Welch et al. BMC Medical Research Methodology 2012, 12:187
http://www.biomedcentral.com/1471-2288/12/187

BMC
Medical Research Methodology

RESEARCH ARTICLE Open Access

Systematic reviews need to consider applicability to disadvantaged populations: inter-rater agreement for a health equity plausibility algorithm

Vivian Welch[1,2*], Kevin Brand[3], Elizabeth Kristjansson[4], Janet Smylie[5,6], George Wells[7,8,9] and Peter Tugwell[10,11,12]

Abstract

Background: Systematic reviews have been challenged to consider effects on disadvantaged groups. A priori specification of subgroup analyses is recommended to increase the credibility of these analyses. This study aimed to develop and assess inter-rater agreement for an algorithm for systematic review authors to predict whether differences in effect measures are likely for disadvantaged populations relative to advantaged populations (only relative effect measures were addressed).

Methods: A health equity plausibility algorithm was developed using clinimetric methods with three items based on literature review, key informant interviews and methodology studies. The three items dealt with the plausibility of differences in relative effects across sex or socioeconomic status (SES) due to: 1) patient characteristics; 2) intervention delivery (i.e., implementation); and 3) comparators. Thirty-five respondents (consisting of clinicians, methodologists and research users) assessed the likelihood of differences across sex and SES for ten systematic reviews with these questions. We assessed inter-rater reliability using Fleiss multi-rater kappa.

Results: The proportion agreement was 66% for patient characteristics (95% confidence interval: 61%-71%), 67% for intervention delivery (95% confidence interval: 62% to 72%) and 55% for the comparator (95% confidence interval: 50% to 60%). Inter-rater kappa, assessed with Fleiss kappa, ranged from 0 to 0.199, representing very low agreement beyond chance.

Conclusions: Users of systematic reviews rated that important differences in relative effects across sex and socioeconomic status were plausible for a range of individual and population-level interventions. However, there was very low inter-rater agreement for these assessments. There is an unmet need for discussion of plausibility of differential effects in systematic reviews. Increased consideration of external validity and applicability to different populations and settings is warranted in systematic reviews to meet this need.

Keywords: Systematic reviews, Applicability, Health equity, Sex and gender, Socioeconomic status

* Correspondence: Vivian.welch@uottawa.ca
[1]Clinical Epidemiology Unit, Ottawa Hospital Research Institute, Ottawa Hospital, Ottawa, ON, Canada
[2]Centre for Global Health, Institute of Population Health, University of Ottawa, 1 Stewart Street, Ottawa, ON K1N6N5, Canada
Full list of author information is available at the end of the article

Background

Health inequity is defined as avoidable and unfair differences in health [1]. Health inequity is caused by multiple, interacting factors described by the WHO Commission on Social Determinants of Health (CSDH) as 1) the socioeconomic and political context, 2) social position, 3) material circumstances and 4) the health system [2]. The health system described by the WHO CSDH is a broad concept which includes factors such as access to health care as well as the effectiveness of health care as well as public health and non-health interventions. Health care, health policy and non-health (e.g. social or financial) interventions may inadvertently increase health inequity if they are less effective in disadvantaged populations due to either prognostic factors (e.g. comorbidities or nutritional deficiencies) or treatment-covariate interactions (e.g. crowded home environment may increase transmission of infectious diseases, low literacy affects ability to benefit from written materials) [3].

Systematic reviews are useful as a basis for evidence-informed policy and practice [4,5]. Systematic reviews represent an opportunity to identify what works to promote health equity because they include studies conducted in a diversity of settings and populations allowing both prognostic factors and treatment-covariate interactions to be explored [6-8]. However, systematic reviews rarely assess whether interventions have intended or unintended effects on health equity. For example, only 1% of a random sample of Cochrane reviews assessed differences in effectiveness of interventions across socioeconomic or demographic factors [9]. Decision-makers such as NICE (National Institute of Clinical Excellence) in the UK and CADTH (Canadian Agency for Drugs and Technology in Health) in Canada are asking systematic reviews to assess the evidence of differential effects across age, sex and socioeconomic status. Failure to assess or consider effects on health equity in systematic reviews may lead to rejection of systematic reviews as a useful source of evidence for policy-makers who seek information on distribution of effects in the population [10,11], or may even lead to implementation of policies and programs which inadvertently increase health inequities [12,13].

Health inequity can be categorized using the PROGRESS-Plus framework –this is an acronym which summarizes factors across which differences in health may be considered inequitable depending on the setting and context: Place of residence; Race/ethnicity/culture; Occupation; Gender; Religion; Education; Socioeconomic status; Social capital [14]. There are additional personal characteristics (e.g. age, disability and sexual orientation) [15,16] that may present barriers to use of health and other public services; some of which are included in antidiscrimination laws. Furthermore, additional details may be helpful in describing vulnerable people depending on the focus of the study. For example, for studies of young people, additional characteristics related to vulnerability may include exclusion from school, looked after children or runaways. The term PROGRESS-Plus was coined to be a more inclusive framework for identifying both broad social and economic determinants of health (PROGRESS) and other characteristics, context and setting that are particularly relevant to the focus of study (Plus) [15]. Promoting health equity remains high on the agenda of local, national and international policy agendas [17].

Systematic reviews have been challenged to consider whether disadvantaged groups will obtain the same or more benefit than the mean effect, while at the same time minimizing the risk of spurious results due to multiple comparisons[12]. One way to reduce the chances of spurious results and increase the credibility of subgroup analyses is to specify comparisons prior to analysis [18-20] . Different authors are likely to argue *pre-hoc* for different groups. Thus, there is a need to assess how often there is disagreement amongst authors, and if so, develop guidelines for making these judgments.

This study aimed to develop and evaluate an algorithm to assess the likelihood of differences in relative effects of interventions in disadvantaged populations (across sex and socioeconomic status) relative to advantaged populations.

Methods

The health equity plausibility algorithm was developed using steps based on Feinstein [21] and Streiner and Norman [22].

Ethics approval

This research study was approved by the University of Ottawa Research Ethics Board (ethics approval certificate #H02-09-11b and #H02-09-11c).

Purpose

The purpose of the health equity plausibility algorithm is to assess the likelihood of differences in relative effect of an intervention across population characteristics defined by PROGRESS-Plus. For example, aspirin reduces the risk of stroke in women but not in men [23]. As another example, the measles antibody response in children is lower among the malnourished [23]. We chose to focus on only two PROGRESS-Plus characteristics because we felt that this would be a difficult task for respondents, and multiple characteristics might lead to lack of attention to additional characteristics. We selected sex and socioeconomic status because they are amongst the most commonly reported differences in effects and burden of

disease in the literature, and we felt respondents would feel the most able to make decisions about these characteristics. We recognize that differences across other characteristics such as ethnicity, educational attainment and occupation are important but these have not been assessed in this study.

It is important to note that the health equity plausibility algorithm is intended to predict likelihood of differences in relative but not absolute effects. Differences in absolute effects can be artifacts of differences in baseline risk, with no attendant difference in actual effect. For example, immigrants and refugees have a lifetime risk of developing active tuberculosis of over 35%. Assuming a relative risk reduction with isoniazid treatment of 93%, treating all immigrants and refugees will reduce lifetime cases of active tuberculosis by 33 in 100 people. In contrast, the typical high income country-born population has a low risk of developing latent tuberculosis: a typical prevalence in this case would amount to only 5 in 100 people. In this case the absolute reduction in number of cases with treatment would be only 2 per 100 people [24]; the difference of 33 per 100 for refugees or immigrants compared to 2 per 100 for high-income country born reflects a difference in baseline rates alone.

Item generation

Items were generated using three methods. First, existing checklists for applicability, transferability and external validity were assessed for factors related to judging likely differences in relative effects (Additional file 1: Appendix 1). Second, factors associated with subgroup analyses across PROGRESS-Plus were assessed in a systematic review of methods for assessing effects on health equity [25]. Third, practitioners and managers were interviewed using convergent interviewing [26] to identify factors associated with success or failure of program implementation in a vulnerable population [27].

Item reduction, questionnaire format, scaling, face validity

One author (VW) developed a draft algorithm using the above items. Items were phrased using wording from previously published checklists where possible. Dichotomous yes/no categories were chosen for the responses for two reasons: 1) a "don't know" category was considered unhelpful in determining likelihood; and 2) the ability of respondents to discriminate more finely than yes/no was uncertain. The draft algorithm was refined with two other authors (PT and GW). The face and conceptual validity was tested by asking four clinician methodologists with experience in clinical epidemiology, systematic reviews and health equity to review the items and judge their clarity and their ability to measure the concept of interest, defined as assessing the likelihood of differences in relative or absolute effects across PROGRESS-Plus characteristics[21]. Subsequent pilot-testing persuaded us to restrict attention to differences in relative effects, thereby dropping further consideration of absolute effects.

Consistency

Inter-rater reliability of the health equity plausibility algorithm was assessed by recruiting methodologists, clinicians and users of systematic reviews to apply the health equity plausibility algorithm to a sample of 10 systematic reviews. Thirty-five respondents (methodologists, clinicians and users of systematic reviews) were recruited by approaching members of the Cochrane Collaboration who attended the 2009 annual Colloquium. Respondents were asked to judge the likelihood of differences for two PROGRESS-Plus items: 1) sex: female vs. male and 2) low socioeconomic status vs. high socioeconomic status, using the survey (Additional file 1: Appendix 2).

Raters were given a summary for each of 10 systematic reviews inclusion criteria and methods for the population, intervention, comparison, outcomes and study designs included (columns 3–7 of Additional file 1: Appendix 3). Raters were not given the results of the systematic reviews.

The health equity plausibility algorithm was presented to raters as a list of three questions on one page, with a checkbox requiring Yes or No answers. Raters were also given examples of how each of the factors might create important differences in the magnitude of relative effects, based on examples from the updated Journal of American Medical Association (JAMA) User's Guide on applying results to individual patients [28].

Inter-rater agreement was assessed using Fleiss' Kappa for multiple raters for 60 assessments (10 systematic reviews X 3 questions X 2 PROGRESS-Plus factors) [29]. We also assessed and interpreted the proportion agreement [30].

The ten systematic reviews chosen for the survey were selected based on the following criteria: 1) proven effective and cost-effective interventions identified by the WHO-CHOICE (World Health Organization CHoosing Interventions that are Cost-Effective) initiative [31]; 2) representation of different types of intervention using the categories defined by the Disease Control Priorities Project of: i) Population-based primary prevention; ii) Personal interventions and iii) Policy instruments [32]; 3) included in the top ten causes of the global burden of disease projected for 2030 [33]; 4) availability of a systematic review with greater than five included studies including a diversity of settings and populations (a meta-analysis was not required since differences across populations can be assessed without a meta-analysis). Candidate reviews were identified by searching

MEDLINE and the Cochrane Library by VW. The ten reviews were chosen by discussion of two reviewers (VW, PT) based on the above criteria. They included nine Cochrane reviews and one non-Cochrane review. The attribute of being Cochrane or non-Cochrane reviews was not a criterion for selection since we did not feel there would be important differences in how systematic reviews approach the issue of differences across subgroups based on this attribute.

Construct validity

The ideal test of construct validity of this health equity plausibility algorithm is comparison with the criterion of whether there are real differences in effects by sex and socioeconomic status, irrespective of whether the instruments (systematic reviews and underlying studies) are able to detect these differences. Comparison to this criterion would be considered concurrent validity. However, since we cannot know the real intervention effects, we used a proxy by considering the results and discussion of the systematic reviews as an indicator of real differences in effects and compared this with the health equity plausibility algorithm to assess construct validity. We extracted these details from the discussion sections of each systematic review by assessing any mention of differences across PROGRESS Plus factors, whether they were supported by evidence or not (e.g. hypothesized differences not supported by evidence were included). We recognize that systematic reviews and underlying primary studies may not be designed or powered to detect such differences (column 2 of Additional file 1: Appendix 3). Therefore, the construct validity of the health equity plausibility algorithm was assessed by comparing whether raters' assessment of likelihood of important differences in effects were concordant with the discussion of applicability and generalizability by the authors of the systematic reviews across sex and socioeconomic status. Only those comparisons where 70% of raters gave the same judgment were compared with the systematic review conclusions about applicability and generalizability. This cut-off of 70% was chosen based on use of this criterion for consensus recommendations [34].

Results

Item generation

A draft health equity plausibility algorithm was developed by three authors (VW,GW,PT) that consisted of four yes/no items: 1) differences in implementation factors across PROGRESS-Plus (e.g. differences in resources); 2) likelihood of variations in the delivery of the intervention across PROGRESS-Plus (e.g. because of poor acceptability, inappropriate literacy level); 3) different mechanism of action of the intervention across PROGRESS-Plus; and 4) differences in expected absolute effects because of higher risk or prevalence across PROGRESS-Plus.

Item reduction, questionnaire format, scaling, face validity

Consultation with five content experts resulted in adding an item about possible differences in the comparator that might lead to differences in effectiveness. Differences in comparator was defined as whether the "control" groups differed in the context of the systematic review studies compared to the context to which the results would be applied. A question about absolute effects was removed, since it did not fit with the focus of this checklist to identify likelihood of differences in relative effects across PROGRESS-Plus. The revised health equity plausibility algorithm consisted of three questions (Table 1).

Inter-rater consistency

Ten systematic reviews were selected for the field test (Additional file 1: Appendix 3): 1) mass media to promote HIV testing [35]; 2) Population level tobacco control [36]; 3) psychological therapy for post-traumatic stress disorder [37]; 4) first-line anti-hypertensive drugs for people with hypertension [38]; 5) surgical interventions for age-related cataract [39]; 6) vaccines for measles, mumps and rubella [40]; 7) antidepressants vs. placebo in primary care [41]; 8) artemisinin-based combination therapy for uncomplicated malaria [42]; 9) primary safety belt laws [43]; and 10) handwashing for the prevention of diarrhea episodes [44].

Of 43 people contacted, 35 filled out the questionnaire (participation rate= 81%). The 35 respondents

Table 1 Health equity plausibility algorithm questions

Question 1:	Are there differences in patient/community/ population characteristics (e.g. underlying pathophysiology, comorbidities, patient attitudes, etc.) that are likely to create important differences in the magnitude of relative effect of the intervention versus the control for the outcome of interest?
Question 2:	Are there differences in the way that the intervention is delivered (e.g. provider compliance, provider skill, technical resources, availability of drugs/treatments) that are likely to create important differences in the magnitude of the relative effect of the intervention versus the control for the outcome of interest?
Question 3:	Are there differences in the comparator across patient, community or population that are likely to create important differences in magnitude of relative effects?

Table 2 Characteristics of 35 raters who assessed health equity plausibility

Experience with systematic reviews	User-6
	Methodologist-17
	Clinician-12
Years of experience	Median: 7
	Range : 2-15
Area of research/expertise	Public health-8
	Musculoskeletal-7
	Dermatology-1
	Child health-1
	Methods-9
	Family medicine-5
	Infectious disease-2
	Reproductive health-1
	Cancer-1

represented a mix of users, methodologists and clinicians with a median of 7 years of experience using or conducting systematic reviews and diverse clinical experience (e.g. public health, musculoskeletal, dermatology, cancer, infectious disease) (Table 2). No questions were skipped.

For the decision about whether important differences existed across sex (e.g. between males and females), the proportion agreement was 66% for patient characteristics, 54% for intervention delivery, and 51% for comparator. Across socioeconomic status, the proportion agreement was 66% for patient characteristics, 79% for intervention delivery, and 59% for comparator. The Fleiss kappa for multiple raters ranged from –0.001 to 0.199 (Table 3 and 4). A kappa of less than 0.2 is considered slight agreement [45], with kappa of 0.2<k<0.4 considered fair; 0.4<k<0.6 moderate, 0:6<k<0.8 substantial and k>0.8 almost perfect agreement.

As part of the survey, respondents were asked to explain the reasoning behind their answers (Table 5). The most common reasons for the answers were based on theory (n=10 raters) and personal experience (n=11). Other reasons for ratings were empirical data (n= 3) and "guesses" (n=4) (8). Respondents stated that other information would have been useful to complete the task, including more information about the interventions and outcomes, clarity on the comparator question and details on the size of difference considered important. Almost one third (10/35) of raters endorsed the importance of considering differences across PROGRESS-Plus in the design of systematic reviews (n=10 people).

Table 3 Sex/Gender: Health equity plausibility ratings for each question, across 10 systematic reviews

	Sex: Proportion judging important differences exist across sex			
Systematic review	*Question 1: Patient differences*	*Question 2: Delivery of intervention*	*Question 3: Comparator*	*Description in systematic review*
INTER-RATER AGREEMENT on Q1 ≥70%				
Mass media for HIV testing	96%	70%	57%	Sex differences not analyzed or discussed
Antidepressants for depression in primary care	92%	67%	50%	Sex not discussed or analyzed
Vaccines for MMR in children	8%	17%	25%	Sex not discussed or analyzed.
Primary safety belt laws	83%	67%	33%	Men have higher uptake of seatbelts
Psychological therapy for PTSD	83%	83%	52%	Studies including only females, all of whom had been assaulted, produced more positive results than the overall results.
INTER-RATER AGREEMENT on Q1 <70%				
Population tobacco control	70%	48%	48%	No differences found across sex
First line antihypertensives	65%	48%	43%	Females represented 45% of population. No subgroup analyses conducted on sex
Surgery for age-related cataract	67%	67%	67%	Sex not discussed
Hand washing for diarrhoea	67%	50%	50%	Analyses were age and sex adjusted, differences not discussed
ACT for malaria	33%	33%	25%	Sex not discussed
Fleiss Kappa	**0.199**	**0.068**	**0.005**	

Notes: PTSD: Post-traumatic stress disorder; SES: socioeconomic status; MMR: Measles, mumps and rubella vaccine.

Table 4 Socioeconomic status (SES): Proportion of respondents judging important differences exist for each question, across 10 systematic reviews

	Proportion of respondents judging important differences exist across SES				
	Average rating	*Question 1: Patient differences*	*Question 2: Delivery of intervention*	*Question 3: Comparator*	*Description in systematic review*
INTER-RATER AGREEMENT on Q1 ≥70%					
Mass media for HIV testing	87%	100%	83%	78%	Radio and television interventions can be used in literate and non-literate communities; therefore applicable to LMIC
Antidepressants for depression in primary care	84%	92%	92%	67%	SES not discussed
Population tobacco control	84%	91%	74%	87%	Price increases are more effective in low-income populations. Smoking restrictions: no SES differences
Hand washing for preventing diarrhoea	89%	83%	92%	92%	SES not discussed
INTER-RATER AGREEMENT on Q1 <70%					
Surgery for age-related cataract	86%	75%	100%	83%	In developing countries, access to expensive machines, volume of surgeries and skill of surgeons may be lower
Psychological therapy for PTSD	75%	78%	91%	57%	SES not discussed
ACT for malaria	72%	75%	92%	50%	SES not discussed
Primary safety belt laws	72%	58%	100%	58%	More effective for lower use groups (e.g. African-American and Hispanic in USA)
First line anti-hypertensives	67%	65%	83%	52%	SES differences not assessed.
Vaccines MMR in children	67%	50%	75%	75%	SES not discussed. "effectiveness demonstrated world-wide"
Fleiss Kappa		**−0.001**	**0.105**	**0.04**	

Construct validity

Our proxy for a "gold standard," namely a discussion of differences in effect across sex or socioeconomic status appearing in the conclusions of systematic reviews is presented in Additional file 1: Appendix 3, column 1. Each of the ten systematic reviews discussed plausibility of different effects for at least one of three factors: patient characteristics, intervention delivery or comparator. For example, the review of primary safety belt laws discussed greater effects in low use groups, including men, Hispanics and African-Americans. The tobacco control review assessed and found no differences across sex. It is important to note that some of the systematic reviews mentioned differences across PROGRESS-Plus factors, even though they were not supported by evidence. For example, the mass media review mentioned differences between literate and non-literate populations, but this interpretation was not supported by their data no part of any subgroup analyses.

Thirty-two assessments out of 60 (53%) attained greater than 70% agreement between raters (bolded text in Additional file 1: Appendix 4). Of those assessments, 28/31 agreed that differences were likely; 3/31 agreed that difference in effects was unlikely.

The type of intervention did not seem to be associated with agreement. Of those systematic reviews with greater than 70% agreement on the first question on patient characteristics, two were pharmacologic interventions (vaccines and antidepressants) and three were behavioural interventions (PTSD, safety belt laws and mass media) (Table 5).

Agreement with conclusions of systematic reviews

Seventeen of the assessments with greater than 70% inter-rater agreement (17/32; 53%) were concordant with our proxy for construct validity (the judgments of applicability described by the authors of the systematic reviews). For example, differences in effects due to patient characteristics (question 1) was judged likely for population tobacco control across socioeconomic status, which is supported by the conclusions of the systematic review that state that price control is more effective for low income populations [7].

Eight of the assessments (8/32; 25%) which reached greater than 70% agreement between raters were not consistent with the judgments of the authors of the systematic reviews.

Seven out of 32 assessments (22%) were classified as agreement "Not Known" since the systematic reviews

Table 5 Comments and reactions to making health equity plausibility judgments

Reason for answer	theory-10; personal experience- 11; empirical data-3 guesses-4;
Other information needed	more intervention specific information and data
	how big are the differences being sought?
	how does comparator overlap with intervention delivery
	was information available from trials?
	consider including community cluster trials
	Need information on how intervention was delivered
General comments	Important to consider these issues in design of SR-5
	Difficult- 4;
	Interesting to consider these issues-5;
	Subjective-6;
	Why only gender and SES- need to consider other factors (e.g. sexual orientation)- 2
	Country differences are important- e.g. if universal drug coverage is available-1
	Accessibility of drugs is less of an issue-1
	Ask questions about heterogeneity-1
	Intervention delivery is important for understanding
	Easy-1

did not discuss plausible or measured differences. For example, 91% of raters judged that important differences were likely across SES due to delivery factors for PTSD, but this was not discussed in the systematic review.

Discussion

The three questions used for this study were derived from questions which are given to users of systematic reviews to make judgments about clinical practice implications, such as in the JAMA User's Guide [28]. Although a high proportion of participants rated differences likely across socioeconomic status and sex, the Fleiss kappa indicated low to no agreement beyond chance between raters. Low kappa values in the presence of high agreement has been called one of the paradoxes of kappa [30,45]. Low kappa values may also indicate that there was low contrast in the systematic reviews or the raters, or that some systematic reviews were easier to agree on than others [46]. The low kappa in this study suggest that when making judgments about likely differences in effects, there is a need to have a deep understanding of the content area to make these ratings, and may require the involvement of multiple stakeholders. This has implications for the design of systematic review author teams and advisory boards, as well as for how to form panels of people who make judgments about applicability for the purpose of guidelines or policy decisions.

Secondly, the low kappa observed may be due to multi-component questions covering several factors, and potential confusion of access to health care, prognostic factors and treatment-covariate interactions. Since these three items were derived from questions recommended to users of systematic reviews, this study suggests a need to further refine these questions to improve understanding and increase inter-rater agreement beyond chance.

The strengths of this study are that it followed established steps in developing a checklist of item generation, pilot-testing and assessed consistency and construct validity in a field study. We selected systematic reviews for the field study based on predetermined criteria to maximize diversity of types of interventions (personal and population level) and disease areas with greater than five included studies including a diversity of settings and populations. These ten systematic reviews also included a mix of plausibility of different effects across the three questions.

This study is subject to some limitations. Firstly, the high proportion of "yes" answers may have been due to the way the questions were framed. Secondly, the rationale for affirmative responses was assessed by self-report on a questionnaire, and few details were provided by raters regarding how theory, data or personal experience was used to make these judgments. Thus, we could not assess whether differences were due to some of the reasons that we expected such as the type of study design (randomized vs. observational) or type of comparator (placebo vs. control). Also, given that personal experience was the most common reason for ratings, the raters individual characteristics may be important (e.g. whether they are women, low income, immigrants), and this could be explored in future studies. Both of these limitations could be assessed in further detail by cognitive interviewing using think out loud protocols to assess how respondents interpreted the questions and could be used to improve the questions [47]. Another possible limitation is the use of only two categories (yes/no), and the omission of a "don't know" category. Raters may have resorted to personal value judgments or random guesses when forced to choose between "yes" and "no". This might be addressed in future research by using more categories to reflect strong agreement to strong disagreement or by using a visual analogue scale; either of these methods would allow an assessment of confidence of raters in the importance of the difference. Finally, we recognize that our proxy for construct validity yielded unsatisfactory results. We suggest that this indicates a need for improved consideration and reporting of applicability in systematic reviews, particularly for populations to whom the results are likely to be applied.

Conclusions

This study reinforces the need to develop, evaluate and promote structured methods for considering applicability to relevant populations in systematic reviews. The low kappa suggests a need for a depth of content expertise and stakeholders on systematic review author teams in making such decisions. The results also suggest a need to improve the clarity of questions intended to help users of systematic reviews make applicability judgments. Planning and design of primary research studies and trials needs to take into account whether there are expected and plausible differences across population groups such as sex using appropriate methods such as stratification and pre-planned subgroup analyses.. The Campbell and Cochrane Equity Methods Group is conducting methodological research on methods such as the use of logic models, subgroup analysis, applicability judgments and process evaluation as tools for assessing the plausibility of effects on health equity [48].

Additional file

Additional file 1: Appendix 1. Applicability and transferability checklists [49-60]. **Appendix 2** Sample equity plausibility algorithm survey provided to raters. **Appendix 3** Characteristics of systematic reviews chosen for testing the equity plausibility algorithm [61]. **Appendix 4** Agreement of equity plausibility ratings between raters for each question and PROGRESS factor, across 10 systematic reviews.

Competing interests

Two members of the author team are members of the Campbell and Cochrane Equity Methods Group which conducts methodological research on assessing and reporting effects on health equity in systematic reviews (PT and VW).

Authors' contributions

PT and VW had the idea for the study. All authors contributed to the design of the study. VW collected and analyzed data for the study. All authors contributed to writing and revising the manuscript and approved the final version.

Acknowledgments

We would like to thank Mark Petticrew, Gord Guyatt, Sandy Oliver, and Inday Dans for comments on the survey items.

Funding

Vivian Welch was supported by a Canadian Institutes of Health Research Canada Graduate Scholarship for her doctoral studies, and a University of Ottawa entrance scholarship.

Author details

[1]Clinical Epidemiology Unit, Ottawa Hospital Research Institute, Ottawa Hospital, Ottawa, ON, Canada. [2]Centre for Global Health, Institute of Population Health, University of Ottawa, 1 Stewart Street, Ottawa, ON K1N6N5, Canada. [3]Telfer School of Management, University of Ottawa, Ottawa, ON, Canada. [4]School of Psychology, Centre for Global Health, Institute of Population Health, University of Ottawa, 1 Stewart Street, Ottawa, ON K1N6N5, Canada. [5]Centre for Research on Inner City Health, Keenan Research Centre of the Li Ka Shing Knowledge Institute, St. Michael's Hospital, Toronto, ON, Canada. [6]Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada. [7]Department of Epidemiology and Community Medicine and Department of Medicine, University of Ottawa, Ottawa, ON, Canada. [8]University of Ottawa Heart Institute, Ottawa, ON, Canada. [9]Clinical Epidemiology Unit, Ottawa Hospital Research Institute, Ottawa Hospital, Ottawa, ON, Canada. [10]Department of Epidemiology and Community Medicine and Department of Medicine, University of Ottawa, Ottawa, ON, Canada. [11]Clinical Epidemiology Unit, Ottawa Hospital Research Institute Ottawa Hospital, Ottawa, ON, Canada. [12]Centre for Global Health, Institute of Population Health, University of Ottawa, 1 Stewart Street, Ottawa, ON K1N6N5, Canada.

Received: 21 January 2012 Accepted: 5 December 2012
Published: 19 December 2012

References

1. Whitehead M: **The concepts and principles of equity and health.** *Int J Health Serv* 1992, **22**:429–445.
2. Kelly MP, Morgan A, Exworthy M, Popay J, Tugwell P, Robinson V, Simpson S, Narayan T, Myer L, Houweling T, *et al*: *The social determinants of health: developing an evidence base for political action: final report to the world health organization commission on the social determinants of health.* Geneva: 2007.
3. White M, Adams J, Heywood P: **How and why do interventions that increase health overall widen inequalities with populations?** In *Social inequality and public health.* 1st edition. Edited by Babones SJ. Bristol: The Policy Press, University of Bristol; 2009:65–82.
4. World Health Assembly: *Resolution 5834 on the ministerial summit on health research.* 2005:5834.
5. Lavis JN, Davies HTO, Gruen RL: **Working within and beyond the cochrane collaboration to make systematic reviews more useful to healthcare managers and policy makers.** *Healthcare Policy* 2006, **1**:21–33.
6. Mackenbach JP: **Tackling inequalities in health: the need for building a systematic evidence base.** *J Epidemiol Community Health* 2003, **57**:162.
7. Thomas S, Fayter D, Misso K, Ogilvie D, Petticrew M, Sowden A, Whitehead M, Worthy G, Thomas S, Fayter D, *et al*: **Population tobacco control interventions and their effects on social inequalities in smoking: systematic review. [Review] [109 refs].** *Tob Control* 2008, **17**:230–237.
8. Chopra M, Munro S, Lavis JN, Vist G, Bennett S: **Effects of policy options for human resources for health: an analysis of systematic reviews.** *Lancet* 2008, **371**:668–774.
9. Tsikata S, Robinson V, Petticrew M, Kristjansson B, Moher D, McGowan J, Shea B, Wells G, Tugwell P: *Do cochrane systematic reviews contain useful information about health equity? [abstract]. XI cochrane colloquium: evidence, health care and culture; 2003 Oct 26 31.* Barcelona, Spain: 2003:77.
10. Petticrew M, Whitehead M, Macintye SJ, Graham H, Egan M: **Evidence for public health policy on inequalities: 1: the reality according to policymakers.** *J Epidemiol Community Health* 2004, **58**:811–816.
11. Lavis JN, Davies H, Oxman A, Denis JL, Golden-Biddle K, Ferlie E: **Towards systematic reviews that inform health care management and policy-making.** *J Health Serv Res Policy* 2005, **10**:35–48.
12. Tugwell P, Petticrew M, Robinson V, Kristjansson E, Maxwell L: **Cochrane and campbell collaborations, and health equity.** *Lancet* 2006, **367**:1128–1130.
13. Petticrew M, Tugwell P, Welch V, Ueffing E, Kristjansson E, Armstrong R, Doyle J, Waters E: **Better evidence about wicked issues in tackling health inequities.** *J Public Health* 2009, **31**:453–456.
14. Evans T, Brown H: **Road traffic crashes: operationalizing equity in the context of health sector reform.** *Inj Control Saf Promot* 2003, **10**:11–12.
15. Kavanagh J, Oliver S, Lorenc T: *Reflections in developing and using PROGRESS-plus. Equity update (2).* Ottawa: 2008:1–3.
16. US Department of Health and Human Services: *Understanding and improving health.* Washington, DC: 2000.
17. Marmot M, Friel S, Bell R, Houweling TA, Taylor S: **Commission on social determinants of health: closing the gap in a generation: health equity through action on the social determinants of health.** *Lancet* 2008, **372**:1661–1669.
18. Sun X, Briel M, Walter SD, Guyatt GH: **Is a subgroup effect believable? Updating criteria to evaluate the credibility of subgroup analyses.** *Br Med J* 2010, **340**:C117.
19. Sun X, Briel M, Busse JW, You JJ, Akl EA, Mejza F, Bala MM, Bassler D, Mertz D, Diaz-Granados N, Vandvik PO, Malaga G, Srinathan SK, Dahm P, Johnston BC, Alonso-Coello P, Hassouneh B, Walter SD, Heels-Ansdell D, Bhatnagar N, Altman DG, Guyatt GH: **Credibility of claims of subgroup effects in randomised controlled trials: systematic review.** *Br Med J* 2012, **344**:e1553.

20. Petticrew M, Tugwell P, Kristjansson E, Oliver S, Ueffing E, Welch V: **Damned if you do, damned if you don't: subgroup analysis and equity.** *J Epidemiol Community Health* 2011, epub ahead of print.
21. Feinstein AR: *Clinimetrics*. New Haven: Yale University Press; 1987.
22. Feinstein AR: *Feinstein AR. Clinical epidemiology: the architecture of clinical research*. Philadelphia: Saunders; 1985:39–52. Philadelphia: Saunders; 1985.
23. Dans AL, Dans LF, Guyatt GH, Richardson S: **Users' Guides to the medical literature: XIV. How to decide on the applicability of clinical trial results to your patient. Evidence-based medicine working group.[see comment].** *JAMA* 1998, **279**:545–549.
24. Greenaway C, Sandoe A, Vissandjee B, Kitai I, Gruner D, Ueffing E, Wobeser W, Menzies D, Schwartzman K: **Tuberculosis: evidence review for newly arriving immigrants.** *Can Med Assoc J* 2010, **183**(12):E939–E951.
25. Welch V, Tugwell P, Wells GA, Kristjansson E, Petticrew M, McGowan J, DeMontigny J, Benkhalti M, Ueffing E: **How effects on health equity are assessed in systematic reviews of interventions.** *Cochrane Database Syst Rev* 2010, (12):MR000028.
26. Driedger SM, Gallois C, Sanders CB, Santesso N, Driedger SM, Gallois C, Sanders CB, Santesso N: **Finding common ground in team-based qualitative research using the convergent interviewing method.** *Qual Health Res* 2006, **16**:1145–1157.
27. Welch V, Smylie JK, Kristjansson E, Brand K, Tugwell P, Wells GA: *What is the role of systematic reviews in tackling health inequity?* University of Ottawa: 2010. PhD Thesis.
28. Dans AL, Dans LF, Guyatt G: **Applying results to individual patients**. In *Users' Guides to the medical literature: a manual for evidence-based clinical practice*. 2nd edition. Edited by Guyatt G, Rennie D, Meade MO, Cook DJ. New York, USA: The McGraw-Hill Companies, Inc; 2008:273–289.
29. Fleiss JL: **Measuring nominal scale agreement among many raters.** *Psychol Bull* 1971, **76**:378–382.
30. Cicchetti DV, Feinstein AR: **High agreement but low kappa: II.** *Resolving the paradoxes. Journal of Clinical Epidemiology* 1990, **43**:551–558.
31. Evans DB, Lim SS, Adam T, Edejer TT: **WHO choosing interventions that are cost effective (CHOICE) millennium development goals team: evaluation of current strategies and future priorities for improving health in developing countries.** *BMJ* 2005, **331**:1457–1461.
32. Jamison DT, Breman JG, Measham AR, Alleyne G, Claeson M, Evans DB, Jha P, Mills A, Musgrove P: *Priorities in health*. New York: Oxford University Press; 2006.
33. Mathers CD, Loncar D: **Projections of global mortality and burden of disease from 2002 to 2030.** *PLoS Med* 2006, **3**:3442.
34. Tugwell P, Boers M, Brooks P, Simon L, Strand V, Idzerda L: **OMERACT: an international initiative to improve outcome measurement in rheumatology.** *Trials* 2007, **8**:38.
35. Vidanapathirana J, Abramson MJ, Forbes A, Fairley C: **Mass media interventions for promoting HIV testing.** *Cochrane Database Syst Rev* 2005, (3):CD004775.
36. Main C, Thomas S, Ogilvie D, Stirk L, Petticrew M, Whitehead M, Sowden A: **Population tobacco control interventions and their effects on social inequalities in smoking: placing an equity lens on existing systematic reviews.** *BMC Publ Health* 2008, **8**:178.
37. Bisson J, Andrew M: **Psychological treatment of post-traumatic stress disorder (PTSD).** *Cochrane Database Syst Rev* 2007, (3):CD003388.
38. Wright JM, Musini VM: **First-line drugs for hypertension.** *Cochrane Database Syst Rev* 2009, Art. No. CD001841.
39. Riaz Y, Mehta JS, Wormald R, Evans JR, Foster A, Ravilla T, Snelligen T: **Surgical interventions for age-related cataract.** *Cochrane Database Syst Rev* 2006, (4). doi:10.1002/14651858.CD001323.pub2. Art. No. CD001323.
40. Demicheli V, Jefferson T, Rivetti A, Price D: **Vaccines for measles, mumps and rubella in children.** *Cochrane Database Syst Rev*, (4). doi:10.1002/14651858.CD004407.pub2. Art. No. CD004407.
41. Arroll B, Elley CR, Fishman T, Goodyear-Smith FA, Kenealy T, Blashki G, Kerse N, McGillivray S: **Antidepressants versus placebo for depression in primary care.** *Cochrane Database Syst Rev* 2009, (3). doi:10.1002/14651858.CD007954. Art. No CD007954.
42. Sinclair D, Zani B, Donegan S, Olliaro P, Garner P: **Artemisinin-based combination therapy for treating uncomplicated malaria.** *Cochrane Database Syst Rev* 2009, .
43. Shults RA, Nichols JL, Dinh-Zarr TB, Sleet DA, Elder RW: **Effectiveness of primary enforcement safety belt laws and enhanced enforcement of safety belt laws: a summary of the guide to community preventive services systematic reviews.** *J Safety Res* 2004, **35**:189–196.
44. Ejemot RI, Ehiri JE, Meremikwu MM, Critchley JA: **Hand washing for preventing diarrhoea.** *Cochrane Database Syst Rev* 2008, (1). doi:10.1002/14651858.CD004265.pub2. Art No. CD004265.
45. Kraemer HC, Periyakoil VS, Noda A: **Tutorial in biostatistics: kappa coefficients in medical research.** *Stat Med* 2002, **21**:2109–2129.
46. Vach W: **The dependence of Cohen's kappa on the prevalence does not matter.** *J Clin Epidemiol* 2005, **58**:655–661.
47. Napoles-Springer A, Santoyo-Olsson J, O'Brien H, Stewart AL: **Using cognitive interviews to develop surveys in diverse populations.** *Medical Care* 2006, **44**:s21–s30.
48. Tugwell P, Petticrew M, Kristjansson E, Welch V, Ueffing E, Waters E, Bonnefoy J, Morgan A, Doohan E, Kelly MP: **Assessing equity in systematic reviews: realising the recommendations of the commission on social determinants of health.** *BMJ* 2010, **341**:c4739.
49. Glasziou P, Guyatt GH, Dans AL, Dans LF, Straus S, Sackett DL: **Applying the results of trials and systematic reviews to individual patients. [see comment].** *ACP J Club* 1998, **129**:A15–A16.
50. National Health and Medical Research Council: *How to use the evidence: assessment and application of scientific evidence*. Canberra, Australia: 2000.
51. Briss PA, Zaza S, Pappaioanou M, Fielding J, Wright-De AL, Truman BI, Hopkins DP, Mullen PD, Thompson RS, Woolf SH, *et al*: **Developing an evidence-based guide to community preventive services-methods. The task force on community preventive services.** *Am J Prev Med* 2000, **18**(Suppl):43.
52. Rothwell PM: **Treating individuals 2. Subgroup analysis in randomised controlled trials: importance, indications, and interpretation.** *Lancet* 2005, **365**:176–186.
53. Wang S, Moss JR, Hiller JE: **Applicability and transferability of interventions in evidence-based public health.** *Health Promot Int* 2006, **21**:76–83.
54. Green LW, Glasgow RE: **Evaluating the relevance, generalization, and applicability of research: issues in external validation and translation methodology. [Review] [91 refs].** *Eval Health Prof* 2006, **29**:126–153.
55. Altman DG, Schulz KF, Moher D, Egger M, Davidoff F, Elbourne D, Gøtzsche PC, Lang T: **The revised CONSORT statement for reporting randomized trials: explanation and elaboration.** *Ann Intern Med* 2001, **134**:663–694.
56. Higgins JPT, Green S (Eds): *Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 [updated March 2011]*. The Cochrane Collaboration; 2011. Available from www.cochrane-handbook.org.
57. Lavis JN, Posada FB, Haines A, Osei E: **Use of research to inform public policymaking.[see comment].** *Lancet* 2004, **364**:1615–1621.
58. Dans AM, Dans L, Oxman AD, Robinson V, Acuin J, Tugwell P, Dennis R, Kang D: **Assessing equity in clinical practice guidelines. [Review] [53 refs].** *J Clin Epidemiol* 2007, **60**:540–546.
59. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gotzsche PC, Ioannidis JP, Clarke M, Devereaux PJ, Kleijnen J, Moher D: **The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration.** *BMJ* 2009, **339**:b2700.
60. *Tackling health inequalities in Europe: an integrated approach*, EUROTHINE: final report. Rotterdam: 2007.
61. Buendia-Rodriguez JA, Sanchez-Villamil JP, Buendia-Rodriguez JA, Sanchez-Villamil JP: **Using systematic reviews for evidence-based health promotion: basic methodology issues.** *Revista de Salud Publica* 2006, **8**(Suppl 2):94–105.

doi:10.1186/1471-2288-12-187
Cite this article as: Welch *et al.*: Systematic reviews need to consider applicability to disadvantaged populations: inter-rater agreement for a health equity plausibility algorithm. *BMC Medical Research Methodology* 2012 **12**:187.