evaluated) that are correctly identified by the rater compared with the expert reference standard and that are not to be further evaluated. Those responses that are rated as a 'NO' when they should have been included are false negatives; those that are rated as a 'YES' when they should have been excluded are false positives. To establish the expert reference standard, two of the investigators (M.C. and S.Z.) rated every title and abstract independently and then met to come to a consensus on all titles and abstracts. Dichotomous responses were tabulated in order to calculate sensitivity and specificity. For the analysis, an 'UNABLE TO DETERMINE' response was converted to a 'YES' response for title filtering since abstracts would be retrieved with either of these responses.

Overall sample size and power are determined by the number of measurements (i.e. raters) and by the number of observations (e.g. research studies) examined. Calculations based on the kappa statistic were conducted to ensure that there were adequate numbers of rated titles and abstracts to estimate reliable kappa statistics, in addition to confidence intervals and standard error values[5]. The number of raters was based on ensuring that there was a sufficient number of pairs to compare the inter-rater reliability results.

Literature retrieval

To complete the filtering process assessment, the investigators chose to search and retrieve literature on the topic of the impact of a reduced-fat diet in children on lipids, coronary heart disease, growth and other outcomes[13]. An analytical framework (schematic diagram) was created to illustrate the complex interactions between a reduced-fat diet and potential outcomes used to guide the systematic literature review. A comprehensive review of the literature on this topic was completed. All articles published from the years 1966 to January 2002, and limited to the English language, human research and the age group 2–18 years, were selected. The databases searched were MEDLINE, CINAHL, EMBASE, PSYCHLIT and the COCHRANE COLLABORATION. For the purpose of our study, we chose a convenience sample of about 5% of the total number of articles identified in the search ($n = 185$). This exceeded the sample size needed to ensure that there were adequate numbers of rated titles and abstracts to estimate reliable kappa statistics, in addition to confidence intervals and standard error values. Thus, 185 titles were reviewed by each rater. In the next step, the sample was subsequently narrowed to 90 abstracts for review by all raters.

Results

Rater demographics

The dietitians ($n = 6$) ranged in age from 27 to 38 years (mean ± standard deviation (SD): 31.3 ± 3.6 years) and had an average of 5.2 years of professional experience. Three dietitians had a Master of Science degree in nutrition, two had a Master of Health Science degree in nutrition and one was studying for a Master of Health Science degree part-time. The graduate students ($n = 6$)

https://doi.org/10.1079/PHN2005877 Published online by Cambridge University Press

ranged in age from 24 to 31 years (mean ± SD: 26.6 ± 2.3 years). Three graduate students were taking their Master of Science degree in nutritional sciences and three graduate students were taking their Master of Health Science degree in community nutrition at the time of this study.

Inter-rater agreement and validity – title filtering

Table 1 shows the percentage agreements and weighted kappa values for the assessment of titles for the pair of experts and for each of the RD pairs compared with the corresponding Grad pairs. Overall, the Grad pairs demonstrated greater agreement than the RD pairs. For groups 1 and 3, there was a significantly greater agreement within the Grad pairs compared with the corresponding RD pairs (RD_1 vs. $Grad_1$, $Z = 3.1$, $P < 0.05$; RD_3 vs. $Grad_3$, $Z = 2.3$, $P < 0.05$). For group 2, there was a trend for the Grad pair to have more consistent agreement than the RD pair, but the difference was not significant. With the exception of RD groups 1 and 3, all other dietitian and student groups had a $\kappa > 0.5$, indicating moderate to substantial agreement[10].

The percentage agreement with the expert reference standard for each rater, along with the sensitivity and specificity of the ratings for each individual and for each pair of raters, are shown in Table 2 for the title filtering process. Compared with the decisions made by the experts, every rater and pair of raters had some false-negative responses (raters excluded some studies that the experts had included). There was wide variation in the raters' responses since certain individual raters from each of the groups better reflected the responses of the expert reference standard.

Inter-rater agreement and validity – abstract filtering

Table 1 reports the percentage agreement and unweighted kappa values for the assessment of the 90 abstracts for the pair of individuals in the expert reference standard and each RD pair compared with the corresponding Grad pair. There was no statistically significant difference in kappa values for the RD compared with any of the corresponding Grad pairs. Moderate agreement was achieved for all dietitian and all graduate student groups[10].

Table 3 shows the percentage agreement with the expert reference standard for each rater, along with the sensitivity and specificity of the ratings for each individual, and the sensitivity and specificity of the ratings for each group of raters for the abstract filtering process. On average, the sensitivity and specificity for the abstract filtering process for the individual dietitians were a bit lower than for the graduate students. Similar to title filtering, certain raters from each of the groups better reflected the responses of the expert reference standard. Additionally, these results indicate that even after having the opportunity for groups to meet together to discuss disagreements, dietitians and students had relatively low sensitivity in their evaluations.

Discussion

At a time of increasing emphasis on guidelines for evidence-based practice and the development of systematic comprehensive reviews of the literature, it is critical to ensure the validity and reliability of the various steps in the guideline development process. Selecting appropriate articles from a search of the literature forms the foundation of any review or guideline developed from selected citations. Over the past 20 years the volume of original research studies has expanded exponentially, resulting in the need to examine methods for keeping up with current literature and synthesising evidence[14]. In the current research, the two experts had only a modestly higher proportion of agreements in their independent decisions compared with any of the trained pairs of raters. In theory, if instructions are totally comprehensive and inclusive, the raters should be in perfect agreement. This was not the case. Among any two raters who independently review a topic, regardless of their expertise, this is an expected finding for a process that has some subjectivity[15]. Any confusion associated with the guideline topic, outcome

Table 1 Within and between group comparisons of inter-rater agreement for title filtering and abstract filtering

	Title filtering			Abstract filtering		
Group	Number of titles	Percentage agreement within pairs	Weighted* kappa (SE; 95% CI)	Number of abstracts	Percentage agreement within pairs	Unweighted kappa (SE; 95% CI)
Expert reference standard (expert 1&2)	185	82	0.74 (0.04; 0.66, 0.83)	90	87	0.73 (0.08; 0.59, 0.87)
RD_1 (dietitian 1&2)	185	66	0.49 (0.06; 0.38, 0.60)	90	81	0.60 (0.09; 0.43, 0.77)
$Grad_1$ (student 1&2)	185	80	0.66 (0.05; 0.56, 0.76)	90	80	0.59 (0.08; 0.43, 0.76)
RD_2 (dietitian 3&4)	185	74	0.64 (0.05; 0.55, 0.74)	90	77	0.53 (0.09; 0.35, 0.71)
$Grad_2$ (student 3&4)	185	79	0.72 (0.04; 0.64, 0.80)	90	77	0.53 (0.09; 0.36, 0.71)
RD_3 (dietitian 5&6)	185	67	0.47 (0.06; 0.36, 0.58)	90	79	0.57 (0.09; 0.40, 0.74)
$Grad_3$ (student 5&6)	185	77	0.66 (0.05; 0.57, 0.76)	90	77	0.55 (0.08; 0.39, 0.71)

SE – standard error; CI – 95% confidence interval.
* Linear weight assigns weight to disagreements in a linear manner reflecting the row and column.

https://doi.org/10.1079/PHN2005877 Published online by Cambridge University Press

Table 2 Sensitivity and specificity for individual raters and groups for the title filtering process

				Groups after consensus†		
Rater	Percentage agreement* with experts	Sensitivity (%)	Specificity (%)	Group	Sensitivity (%)	Specificity (%)
Dietitian$_1$	82	72	86	RD1	79	92
Dietitian$_2$	84	87	83			
Dietitian$_3$	86	85	86	RD$_2$	90	86
Dietitian$_4$	79	90	73			
Dietitian$_5$	92	85	95	RD$_3$	85	87
Dietitian$_6$	79	84	77			
Student$_1$	88	84	90	Grad$_1$	74	93
Student$_2$	83	77	86			
Student$_3$	87	90	85	Grad$_2$	89	90
Student$_4$	85	90	82			
Student$_5$	84	90	82	Grad$_3$	80	90
Student$_6$	87	74	93			

* Percentage agreement refers to the individual dietitians or graduate students agreeing with the expert reference standard on how to classify a title.
† Groups after consensus refers to the calculation of sensitivity and specificity based on mutual decisions on eligibility compared with the expert reference standard.

measures or any part of the process could have influenced the raters' results.

Grad pairs were more consistent than RD pairs in making identical, independent decisions to decide whether to further evaluate a research paper based on the title. Consistency in the proportion of disagreements within a rater category was observed, except for RD$_2$ who had a statistically significant higher inter-rater agreement than the other RD pairings. The high proportion of disagreements among both groups may be explained by several factors. First, the raters were not previously familiar with the literature filtering process and the guideline topic. Second, a three-category response for titles was used ('YES', 'NO' or 'UNABLE TO DETERMINE'), which increased the potential for divergent evaluations within a pair.

For the abstract filtering process, Grad and RD pairs were similar in their inter-rater agreement. Despite this similarity, disagreements within pairs ranged from 19 to 23% of the 90 abstracts that were assessed for eligibility. This was a surprisingly large disagreement considering that these same raters had previously been through the process of title filtering. The continuing disagreements may have been due to the complexity of the topic of review. Despite the guideline methodology being exhaustive, this comprehensiveness may have been a confounding factor in the selection of eligible literature. Although there are advantages associated with this methodology, one could consider treating several lines of evidence from the analytical framework as separate reviews to limit the potential for disagreements which likely vary in proportion to the number and complexity of outcome measures. In addition, some raters may have been inherently more cautious when making decisions, resulting in the over-inclusion of studies, while others may be more inclined to exclude studies. Thus, despite what we believed were clear and precise instructions for inclusion and exclusion of abstracts, the process turned out to be very subjective and identified the need for refined information.

Table 3 Sensitivity and specificity for individual raters and groups for the abstract filtering process

				Groups after consensus†		
Rater	Percentage agreement* with experts	Sensitivity (%)	Specificity (%)	Group	Sensitivity (%)	Specificity (%)
Dietitian$_1$	76	62	89	RD$_1$	67	89
Dietitian$_2$	81	71	91			
Dietitian$_3$	78	82	73	RD$_2$	89	78
Dietitian$_4$	81	84	78			
Dietitian$_5$	79	91	67	RD$_3$	91	76
Dietitian$_6$	84	89	80			
Student$_1$	77	77	86	Grad$_1$	78	91
Student$_2$	88	84	91			
Student$_3$	86	84	87	Grad$_2$	87	80
Student$_4$	78	84	71			
Student$_5$	83	93	73	Grad$_3$	89	80
Student$_6$	80	71	89			

* Percentage agreement refers to the individual dietitians or graduate students agreeing with the expert reference standard on how to classify an abstract.
† Groups after consensus refers to the calculation of sensitivity and specificity based on mutual decisions on eligibility compared with the expert reference standard.

https://doi.org/10.1079/PHN2005877 Published online by Cambridge University Press

The current data show that allowing discussion among pairs to reach consensus was useful for resolving differences in ratings and interpretation between individuals, and in general responses became more similar between the groups and to the expert reference standard. Consensus meetings are commonly used to resolve differences among two raters[1,16], but there has been minimal discussion about the ease or difficulty of this process. Despite the fact that the overall proportion of rater disagreements was similar between pairs, there was a difference in the specific citations that were rated identically to the expert reference standard. Therefore, different judgements were made by the pairs of raters during their consensus meetings. There was not a consistent, systematic training error that could explain the different responses compared with the expert reference standard.

Sensitivity was measured to examine the validity of the choice to include a study compared with the expert reference standard. Too many false positives suggest that the raters did not fully understand the eligibility criteria. Overall, the proportion of false-negative responses for the individual dietitians was similar to that for the students for both title and abstract filtering. Furthermore, the false-negative responses were similar for the RD and Grad pairs for title filtering after the individual raters met with their pair partner, but were higher for the dietitian compared with the student pairs for abstract filtering. These data suggest that there was the potential for removing eligible studies by both groups of raters at either filtering step.

In order to eliminate some of the effects of relatively high disagreements and false-negative responses, one could consider going directly to the abstract filtering step. Despite the added burden of having to read more abstracts, there could be less chance of removing eligible studies early in the process due to limited information. Ambiguity within titles and abstracts can lead to high respondent burden as this makes decisions more difficult, while leading to more false-negative and false-positive responses. Recent initiatives such as the QUOROM and CONSORT statements provide guidelines on what should be contained within an abstract for a meta-analysis and a randomised controlled trial, respectively[17,18]. These initiatives should make abstract filtering an easier process in the future.

A factor that could potentially confound the interpretation of the results is the effect of training. Our hypothesis was that any student or dietitian could make similar rating judgements, if specifically trained using a standardised manual, protocol and set of verbal instructions. Similar to previous research[19], the current study involved training with each of the graduate students and dietitians on the details of the study and the topic of the review. After the initial training, e-mail correspondence or phone calls were conducted with any rater who needed further clarifications. Any major issues or concerns of one rater were addressed in communications to all raters. An aim of our study was to find a balance between training the individuals to follow a protocol and allowing them to make independent decisions.

The current research was conducted with a relatively small sample of articles, thus one cannot precisely determine the impact of a larger sample size of raters or evaluated studies with respect to the outcome. Including more individuals as experts would not necessarily improve upon the results, as this was a consensus decision that served as a validity measure to the ratings by professionals and non-professionals. There is no research to tell us how many people are necessary to take on this expert task. The generalisability of the results can be enhanced if the filtering process is further evaluated across different centres and countries. Despite not knowing how the forms and instructions would function with different groups and nutrition topics, the current work does provide preliminary findings on the reproducibility of the title and abstract filtering processes within professionals and non-professionals.

Since each rater evaluated 185 titles and 90 abstracts, respondent burden may have influenced the responses. The process of filtering 185 titles took an average of 3 to 4 h per rater. Filtering of 90 abstracts took from 4 to 15 h per rater. Although it was suggested that each rater evaluate small batches at a time to prevent fatigue, raters were not explicit in their feedback on whether they actually followed this advice. Certain raters rated and re-reviewed their responses several times before coming to a final decision. This may account for some of the time variance.

Since the Cochrane Collaboration suggests using professionals to conduct literature filtering, it was important to question this practice, for practical reasons. For the topic examined in this review, the initial number of citations identified was over 1000. An expert professional (PhD scientist) may not have the time to review this large a number of citations and alternative strategies for filtering the literature are needed. Even though the raters in this study were designated as non-experts in guideline development, all had either an education or practical background in dietetics and nutrition. Clearly these raters were different from the kind of experts (e.g. senior researchers, professors, etc.) who would normally sit on guideline committees. Similar educational backgrounds between the professional (dietitians) and the non-professional groups (nutrition graduate students) could have influenced the differences in inter-rater agreement between the groups. Despite the similarities between the groups there were also differences. Dietitians had years of professional experience whereas graduate students had no practical nutrition experience. Future research can focus on examining the similarities between groups with more diverse backgrounds such as a comparison between professionals and undergraduate students.

https://doi.org/10.1079/PHN2005877 Published online by Cambridge University Press

Results from the current study demonstrated the similarities and differences in agreement among the different pairs, including the experts. The selection of more specific topics of research to filter, rather than the relationship between diet and many health outcomes, may lead to a lower proportion of disagreements than observed in this study and provide a clearer understanding of the factors that contributed to the differences we observed.

Conclusions/applications

With a growing literature base of articles on how to conduct systematic reviews and develop evidence-based guidelines, the current work provides some preliminary and important findings on the initial stages of the literature filtering process. This study demonstrated that all pairs had disagreements (and risked eliminating potentially important articles) at each filtering stage. Because there were false-negative results for all groups at both the title and abstract filtering stage, it seems appropriate to question whether expert raters can accomplish this task better than non-experts. Since false-negative results may be a consequence of difficulties with the particular topic of review, further research should focus on examining the literature filtering process with a simpler topic of review and different raters to appropriately answer whether any group of non-experts can be utilised to perform this task.

Acknowledgements

We gratefully acknowledge the work of the raters who participated in this study.

Financial support: M.C. received a Danone Institute of Canada Doctoral Award in Nutrition Education and Communication. W.U. is supported by a Canadian Institutes for Health Research New Investigator Career award.

References

1 Cochrane Collaboration Handbook. In: Mulrow CD, Oxman AD, eds. *The Cochrane Library*, 4th ed [database on disk and CD-ROM]. Oxford: The Cochrane Collaboration, Oxford Update Software, 1997.

2 Splett P. *Developing and Validating Evidence-Based Guides for Practice: A Tool Kit for Dietetics Professionals*. Chicago, IL: American Dietetic Association, 1999.

3 Brunner E, Rayner M, Thorogood M, Margetts B, Hooper L, Summerbell C, *et al.* Making public health nutrition relevant to evidence-based action. *Public Health Nutrition* 2001; **4**: 1297–9.

4 Streiner DL. Learning how to differ: agreement and reliability statistics in psychiatry. *Canadian Journal of Psychiatry* 1995; **40**: 60–6.

5 Haas M. Statistical methodology for reliability studies. *Journal of Manipulative and Physiological Therapeutics* 1991; **14**: 19–32.

6 Cohen J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 1960; **20**: 37–46.

7 Fitzmaurice G. Statistical methods for assessing agreement. *Nutrition* 2002; **18**: 694–6.

8 Rigby AS. Statistical methods in epidemiology. V. Towards an understanding of the kappa coefficient. *Disability and Rehabilitation* 2000; **22**: 339–44.

9 Cicchetti DV, Allison T. A new procedure for assessing reliability of scoring EEG sleep recordings. *American Journal of EEG Technology* 1971; **11**: 101–9.

10 Landis JR, Koch JJ. The measurement of observer agreement for categorical data. *Biometrics* 1977; **33**: 159–74.

11 Fletcher RH, Fletcher SW, Wagner EH. *Clinical Epidemiology: The Essentials*, 2nd ed. London: Williams and Wilkins, 1988.

12 Gordis L. *Epidemiology*. Philadelphia, PA: WB Saunders Co., 2000.

13 Cooper MJ, Zlotkin SH. An evidence-based approach to the development of national dietary guidelines. *Journal of the American Dietetic Association* 2003; **103**: S28–33.

14 Margetts BM, Vorster HH, Venter CS. Evidence-based nutrition. *South African Journal of Clinical Nutrition* 2002; **15**: 7–12.

15 Oxman AD, Guyatt GH, Cook DJ, Jaeschke R, Heddle N, Keller J. An index of scientific quality for health reports in the lay press. *Journal of Clinical Epidemiology* 1993; **46**: 987–1001.

16 Arrive L, Renard R, Carrat F, Belkacem A, Dahan H, Le Hir P, *et al.* A scale of methodological quality for clinical studies of radiologic examinations. *Radiology* 2000; **217**: 69–74.

17 Moher D, Cook DJ, Eastwood S, Olkin I, Rennie D, Stroup DF. Improving the quality of reports of meta-analyses of randomised controlled trials: the QUOROM statement. Quality of Reporting of Meta-analyses. *Lancet* 1999; **354**: 1896–900.

18 Moher D, Schulz KF, Altman DG; CONSORT group. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. *Journal of the American Podiatric Medical Association* 2001; **91**: 437–42.

19 Sands ML, Murphy JR. Use of kappa statistic in determining validity of quality filtering for meta-analysis: a case study of the health effects of electromagnetic radiation. *Journal of Clinical Epidemiology* 1996; **49**: 1045–51.

https://doi.org/10.1079/PHN2005877 Published online by Cambridge University Press