Public Health Nutrition: 11(10), 1006–1014 doi:10.1017/S1368980007001334

Reproducibility of systematic literature reviews on food, nutrition, physical activity and endometrial cancer

RL Thompson[1,*†], EV Bandera[2], VJ Burley[3], JE Cade[3], D Forman[3], JL Freudenheim[4], D Greenwood[3], DR Jacobs Jr[5], RV Kalliecharan[3], LH Kushi[6], ML McCullough[7], LM Miles[8], DF Moore[2,9], JA Moreton[3], T Rastogi[10] and MJ Wiseman[1,8]

[1]Institute of Human Nutrition, University of Southampton, Southampton, UK: [2]The Cancer Institute of New Jersey, University of Medicine and Dentistry of New Jersey–Robert Wood Johnson Medical School, New Brunswick, NJ, USA: [3]Centre for Epidemiology and Biostatistics, University of Leeds, Leeds, UK: [4]Department of Social and Preventive Medicine, School of Public Health and Health Professions, University at Buffalo, Buffalo, NY, USA: [5]Division of Epidemiology, University of Minnesota School of Public Health, Minneapolis, MN, USA: [6]Division of Research, Kaiser Permanente, Oakland, CA, USA: [7]Epidemiology and Surveillance Research, American Cancer Society, Atlanta, GA, USA: [8]World Cancer Research Fund International, London, UK: [9]Biostatistics, School of Public Health, University of Medicine and Dentistry of New Jersey, New Brunswick, NJ, USA: [10]Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, MD, USA

Submitted 22 November 2006: Accepted 9 October 2007: First published online 6 December 2007

Abstract

Objective: Despite the increasing dependence on systematic reviews to summarise the literature and to issue public health recommendations, the formal assessment of the reliability of conclusions emerging from systematic reviews has received little attention. The main goal of the present study was to evaluate whether two independent centres, in two continents, draw similar conclusions regarding the association of food, nutrition and physical activity and endometrial cancer, when provided with the same general instructions and with similar resources.

Design: The assessment of reproducibility concentrated on four main areas: (1) paper search and selection; (2) assignment of study design; (3) inclusion of 'key' papers; and (4) individual studies selected for meta-analysis and the summary risk estimate obtained.

Results: In total 310 relevant papers were identified, 166 (54 %) were included by both centres. Of the remaining 144 papers, 72 (50 %) were retrieved in the searches of one centre and not the other (54 in centre A, 18 in centre B) and 72 were retrieved in both searches but regarded as relevant by only one of the centres (52 in centre A, 20 in centre B). Of papers included by both centres, 80 % were allocated the same study design. Agreement for inclusion of cohort-type and case–control studies was about 63 % compared with 50 % or less for ecological and case series studies. The agreement for inclusion of 138 'key' papers was 87 %. Summary risk estimates from meta-analyses were similar.

Conclusions: Transparency of process and explicit detailed procedures are necessary parts of a systematic review and crucial for the reader to interpret its findings.

Keywords
Systematic literature review
Endometrial carcinoma
Diet
Epidemiological studies
Reproducibility

Among non-communicable diseases cancer is the second leading cause of death, estimated as responsible for 7·6 million deaths in 2005, second to cardiovascular disease (17·5 million deaths)[1]. The incidence of cancer can be reduced by 30–40 % by dietary and other lifestyle changes[2]. The published literature on diet and cancer has increased almost tenfold from 1168 articles for the years 1966 to 1975 to 9820 for 1996 to 2005 (using the terms 'cancer' and 'diet' in PubMed).

The World Cancer Research Fund and the American Institute for Cancer Research (WCRF/AICR) have published a second report on *Food, Nutrition, Physical Activity and the Prevention of Cancer: A Global Perspective*[3], based on cancer site-specific systematic literature reviews (SLRs), to explore causal dietary, physical activity or nutritional links with cancer. These were carried out in independent academic institutions in Europe

† Correspondence address: British Nutrition Foundation, High Holborn House, 52–54 High Holborn, London WC1V 6RQ, UK.

*Corresponding author: Email r.thompson@nutrition.org.uk

and the USA[4]. This updated report builds upon the knowledge base that resulted in the 1997 report, *Food, Nutrition and the Prevention of Cancer: A Global Perspective*[2], and further provide summary risk estimates from SLRs, where permissible. A Methodology Task Force was convened to guide the development of a manual with a detailed set of guidelines for conducting systematic reviews of observational epidemiological studies and intervention studies on aetiology of cancer in terms of food, nutrition and physical activity[3].

SLRs and meta-analyses are increasingly used to combine the results of epidemiological studies to provide an overall assessment of dietary risk factors, with the ultimate goal of issuing public health recommendations for cancer prevention. A systematic review uses a predefined, explicit methodology to minimise bias. Meta-analyses from SLRs can give more powerful and less biased estimates of effect than individual studies or non-systematic reviews[5,6]. However, despite its growing use for summarising the literature and for public policy, to our knowledge, assessment of the reliability of conclusions emerging from SLRs on diet and cancer has received essentially no attention.

Validity and reproducibility of systematic reviews are important because systematic reviews carry more weight than single studies as an evidence base for new policies and treatments. The validity and reproducibility of systematic reviews have not been extensively studied, but discrepancies in conclusions made by different reviews of the same topic have been discussed[7,8]. In one assessment of systematic reviews of observational studies investigating oral contraceptives and rheumatoid arthritis, the authors noted different effect sizes of the individual studies used in the meta-analysis[9]. They also reported concurrence between the conclusions of reviews by authors who had also published primary research in this area[9]. In meta-analysis of published articles, absent publication bias, identification of papers and selection of relevant studies for inclusion are critical issues.

The main goal of the present study was to answer the question: will two independent centres, in two continents, identify the same studies and draw the same conclusions regarding the association of food, nutrition and physical activity and endometrial cancer, when provided with same general instructions and with availability to similar resources?

Methods

Conducting the SLRs

Two centres independently conducted an SLR of food, nutrition and physical activity and the risk of endometrial cancer. Research teams were chosen in the USA and UK to ensure that differences in process and interpretation between Europe and the USA (such as availability of literature databases) were addressed. A manual (WCRF/AICR SLR Specification Manual version 7) was provided. The specification manual required review centres to report results for all study designs, including case series, ecological, cross-sectional, case–control, cohort and intervention studies from published peer-reviewed articles. Foreign-language papers identified by the SLR centres were translated. An algorithm was developed for the review centres to ensure that study designs were defined consistently and in a standardised manner. Study design was assessed by answering a series of yes/no questions whereby the user was directed to a correct descriptor of study design. Centres were asked not to exclude relevant studies on the basis of perceived quality. To limit bias, inclusion of relevant papers, study design assignment and data extraction (of study characteristics, quality issues and results) had to be completed independently by two reviewers at each site and differences resolved between the reviewers or with a third party.

Centres were required to develop their own search strategies to search a list of predefined bibliographic databases from date of inception. A list of relevant exposures was provided, and the centres were also allowed to use additional resources. The centres were asked to hand search journals not included in electronic databases, as well as reference lists of included papers. Where a reference was located on more than one database, only one copy of the reference was kept and the database from which the reference was first identified was recorded as the source. Centres were responsible for developing their own data extraction forms for recording study characteristics, quality issues and results of studies. The centres were responsible for deciding when it was appropriate to carry out a meta-analysis and the methods used.

The centres did not communicate with each other during the review process, but had access to the same review coordinator (R.L.T.) who provided guidance on the instructions in the specification manual. Each centre's protocol and final report was peer-reviewed by a different team of experts in cancer/nutrition, SLR and statistics. The protocols and final reports were also peer-reviewed by WCRF/AICR to ensure that instructions in the specification manual had been followed.

Assessing reproducibility

The overall aim was to assess whether two independent centres, in two continents, draw the same conclusions regarding association of food, nutrition and physical activity and endometrial cancer, when provided with the same manual and similar resources.

The specific research questions addressed were:

- Were the same papers identified as relevant by both centres? If not, why were papers included by one centre and not by the other centre?

https://doi.org/10.1017/S1368980007001334 Published online by Cambridge University Press

- Did the centres assign the same study design to papers?
- Were papers identified as 'key' by one centre also included as relevant by the other centre?
- For exposures linked to endometrial cancer in the first WCRF/AICR report[2], how did the results of the included studies and pooled risk estimates compare?

The two review centres were unaware of the specific issues being evaluated. They were told to produce SLRs and meta-analyses, when appropriate, of all relevant data. A protocol with the proposed methodology was submitted to WCRF in June 2004 and a final report was submitted in December 2004. Databases with papers found and data extraction sheets were also sent to WCRF. The review coordinator at WCRF compiled and compared results from the two centres.

The review coordinator (R.L.T.) determined the reasons for papers not being included as relevant by both centres, by examining lists of articles retrieved in searches, reviewing databases and search terms, and determining at what stage each team excluded the paper from consideration (e.g. after reviewing titles/abstracts or the full paper). Each centre was further asked to identify 'key' papers they would be concerned about if they were missing from an SLR on 'food, nutrition, physical activity and risk of endometrial cancer' carried out by another review team. Agreement from meta-analyses was assessed by comparing summary risk estimates and 95% confidence intervals from each centre. To limit the number of exposures compared, we restricted the exposures to those linked to endometrial cancer in the first WCRF/AICR report[2] (fruit, non-starchy vegetables, animal fats (as foods), saturated fat, body mass index (BMI)).

Statistical analysis

Percentage agreement and kappa statistics were used to assess agreement for inclusion of papers and for assigned study design. A kappa statistic of 0·75 is regarded as excellent agreement and a value of 0·4–0·75 as fair to good agreement[10].

Results

Search and assessment of relevance of papers

Both centres conducted their searched between June and July 2003. Figure 1 presents a comparison of the papers retrieved and included as relevant by the two centres. From the combined searches of the two centres 9695 records were downloaded from databases, 720 were regarded as potentially relevant when reading titles and abstracts, and 310 were regarded as relevant upon reading the full paper. Centre A regarded 272 papers as relevant and centre B regarded 204 papers as relevant. A total of 166 (54%) papers were identified as relevant by both centres, an additional 106 were regarded as relevant by centre A and an additional 38 were regarded as relevant by centre B. Agreement was also assessed by language of the paper. The agreement for the 262 English-language papers was 58% compared with 27% for the 48 non English-language papers.

Centre A searched 17 databases and centre B searched 13 databases. The major source of relevant papers for both centres was Medline (82% for each centre). Non-database sources (including bibliographies in published papers) contributed nearly 10%. Of the other databases searched, Embase, ISI Web of Science, LILACS, Pascal and Old Medline identified the greatest number of relevant papers.

The discrepancy in included papers was a result of:

- Papers picked up in the searches of one centre but not the other (54 in centre A, 18 in centre B).
- Papers found in both searches and regarded as relevant by one centre, but not the other, when reading titles and abstracts or full copies of papers (52 in centre A, 20 in centre B).

Kappa statistics were computed at various stages. The value of κ was 0·45 for the selection of potentially relevant papers from the total number downloaded from databases. For the selection of included (relevant) papers from those identified as potentially relevant, κ was 0·55. For the overall process (selection of included (relevant) papers from those downloaded), κ was 0·69.

Table 1 shows the source of papers included as relevant by one centre, but not retrieved in the search of the other centre. Medline and hand searching contributed the most to the discrepancy between the centres. The hand search by centre A found 19 papers while hand searching by centre B found four papers. Eight papers retrieved by centre A were in databases that centre B did not search. For the papers retrieved in both searches but regarded as relevant by one centre only (52 in centre A and 20 in centre B, see Fig. 1) we assessed whether the discrepancy occurred while reading titles/abstracts or the full paper. For 63 out of 72 papers the discrepancy occurred while reading the title/abstract. Difference in assessment of relevance when reading the full paper was less of a problem; and occurred for nine papers.

Allocation of study design

Of the 166 papers included by both centres, 133 (80%) were assigned the same study design by each centre ($\kappa = 0{\cdot}62$). The main source of discrepancy was in the classification of cohort study subtypes (e.g. case-cohort vs. prospective). This was due to lack of essential information in some of the original papers regarding methodology, making the allocation of study design difficult. In a secondary analysis where cohort-type studies were assessed as one group, the agreement for all study designs was 93% ($\kappa = 0{\cdot}85$). Table 2 shows that for

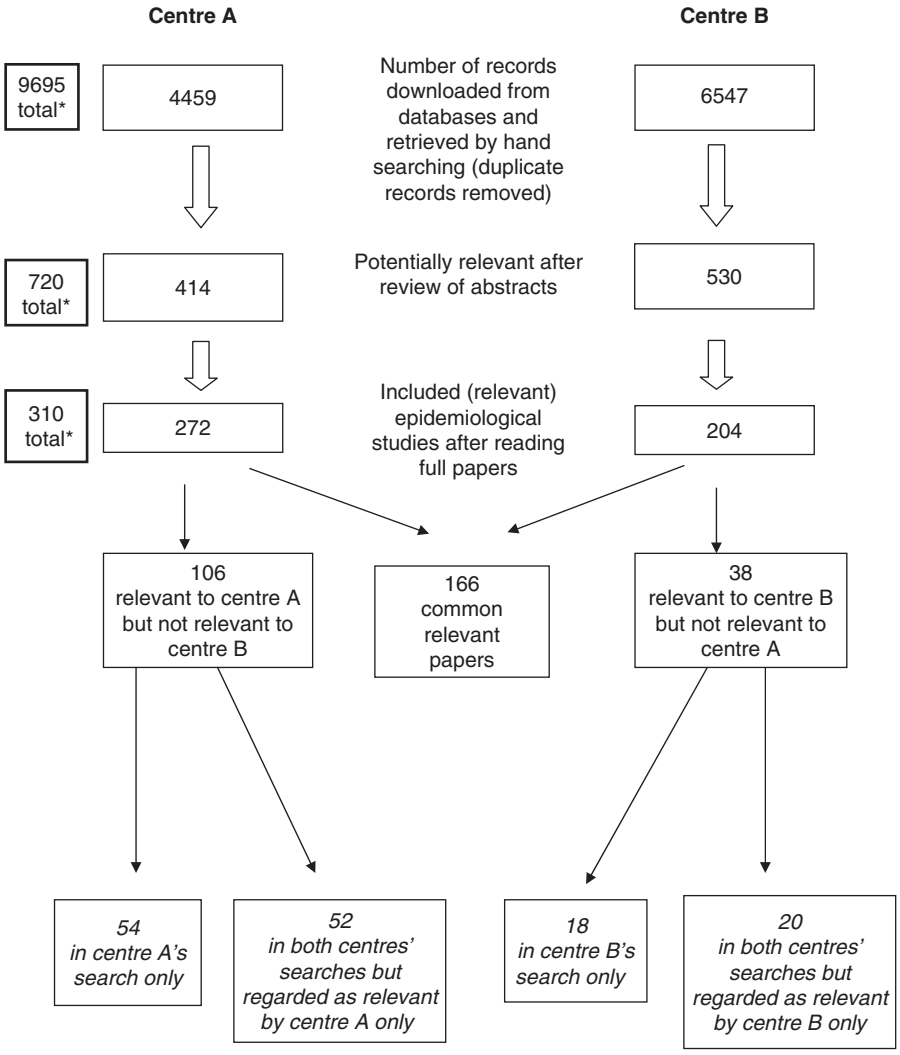

Fig. 1 Flow diagram of the number of papers found and included by each centre; *total number of unique records from both centres

Table 1 Sources of papers included as relevant by one centre, but not retrieved in the search of the other centre

Source	In centre A's search only	In centre B's search only
Medline	24	8
Embase	2	4
Web of Science	3	2
LILACS	1	0
Pascal	2	N/A
Pre-Medline	4	0
Old Medline	6	N/A
Hand searching	19	4
Total	54	18

N/A, did not search database.

cohort-type and case–control studies the agreement was more than 60%, whereas for case series studies it was only 14%. The disagreement in case series classification was because one of the centres classified baseline data in cohort studies of endometrial cancer survivorship as case series.

Agreement for 'key' papers

Figure 2 shows that a total number of 138 papers were identified by either centre as 'key' and 120 (87%) were included by both centres.

Comparison of risk estimates

The centres were responsible for deciding when it was appropriate to carry out meta-analyses. The criteria used by the centres were similar. Both included studies if risk estimates (dichotomous or quantiles) were reported with 95% confidence intervals. Centre A also included continuous risk estimates. Both centres used the comparison of extreme categories method for meta-analysis, which uses a risk estimate for the highest versus lowest quantile of dietary exposure. Summary estimates for each centre using a random effects model for several exposures linked to endometrial cancer in the first WCRF/AICR report[2] were compared. Centre A carried out study design-specific meta-analysis (with the exception of BMI where separate and combined analyses were undertaken).

https://doi.org/10.1017/S1368980007001334 Published online by Cambridge University Press

Table 2 Comparison of inclusion of papers as relevant between centres by study design

Study design	Total no. of included papers	No. (%) included as relevant by both centres	No. included as relevant by centre A, but not centre B	No. included as relevant by centre B but not centre A
Case series	59	8 (14)	41	10
Ecological	8	4 (50)	1	3
Cross-sectional	1	0 (0)	0	1
Case–control	183	116 (63)	50	17
Cohort-type*	59	38 (64)	14	7
Intervention	0	0	0	0
Total	310	166 (54)	106	38

*Cohort-type includes prospective cohort, case-cohort, retrospective cohort and nested case–control studies.

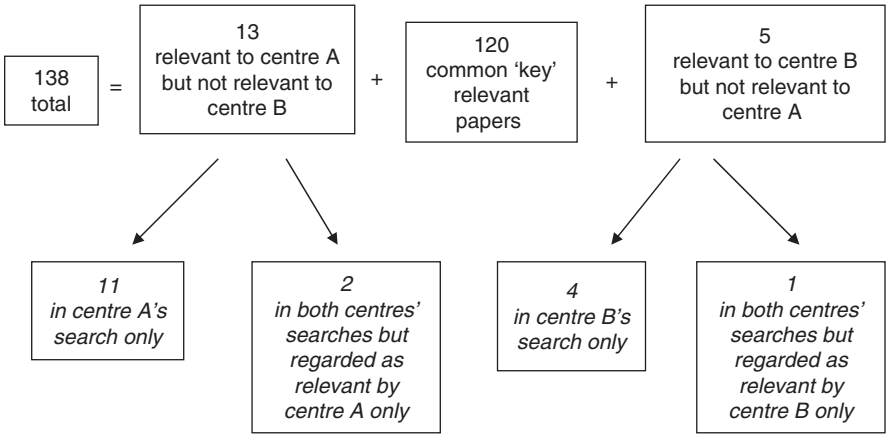

Fig. 2 Flow diagram of the papers considered as 'key' by one centre but not included by the other center

Due to the small number of cohort studies (no more than two) identified for all exposures assessed apart from BMI, centre A performed analyses only for case–control studies. Centre B combined case–control and cohort-type studies in the same analysis. Meta-analysis was not carried out for saturated fatty acids by centre A.

Table 3 shows similar numbers of studies reporting risk estimates in the direction of increased or decreased risk. Studies included by one centre only were further evaluated. Studies that did not report results as odds ratios or relative risks were included as relevant by both centres, however; only one centre included these studies in their report tables. Eleven studies were not picked up in the searches of both centres. Where a different risk estimate was used, this was due to different exposure definitions being used (e.g. total vegetables and cooked vegetables) or a different analysis model; for example, one centre may have chosen an age-adjusted risk estimate and the other chose the most adjusted risk estimate. Similar numbers of studies were included in the meta-analyses.

The summary estimates from both centres were very similar with the exception of animal fat (Fig. 3). For animal fat, a greater summary odds ratio (1·85 vs. 1·37) was reported in the centre A analysis. The centre B analysis included two cohort studies in addition to the same four case–control studies as centre A, which led to the lower summary estimate and greater heterogeneity. On further evaluation, the discrepancy was related to centre B including studies that reported animal fat as both a food group and a nutrient, whereas centre A included only animal fat as a food group. For BMI, both pooled estimates included cohort-type and case–control studies and the summary risk estimates were close to 3·0 for both centres.

Although the estimates for non-starchy vegetables were almost identical, significant heterogeneity was detected by the centre B analysis, but not by centre A (Table 3). Centre B included studies reporting results for green vegetables in the meta-analysis, whereas centre A restricted analysis to total vegetables. This resulted in centre B including two studies with the lowest and highest effect sizes, which increased the heterogeneity.

Discussion

As an initial task for the WCRF/AICR second report on *Food, Nutrition, Physical Activity and the Prevention of Cancer: A Global Perspective*[3], we conducted an assessment of the reproducibility of conclusions from systematic reviews of epidemiological literature, using the example of diet, nutrition, physical activity and endometrial cancer. Our findings suggested that while

https://doi.org/10.1017/S1368980007001334 Published online by Cambridge University Press

Table 3 Results for exposures linked to endometrial cancer in the first WCRF/AICR report (pooled estimates are random effects)

Exposure	Centre	No. of relevant studies	No. of studies showing decreased risk (↓), increased risk (↑) and no effect (=)	Studies only included by one centre	No. of studies in which a different estimate from the same study was used	No. of studies included in meta-analysis (no. of cc studies)	Effect size for highest vs. lowest category	*P* value for heterogeneity	Reasons not included in meta-analysis (no. of studies)
Fruit	A	12 cc	6 ↓; 6 ↑	1 cc	4 cc	9 (9)	1·08 (0·55, 2·10)	<0·0001	No CI (3)
	B	11 cc	6 ↓; 5 ↑			8 (8)	0·93 (0·73, 1·18)	0·001	No CI (1)
									Cont. estimate (2)
Non-starchy vegetables	A	9 cc	6 ↓; 2 ↑; 1 =	4 cc	2 cc	7 (7)	0·68 (0·53, 0·86)	0·15	No CI (2)
	B	11 cc	7 ↓; 3 ↑; 1 =			8 (8)	0·70 (0·53, 0·92)	<0·001	No CI (1)
									Cont. estimate (2)
Animal fats	A	6 cc	5 ↑; 1 unclear	2 cohort	None	4 (4)	1·85 (1·23, 2·79)	0·07	No CI (2)
	B	2 cohort	1 ↓; 1 =	1 cc		6 (4)	1·37 (0·82, 2·27)	<0·001	No CI (1)
		5 cc	5 ↑						
Saturated fat	A	1 cohort	1 ↓	None	None	NP			
		6 cc	1 ↓; 5 ↑						
	B	1 cohort	1 ↓			4 (3)	1·11 (0·74, 1·66)	0·05	No CI (1)
		6 cc	1 ↓; 5 ↑						Cont. estimate (2)
									Not known (1)
Body mass index	A	9 cohort	9 ↑	6 cohort	1 cohort	36 (29)	2·96 (2·49, 3·52)	<0·001	No CI (8)
		35 cc	35 ↑	23 cc	5 cc				
	B	11 cohort	11 ↑			32 (24)	3·03 (2·38, 3·85)	<0·001	No CI (17)
		44 cc	43 ↑; 1 unclear						Not known (6)

WCRF, World Cancer Research Fund; AICR, American Institute for Cancer Research; cc, case–control; CI, confidence interval; Cont. estimate, continuous estimate; NP, not performed.

Fig. 3 Graphical plot of summary odds ratio (with 95% confidence interval shown by horizontal bars) from each centre for fruit, non-starchy vegetables, animal fats and body mass index

our SLRs conducted at two independent centres on different continents showed some differences in terms of the number of citations retrieved or decisions on relevance, the overall conclusions, particularly regarding which studies were most important and pooled risk estimates, were comparable. In the assessment of reproducibility we attempted to answer a series of questions based on the search, assignment of study design, inclusion of 'key' papers and results of meta-analyses.

Only 54% of the papers identified as relevant were included by both centres. The two reasons for this discrepancy were: (1) papers were not retrieved in the search, due to different databases being searched and different search terms; and (2) papers were identified in the search but were subsequently excluded on relevance. The centres had different interpretations of relevance. For example, older papers from Old Medline and multiple publications from the same study reporting the same or similar data were regarded as relevant by centre A (but were excluded from meta-analyses); likewise, centre A retrieved more case series publications, but again these were not included in the meta-analyses. Centre B regarded papers as relevant if they had extractable data

https://doi.org/10.1017/S1368980007001334 Published online by Cambridge University Press

that were included in the report tables. Centre B also excluded papers containing duplicate data and papers that did not have extractable data. Whether duplicate papers are included and then excluded at the analysis level, or excluded initially, clearly should not affect conclusions. What is important is that authors clearly indicate what the inclusion/exclusion criteria were and the decision process used. The reason for this discrepancy is unclear. However, this may be a result of the poorer agreement for case series studies, as many of the foreign-language papers were case series.

The assignment of study design generally concurred although there was some discrepancy in identifying different types of cohort studies, often due to ambiguity in the original source. We believe that if study quality had been one of the inclusion criteria, discrepancies in study design allocation would have been reduced considerably.

When the assessment was restricted to 'key' papers the agreement increased to 87 %. Precise instructions on identifying 'key' papers were not given to the SLR centres and centres found the classification of 'key' considerably subjective. This was particularly true for studies with several publications: choosing one of them as the 'key' paper for that particular study was especially challenging. None the less, there does appear to be better agreement on those papers considered most relevant. However, due to the subjective nature of the assessment, this result should be regarded with caution.

Finally we compared the study results and analyses from each centre for exposures linked to endometrial cancer in the first WCRF/AICR report[2]. Although only 54 % of papers used were common to both centres, similar results and summary estimates were found. The discrepancies in results were mainly due to different interpretations of exposure definitions, choice of analysis for inclusion in the meta-analyses, and presentation of results in report tables, rather than missing studies in the search process. Differences in the width of confidence intervals were observed, although there was substantial overlap. Larger confidence intervals were related to inclusion of more than one study design, a more heterogeneous exposure definition and more studies included in the meta-analysis. More homogeneous exposure definitions and separate analyses for different study designs by centre A were associated with less heterogeneity.

As part of this reliability test, one clear lesson we learned was the wealth of the epidemiological literature on diet and cancer. Endometrial cancer was chosen as a site with relatively limited literature but a reasonable number of studies that had examined these questions, and thus best suited to conduct focused SLRs and conduct reliability assessment. We clearly underestimated the volume of manuscripts and the time that data extraction and analyses would take. Initially the project was allocated to take 4 months, which was extended to 7 months and centres still struggled to conduct searches, review thousands of citations, decide on relevance, obtain manuscripts, conduct data extraction on each relevant paper, tabulate data, select unique analyses from each study, and conduct meta-analyses on relevant exposures by the allocated time. Not only did we find many more citations and manuscripts that we anticipated, we found many other unexpected time-consuming tasks, such as identifying papers coming from the same study (sometimes this was much more difficult than it might appear) or extracting the large volume of data from some of the included manuscripts. For example, data extraction for some manuscripts resulted in hundreds of rows of data, as centres were asked to extract all relevant data, including all statistical models presented, as well as subgroup analyses for all relevant exposures, which included all dietary, nutrition and physical activity variables. Reliability would undoubtedly have been better if centres had more focused associations to evaluate and more time to conduct the SLRs and meta-analyses.

Other authors have addressed discrepancies between reviews of the same topic. Comparisons of reviews on critically ill patients[7], oral contraceptives and rheumatoid arthritis[9], complementary medicine[11], treatment for sciatica[12], diagnosis of angina pectoris[13], treatment for recurrent spontaneous abortion[14] and the impact of low-fat diets in children[15] have been published. There is a lack of information on epidemiological studies on diet and cancer. These studies also report substantial differences in the number and type of papers retrieved. One study reported similar conclusions from two different reviews on the same topic although different methods of meta-analysis had been used[14]. Other studies reported conflicting conclusions[7,9,11,12].

We did not attempt to evaluate validity of the SLRs because it was not possible to determine the true number of relevant studies that have been published. We conclude that some element of subjectivity is inevitable in search and review strategies. The definition of 'relevance' is not universal; however, detailed discussion on the purpose for the review and on manuscript criteria will aid in focusing on the most important papers to include.

The question being addressed by the WCRF/AICR reviews is very broad and multiple exposures are being investigated, rather than a small defined list as is used for most Cochrane reviews. It could be argued that a broader question might lead to a greater possibility of discrepancies than reviews with a narrower question. As centres included all studies regardless of quality this may have reduced disagreements on relevance, as assessing quality is subjective. However, the meta-analysis requires more detail to be provided, thereby reducing the chance of poorer-quality studies being included in meta-analyses.

Our assessment of reproducibility early in the process led to a number of changes to the specification manual for conducting SLRs for the second WCRF/AICR report

(the latest version is available online[16]). A standard search strategy to search Medline was developed. Furthermore, the list of included papers was circulated to principal investigators of large studies to scan for missed papers (particularly from cohort studies). The study algorithm has been revised to clarify the distinction between different cohort-type studies and is now part of the specification manual. Separate analyses for cohort and case–control studies were required to be conducted. Precise instructions and Stata codes on how to carry out meta-analyses were also included in the instruction provided to the SLR centres[17,18]. Thus a more robust process has been developed as a result of direct feedback from this reliability study. The time frame for the overall project was also extended. Other changes to the methodology were implemented as a result of feedback from the centres but are not the subject of this paper.

This study was carried out in relation to dietary factors that were associated with endometrial cancer in the 1997 report. It is possible that the reproducibility results of other dietary exposures might not be as similar. The number of studies eligible for meta-analysis was not large and hence a difference in the risk estimate of one or two studies may have a large impact on the summary estimate. For other cancers such as breast and colorectal, where many more studies are able to be included in meta-analysis, differences in one or two studies may not have a large impact unless they came from very large studies that contributed a high percentage of weight to the overall summary estimate.

We conclude that reproducibility of systematic literature reviews on diet and cancer across two independent centres with access to similar resources was good, overall. Papers retrieved and included in SLRs will inevitably vary to some degree based on subjectivity of perceived study relevance. However, in this study where two centres were provided with general guidelines, similar 'key' papers were identified and meta-analyses arrived at similar conclusions. Transparency in the review process is critical and authors need to explain each step so that the reader can make his/her own conclusions, both regarding the epidemiological evidence and quality of the meta-analysis.

Acknowledgements

This work was funded by WCRF/AICR. We also wish to acknowledge the contribution of other team members and research staff of the SLR centres, and of members of the Secretariat and Methodology Task Force of the WCRF/AICR expert report.

Conflict of interest: None.

Role and nature of each author's contribution: R.L.T. – review coordinator, concept and design of study, analysis, preparation of first draft; E.V.B. – project leader for US centre, data collection, involved in preparation of US systematic review, critical comments on drafts of manuscript; V.J.B. – project manager for UK centre, data collection, involved in preparation of UK systematic review, critical comments on drafts of manuscript; J.E.C. – project leader for UK centre, data collection, involved in preparation of UK systematic review, critical comments on drafts of manuscript; D.F. – project leader for UK centre, data collection, involved in preparation of UK systematic review, critical comments on drafts of manuscript; J.L.F. – member of US centre, data collection, involved in preparation of US systematic review, critical comments on drafts of manuscript; D.G. – member of UK centre, statistician, involved in preparation of UK systematic review, critical comments on drafts of manuscript; D.R.J. – member of US centre, statistician, involved in preparation of US systematic review, critical comments on drafts of manuscript; R.V.K. – member of UK centre, data collection, involved in preparation of UK systematic review, critical comments on drafts of manuscript; L.H.K. – project leader for US centre, data collection, involved in preparation of US systematic review, critical comments on drafts of manuscript; M.L.M. – member of US centre, data collection, involved in preparation of US systematic review, critical comments on drafts of manuscript; L.M.M. – concept and design of study, critical comments on drafts of manuscript; D.F.M. – member of US centre, statistician, involved in preparation of US systematic review, critical comments on drafts of manuscript; J.A.M. – project manager for UK centre, data collection, involved in preparation of UK systematic review, critical comments on drafts of manuscript; T.R. – member of US centre, data collection, involved in preparation of US systematic review, critical comments on drafts of manuscript; M.J.W. – project director, concept and design of study, critical comments on drafts of manuscript.

References

1. World Health Organization (2005) *Preventing Chronic Diseases: A Vital Investment*. Geneva: WHO.
2. World Cancer Research Fund/American Institute for Cancer Research (1997) *Food, Nutrition and the Prevention of Cancer: A Global Perspective*. Washington, DC: AICR.
3. World Cancer Research Fund/American Institute for Cancer Research (2007) *Food, Nutrition, Physical Activity, and the Prevention of Cancer: A Global Perspective*. Washington, DC: AICR.
4. Heggie SJ, Wiseman MJ, Cannon GJ, Miles LM, Thompson RL, Stone EM, Butrum RR & Kroke A (2003) Defining the state of knowledge with respect to food, nutrition, physical activity, and the prevention of cancer. *J Nutr* **133**, 3837S–3842S.
5. Egger M, Davey-Smith G & Altman D (2001) *Systematic Reviews in Health Care*. London: BMJ Books.
6. Glasziou P, Vandenbroucke J & Chalmers I (2004) Assessing the quality of research. *BMJ* **328**, 39–41.
7. Cook DJ, Reeve BK, Guyatt GH, Heyland DK, Griffith LE, Buckingham L & Tryba M (1996) Stress ulcer prophylaxis in critically ill patients. Resolving discordant meta-analyses. *JAMA* **275**, 308–314.

8. Jadad AR, Cook DJ & Browman GP (1997) A guide to interpreting discordant systematic reviews. *CMAJ* **156**, 1411–1416.
9. Pladevall-Vila M, Delclos GL, Varas C, Guyer H, Bruges-Tarradellas J & Anglada-Arisa A (1996) Controversy of oral contraceptives and risk of rheumatoid arthritis: meta-analysis of conflicting studies and review of conflicting meta-analyses with special emphasis on analysis of heterogeneity. *Am J Epidemiol* **144**, 1–14.
10. Kirkwood BR & Sterne JAC (2003) *Essential Medical Statistics*, 2nd ed. London: Blackwell Science.
11. Linde K & Willich SN (2003) How objective are systematic reviews? Differences between reviews on complementary medicine. *J R Soc Med* **96**, 17–22.
12. Hopayian K & Mugford M (1999) Conflicting conclusions from two systematic reviews of epidural steroid injections for sciatica: which evidence should general practitioners heed? *Br J Gen Pract* **49**, 57–61.
13. Gomez Gras E, de Villar Conde E, Lacalle Remigio JR, Pérez de la Blanca EB, Reyes Domínguez A, Alvarez Gil R, Pérez Lozano MJ & Marín León I (1999) A reproducibility and validity study of a systematic review on ischemic cardiopathy. The Study Group of the Quality of Care (GRECA). *Med Clin (Barc)* **112**, Suppl. 1, 74–78.
14. Recurrent Miscarriage Immunotherapy Trialists Group (1994) Worldwide collaborative observational study and meta-analysis on allogeneic leukocyte immunotherapy for recurrent spontaneous abortion. *Am J Reprod Immunol* **32**, 55–72.
15. Cooper M, Ungar W & Zlotkin S (2006) An assessment of inter-rater agreement of the literature filtering process in the development of evidence-based dietary guidelines. *Public Health Nutr* **9**, 494–500.
16. World Cancer Research Fund/American Institute for Cancer Research (2007) Systematic Literature Review Specification Manual – version 15. http://www.dietandcancerreport.org/downloads/SLR_manual.pdf (accessed November 2007).
17. Greenland S & Longnecker MP (1992) Methods for trend estimation from summarized dose–response data, with application to meta-analysis. *Am J Epidemiol* **135**, 1301–1309.
18. Chêne G & Thompson SG (1996) Methods for summarizing the risk associations of quantitative variables in epidemiologic studies in a consistent form. *Am J Epidemiol* **144**, 610–621.